Maastricht University

The Maastricht history-taking and advice checklist : studies of instrumental utility

Citation for published version (APA):

Kraan, H. F., & Crijnen, A. A. M. (1987). *The Maastricht history-taking and advice checklist : studies of instrumental utility*. [Doctoral Thesis, Maastricht University]. Rijksuniversiteit Limburg.

Document status and date:
Published: 01/01/1987

Document Version:
Publisher's PDF, also known as Version of record

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 19 Apr. 2024

THE MAASTRICHT HISTORY-TAKING

AND

ADVICE CHECKLIST:

STUDIES OF INSTRUMENTAL UTILITY

PROEFSCHRIFT

ter verkrijging van de graad van doctor in de geneeskunde
aan de Rijksuniversiteit te Maastricht,
op gezag van de Rector Magnificus, Prof.Dr. F.I.M. Bonke,
volgens besluit van het College van Dekanen
in het openbaar te verdedigen
op donderdag 12 november 1987 des namiddags om 15.00 uur.

door

Herro Foeke Kraan
geboren in 1944
te Wymbritseradeel

en

Alfons Arjen Marie Crijnen
geboren in 1956
te Eindhoven

PROMOTORES	Prof. Dr. M.A.J. Ramme
	Prof. Dr. M.W. de Vries
	Prof. Dr. W.H.F.W. Wijnen
	Prof. Dr. H. Philipsen
REFERENTEN	Prof. Dr. J.J.C.B. Bremer
	Prof. Dr. W. Brouwer
	Prof. A.D. Cox, M. Phil., F.R.C.P., F.R.C. Psych.
	Prof. Dr. M. Lipkin Jr.
	Dr. H.G. Schmidt
	Prof. Dr. J.A.M. Schouten

THE MAASTRICHT HISTORY-TAKING AND ADVICE CHECKLIST: STUDIES OF INSTRUMENTAL UTILITY

H.F. Kraan and A.A.M. Crijnen

Copyright (c) 1987 by Herro F. Kraan and Alfons A.M. Crijnen.

No part of this book may be reproduced in any form by print, photoprint, microfilm or any other means without the prior written permission from the authors.
ISBN 90-9001903-0

Cover picture: Martha Scheeren
Printed by Velder van den Hezelaer, Amsterdam

This book is distributed by Lundbeck, Amsterdam, The Netherlands

TABLE OF CONTENTS

Acknowledgements 9

1. Introduction to studies of the instrumental utility of the Maastricht History-taking and Advice Checklist. 13
H.F. Kraan and A.A.M. Crijnen

2. The medical interview and related skills. 29
H.F. Kraan and A.A.M. Crijnen

3. The medical interview: effects on the patient and the physician. 69
A.A.M. Crijnen and H.F. Kraan

4. The construction of the Maastricht History-taking and Advice Checklist. 81
H.F. Kraan and A.A.M. Crijnen

5. Assessing instrumental utility: issues of validity, reliability and scalability. 119
A.A.M. Crijnen and H.F. Kraan

6. Measuring Patient Satisfaction with the Communication. 145
A.A.M. Crijnen and H.F. Kraan

7. Scalability and reliability of the Maastricht History-taking and Advice Checklist in General Practice. 173
A.A.M. Crijnen and H.F. Kraan

8. Convergent and divergent validity of four measures of medical interviewing skills: a multitrait-multimethod approach. 203
A.A.M. Crijnen and H.F. Kraan

9. Interviewing skills and medical competence. 233
A.A.M. Crijnen, G.J. Post, H.F. Kraan, C. van der Vleuten, T. Imbos and J. Zuidweg

10. Scalability and reliability of the Maastricht History-taking and Advice Checklist in Primary Mental Health Care. 249
H.F. Kraan and A.A.M. Crijnen

11. Content validity of the Maastricht History-taking and Advice Checklist in Primary Mental Health Care. 279
H.F. Kraan and A.A.M. Crijnen

12. The convergent and divergent validity of the Maastricht History-taking and Advice Checklist in Primary Mental Health Care. 305
H.F. Kraan and A.A.M. Crijnen

13. Construct validity studies with the Maastricht History-taking and Advice Checklist in Primary Mental Health Care. 331
H.F. Kraan and A.A.M. Crijnen

14. Summary and conclusions. 355

15. Samenvatting en conclusies (Nederlands). 367

Appendices:

A Items and Manual for observers MAAS-General Practice. 379
A.A.M. Crijnen, H.F. Kraan, J. Zuidweg and J. van Dalen.

B Items and Manual for observers MAAS-Primary Mental Health Care. 405
H.F. Kraan, A.A.M. Crijnen, J. Zuidweg and J. van Dalen.

C MAAS-Self-evaluation in General Practice. 423

D MAAS-Self-evaluation in Primary Mental Health Care. 423

E Global Self-Rating Scale. 425

F Global Expert-Rating Scale. 427

G Generalizability analysis of Rasch homogeneous scales. 429

H Measurement of problem-solving in Primary Mental Health Care. 435

Curricula vitarum: H.F. Kraan 437
A.A.M. Crijnen

ACKNOWLEDGEMENTS

The first to whom we must express our thanks are our promotores Prof. Dr. M.A.J. Ramme and Prof. Dr. M.W. DeVries. With their endless and - sometimes - naïve faith in our competency Marius and Marten provided and secured the opportunity for us to carry out our study, it not being the mainstream of research in social and community psychiatry. Marius and Marten, sometimes differing in opinion from us concerning professional interests and methodological issues, stimulated creative thinking and urged for the excercise of simplicity and for highlighting the main topics instead of details.

Our co-promotores, Prof. Dr. W.H.F.W. Wijnen and Prof. Dr. H. Philipsen, offered us a multifaceted and inspiring support. Wijnand and Hans made us aware of our inclination towards abundant theoretical digressions or reckless methodological adventures. After consultations with them, we went our own way once again.

Furthermore, we owe our thanks to our referents, Prof. Dr. A.D. Cox, Prof. Dr. J.A.M. Schouten, Prof. Dr. M. Lipkin jr., Prof. Dr. J.J.C.B. Bremer, Prof. Dr. W. Brouwer and Dr. H.G. Schmidt, for their approval of this thesis and for their often extensive and enriching criticism.

The native ground of this thesis, however, has been the research program "Research in Medical Education" (program directors Dr. H.G. Schmidt, Prof.Dr. W.H.F.W. Wijnen) of which our research project "Measurement of clinical competency in the psychomedical domain" is a part. Our colleagues, Cees van der Vleuten, Tjaart Imbos and Wim Gijselaars, refused to carry out most of the statistical analyses (except the Rasch analyses). Instead, they pursued a more ambitious goal: the transformation of two clinicians into researchers. Indeed, they succeeded in teaching us methodology, advanced statistics, psychometrics and data analysis with statistical packages. We are very much indebted for their extensive investment in our academic careers.

Room 1.099, sometimes called the "research factory" by outsiders, served as our home, where we shared our researchers' life with our close co-workers Jaap Zuidweg, Leny Meertens, Piet Portegijs and Nicole

Zengerink and, at an earlier stage, Egbert van Wijk and Henriëtte Cuperus. They gave assistance in carrying out the experiments with simulated patients, in rating numerous interviews and in doing many organizational tasks, but - in particular - in providing moral support and setting limits to new ideas and unrealistic plans.

We are greatly indebted to the co-workers of the Skillslab, especially Jan van Dalen, Pie Bartholomeus and, at an earlier stage, Jutta König and Rob Groeneveld. They enabled us to carry out our experimental interviews with simulated patients and contributed much to the construction of the MAAS.

The group charged with the General Practitioner Residency Program was also crucial to the success of our experiments. Vic Dubois, Yvonne van Leeuwen, Jacques van Thiel, Herman Muller and Gijs Bak motivated their groups of residents to act as subjects in our experiments. Moreover, they offered help as experts and as observers of interviews.

Special honour is due to Trees Soute, who indefatigably collected our "products of mind" on the computerized typewriter and maintained an overview of the numerous versions of chapters and articles. Of course, she (and we) were also supported by the other members of the secretariate, in particular by Miranda van den Boorn, Gina Habets and JoAnn van Rooijen.

Furthermore, we are very grateful to colleagues from the Medical Faculty and from the RIAGG who contributed to our research as members of expert panels or as observers of videotaped interviews: Riet Schijns, Tillie Verstappen, Aad van Marrelo, Louk Gerainy, Frans Hermans, Pierre Verkoyen, Ger Brouns, Rob Rotteveel, Maarten Dormaar, Herman Baars, Jaap Berkhout, Wim van Zutphen and Piet Kerkhof.

We recall fruitful working relations with Gerrie Post and Riet Drop, who provided us with data from their project "The follow-up study of the physicians graduated in Maastricht" to be used in our validity studies. We also offer our thanks to Marianne Vroegop who deserves gratitude for her support in the preparation of the early Manual for observers using the MAAS.

In many respects simulated patients played crucial roles in our research; we are greatly indebted many thanks to them, especially to

Mrs. A. Bartelds, who gradually became famous for a depression that was not her own. Furthermore, we offer our thanks to Bart Hermens and Gerrie van Wunnik for their technical assistance.

The international character of our research project is witnessed by the collaboration with the U.S. Task Force on the Medical Interview (SREPCIM). Duscussions with their members were truly inspiring. Another important international contribution came from Carol Herman who transformed our "pidgin english" into a more readable language, and from Russell Hurlburt (University of Arizona) and Benjamin Maoz (University of Beersheva), who contributed inspiring and valuable advice.

We owe many thanks to Lundbeck Holland for the provision of funds for the publication of this thesis and to the Ludgardina Bouwman Stichting who kindly "supported the word processing" by the authors.

Last, but not least, we are much obliged to "our important others" who have convincingly reminded us of the relativity of writing this thesis.

CHAPTER 1 INTRODUCTION TO STUDIES OF THE INSTRUMENTAL UTILITY OF THE MAASTRICHT HISTORY-TAKING AND ADVICE CHECKLIST (MAAS)

H.F. Kraan and A.A.M. Crijnen

1.0 "Reasons for encounter".

Our interest in medical interviewing skills has lead us, over a period of 4 years, to construction and analysis of a method of measurement for this phenomenon.

In the early seventies, Kraan attended the courses on medical interviewing organized at the Department of Medical Psycholgy, University of Amsterdam. Crijnen was engaged in similar types of study groups organized and supervised by Prof. Dr. Paul Sporken at the Maastricht Medical School during the mid seventies. Early participation in these groups established our conviction of the importance of the medical interview for both patient and physician. Moreover, it witnessed the desire to enhance the quality of our own interviewing skills and to gain insight into the intricacies of a consultation process.

The present study has given us the opportunity to scrutinize this extensive domain and thus become acquainted with a gamut of opinions, theories and studies about medical interviewing skills. The development of the research design and experimental settings allowed us, furthermore, to study scientific methods and to apply them, through a creative process, to a feasible experiment with a simulated consultation hour.

The construction of the Maastricht History-taking and Advice Checklist (MAAS) and the subsequent studies in reliability and validity were accomplished due to the enthusiasm of inexperience since neither of us realized beforehand the dangers, pitfalls and demands in terms of time, energy and knowledge that were to be met.

The studies concerning the MAAS are a product of an on-going research-project called "measurement of competency in psychomedical care". This project was started in 1982 to evaluate the effects of an

innovative clerkship in mental health care at the Medical School, University of Limburg, Maastricht. Originally, we set out to design evaluative methods in various competency aspects, such as medical interviewing skills, clinical problem-solving and attitudes. Because of outside pressure from the faculty staff to construct an evaluative test for medical interviewing skills in general and of time constraints, we have devoted most of our time to measurement of medical interviewing skills.

1.1 The medical interview and related skills: definitions and delimitations.

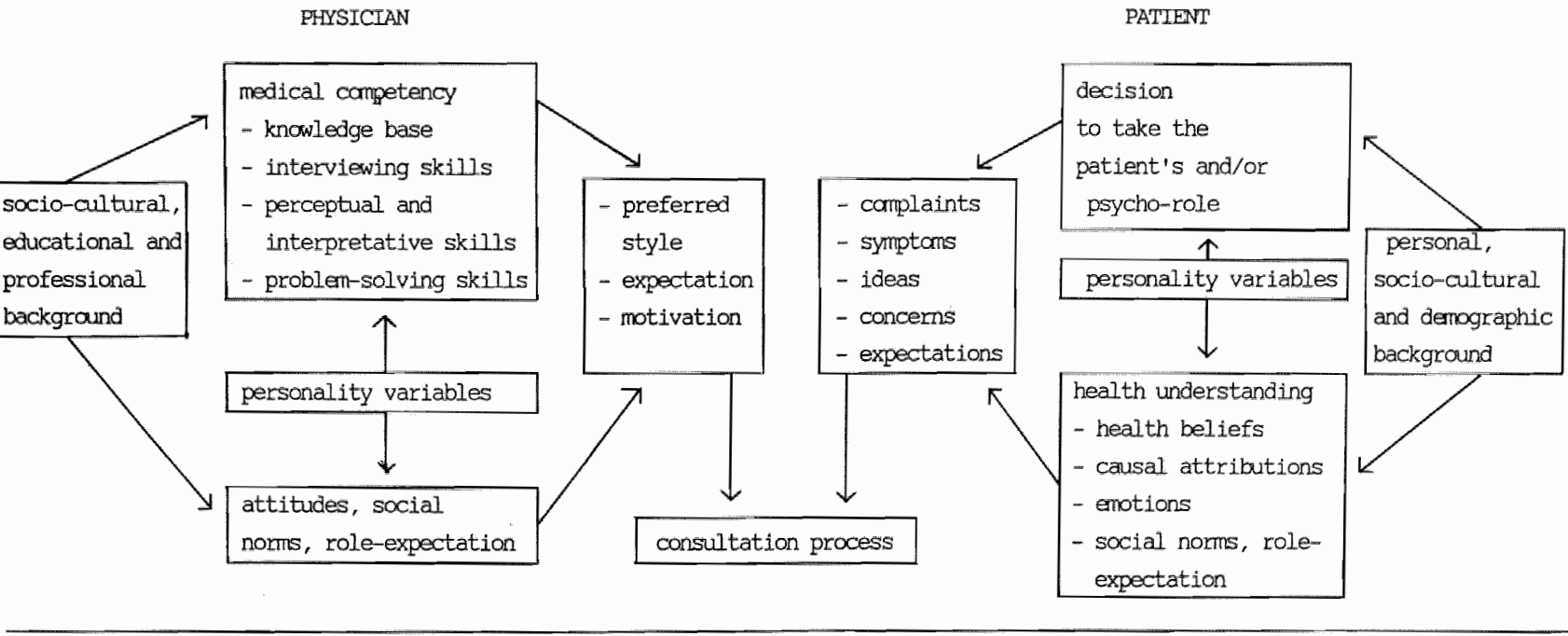
The medical interview has a function in medical consultation. In this chapter, we first define the medical interview from this functional perspective. We then go deeper into the relationship between medical interview and medical consultation. Medical interviewing is also to be considered as part of the physician's competency. We therefore provide a second definition of medical interview, from a competency perspective.

The medical interview has been defined by Schouten et al. (1981) from a functional perspective:

- collection of information for diagnostics and further patient management
- conveyance of information to the patient about diagnosis, aetiology and prognosis
- establishment and maintenance of a physician-patient relationship of trust and acceptance allowing the achievement of the two previous functions

Medical interview and medical consultation are difficult to distinguish (e.g. Pendleton et al., 1983). However, medical consultation is a term with a broader meaning than medical interview. It encompasses the patient's presentation of a (medical) problem and the physician's response in terms of diagnostics as well as preventive and curative advice.

Fig. 1.1 Physician and patient variables determining the medical consultation process.



Many variables, from both physician and patient, determine the medical consultation process. Figure 1.1, which presents an overview of these variables, is adapted from Pendleton's (1983) attempt to design a model of medical consultation.

The physician's contribution to the consultation is primarily determined by his educational background which determines his medical competency. Besides knowledge base, Fabb and Marshall (1983) distinguish interviewing, manual/perceptual skills, and problem-solving skills in medical competency. In addition, the physician's social norms and expectations as to his professional role (Parsons, 1951) and his attitudes towards patients and care-giving (Ajzen et al., 1980; Goldberg et al., 1980; Grol et al., 1985; Verhaak, 1986), determine his contribution to the consultation process.

The patient's contribution to the consultation consists, of course, of his account of his complaints and symptoms, but also of his concerns and ideas about health, illness and treatment. This presentation is strongly influenced by health beliefs (Becker and Maiman, 1975; Rosenstock, 1974), attributions (Rodin, 1978; Stoeckle et al., 1980) and attitudes/social norms (Ajzen et al., 1980). Like the physician's contribution, these variables are moulded by personality traits and socio-cultural background variables. Moreover, the decision making to enter into the patient-role (Parsons, 1951; Philipsen, 1969; Zola, 1973) or, particularly in the case of mental health care, into the "psycho-role" (Siegler et al., 1976), is an important variable. The majority of these patient variables, also topics of the medical interview, are discussed in the next chapter.

This complex model makes a further delineation necessary. We restrict ourselves therefore to the physician's competency aspect of the interviewing skills.

We therefore define the medical interview from a competency perspective as the dimension in consultation that is determined by the physician's interviewing skills.

Although this definition has the flavour of a tantology, it serves as a further restriction of our field of study to the medical interviewing skills and, in particular, to their measurement.

1.2 The medical interview and medical interviewing skills.

Several arguments are widely recognized as underscoring the importance of teaching and research into the medical interview and related skills:

- The medical interview is the medium of every consultation. Physicians spend about 60% of their working time interviewing (Lipkin, et al., 1984).
- As a function of the medical interview, accurate data collection is a prerequisite for accurate diagnosis (Goldberg et al., 1982; Elstein et al., 1978; Cox et al., 1981).
- Patient's satisfaction with a medical consultation is dependent on the feelings of the patient being understood and taken seriously (Pendleton, 1983).
- Provision of information about the nature of the illness, its causes, its prognosis and about its further management leads to patient compliance and therefore enhances the effectiveness of health care (DiMatteo et al., 1982).
- Medical interviewing skills contribute to establishing and maintaining the physician-patient relationship (Pendleton et al., 1984).
- Medical interviewing skills are indispensable for the detection of mental health problems (Goldberg et al., 1980), prevention of somatic fixation (Grol et al., 1981) and basic treatment of psychosocial problems (Ivey, 1983).

These arguments stress the importance of medical interviewing skills in health care, medical education and research. The need, therefore, for measurement and evaluation methods is evident.

1.3 Measuring interviewing skills.

Measurement is not an aim in itself. Objectives, subject matter and criteria for measurement, as well as psychometric requirements, should be clearly settled before an appropriate measurement method can be selected or constructed.

- Objectives of measurement.

Our measurement method of interviewing skills should be used as an evaluation instrument and as an aid to education in interviewing skills. These objectives have two strong implications for the selection of interviewing skills to be included in the method. First, it implies that these skills can be acquired through training and education (Van Dalen et al., 1981) in the sense that the interviewer:

- knows when pertinent skills should be applied
- knows how to apply these skills
- can actually apply these skills
- knows which effects could result from these skills

These teachability conditions narrow down our definition of medical interviewing skills somewhat because it excludes behavior during medical interviews that cannot be taught (some non-verbal behavior, idiosyncratic behavior related to the person of the interviewer). Nevertheless, this behavior may be important in relation to the outcome of the interview.

The second implication is that the focus of measurement is on the physician's behavior during the interview.

- Subject matter of measurement.

The physician's interviewing skills are the subject of measurement. To be more specific we delineate this field even further. First, we are interested in medical interviews in primary care because of the superordinate objectives of the medical school where this research project has been carried out.

Second, we selected "initial" interviews where the patient comes up with a problem for consultation that (s)he considers as "new". "Initial" interviews are more uniform in their structure than "follow-up" interviews and, therefore, are easier to fit into an educational model for interviewing skills (see next item). Moreover, initial interviews are of paramount importance in primary care where a patient's career and medical decision-making start.

- Criteria for measurement.

Criteria setting is indispensable for the step from measurement to evaluation. To make comparable judgments, standards should be

available. This is discussed in chapter 4.

For educational purposes, we constructed an educational model stating the interviewing skills in concrete terms to be mastered. This model, which is an indispensable prerequisite for educational programs (Carroll et al., 1980), is given in the next chapter.

- Psychometric requirements.

The psychometric requirements to transform medical interviewing skills as concept-defined into variables, are described by De Groot (1961). He mentions four criteria making up the "instrumental utility" of these variables:

a) reliability: is the measurement consistent under varying conditions of subjects, interview situations, cases/problems and observers?
b) scalability: can the operationalizations of interviewing skills, in short, the items, be fitted in a measurement scale?
c) validity: are the measurement scores real reflections of the physician's interviewing ability?
d) practicability: is the measurement practical within the objectives of evaluation and education?

In this thesis, these conditions of measurement are applied in the construction of a measurement method for interviewing skills.

1.4 The Maastricht History-taking and Advice Checklist (MAAS).

The afore-mentioned considerations have resulted in the construction of two observation instruments intended to measure the interviewing skills of physicians during initial consultations in general practice (MAAS-GP) and in primary mental health care (MAAS-PMHC).

The items are divided into 6 scales. The first 3 scales pertain to skills used in the three characteristic phases of initial medical consultations. The remaining three scales measure skills pertaining to process aspects of interviewing.

Scale I, "exploring reasons for encounter", measures the ability to clarify the patient's complaint(s) and to explore the motives in the pre-patient phase leading to the visit to the physician. This is the

patient-centered part of the medical interview.

Scale II consists of "history-taking skills". Through this type of questioning, the physician is able to generate hypotheses and to explain the patient's complaint(s) in medical terms.

In the MAAS-primary mental health care, this scale is divided into three parts. In addition to the scale "history-taking in sensu strictu", two further scales have been added. "Psychiatric exploration" is a scale that measures the skills for exploring possible psychiatric symptoms in order to make a psychiatric assessment of the patient. "Socio-emotional exploration" is the other scale added; this measures the extent to which the physician explores aetiological conditions or consequences of mental health problems.

Scale III deals with the interviewing skills involved when the physician is "presenting solutions". It pertains to the exchange of information about the medical problem (cause, prognosis) and to the negotiation between physician and patient about the problem's definition and solutions to the problem (advice, referral, etc.).

By Scale IV, the way physicians "structure the interview" (introduction, balance of patient- and physician-centered styles of interviewing, sequence of the different phases, closing), is judged.

Scale V consists of "interpersonal skills" which establish rapport with the patient. Moreover, the quality of the physician's response to emotions is assessed.

Scale VI pertains to "communicative skills", which aim to promote the exchange of information between physician and patient.

1.5 Research questions concerning the reliability, scalability and validity studies.

A method to measure interviewing skills is regarded as reliable when the method shows consistency and stability over varying conditions of measurement. The method has to prove its constancy when used by different observers, when used in different measurement situations and when used with different patients. This latter facet, also called "inter-case reliability" (Swanson, 1981), is important because patients differ in the nature of their medical problems and in the way they

present their problems to the physician.

We study the reliability of measurement by investigation of the sources of unreliability, such as differences in observers, in the nature of cases and their presentation by patients and in the method itself. This is carried out by an analysis of variance, called "generalizability analysis" (Mitchell, 1979).

Weaknesses in our method of measurement are detected by investigating to what extent its items fit in scales. This research of "scalability" is carried out with the so-called Rasch-analyses (Wright et al., 1979).

A method to measure interviewing is valid when it really measures what it is intended to measure. Validity studies are carried out in accordance with three questions:

a) Is the content of the method representative for the measurement of all the interviewing skills included in our educational model of initial interviewing skills? This model is presented in chapter two. A study of the representativeness of the item domain or, in other words, its content validity, is carried out within the MAAS-primary mental health care.

b) Validity can also be studied by comparing the measurement properties of the MAAS with those of comparable methods.
When, according to predefined hypotheses, the MAAS shares measurement properties, then it supports the so-called convergent validity. When the MAAS measures, also according to part of our hypotheses, different properties from the compared methods, then this is a sign of the so-called divergent validity. These aspects are investigated in a so-called multitrait-multimethod matrix (Campbell et al., 1959), where the MAAS is compared with 3 related measurement methods of interviewing skills.

c) Validity of a method is also established when, in measurement, "theoretical constructs" are confirmed.
In our study, two questions about such "theoretical constructs" are answered:
- Is the competency aspect "interviewing skills" as measured with the MAAS really different from dimensions such as medical

knowledge, medical problem-solving or attitudes towards care-giving? In this study, conducted using only the MAAS-general practice, the MAAS is compared with other methods which measure different aspects of medical competency.

- Is it possible, with the MAAS, to measure dimensions of interviewing skills that cause hypothesized changes in the patient's interview behavior and his satisfaction with the interview? These changes are measured by means of the so-called Patient Satisfaction Checklist (PSCC). They concern variables such as the patient's feeling of being facilitated to state his problems in his own words; the feeling of being disrupted in this; increased informedness and insight of the patient into his own medical condition; feelings of being too strongly directed towards a solution of the physician's choice and the patient's intention to comply with the proposed preventive or curative advices.

These research questions are answered for the MAAS-General Practice as well as for the MAAS-Primary Mental Health Care.

1.6 Research settings and subjects.

Throughout the entire research project, reliability and validity studies are performed with residents in general practice and medical students interviewing actual and simulated patients. Within the more restricted scope of this thesis, only studies with residents in general practice interviewing simulated patients are included. The studies with the Patient Satisfaction with Communication Checklist have been carried out partly with actual patients.

Simulated patients can be defined as individuals trained to portray the history, physical findings and emotions of an actual patient (Norman, 1985). In chapter 5, a discussion of the validity of simulated patients for these studies is included.

1.7 Survey of chapters.

We summarize below the content of the chapters. It is important to delineate which reliability and validity studies are performed with

what kind of subjects/interviewers. We further indicate the primary author responsible for the pertinent chapter.

Chapter 2 describes the theoretical underpinning of our educational model for initial medical interviews. From this model, the physician's interviewing skills are derived (Kraan).

Chapter 3 is a review of the literature concerning the effects of the medical interview on the patient in terms of outcome variables (the patient's satisfaction, recall, insight into conveyed information, intention to comply, etc.) (Crijnen).

Chapter 4 discusses the construction of the MAAS and its variants. The selection of the item domain from the theory is discussed. The MAAS is compared with more than 20 other methods of measuring interviewing skills as described in the literature. The decision to construct the MAAS as a new observation instrument is discussed (Kraan).

Chapter 5 describes the methodology of scalability, reliability and validity research as used in this thesis. It expands further on the validity of "simulated" patients as used in this thesis (Crijnen).

Chapter 6 is an article describing the construction and scaling as well as the reliability measures of the Patient Satisfaction with Communication Checklist (PSCC). It is a variant of the MAAS included for use by patients to judge the physician's interviewing skills and their effects. The PSCC is pre-tested in about 240 medical interviews (Crijnen).

Chapter 7 describes the results of reliability and scalability studies of the MAAS-General Practice in a sample of 40 residents in general practice, each interviewing two simulated patients, one presenting pre-cardial pain and the other inception of diabetes mellitus (Crijnen).

Chapter 8 studies convergent and divergent validity by comparing the MAAS-General Practice with two self-evaluation methods and experts' judgments of interviewing skills. The research situation as described in the previous chapter is again used (Crijnen).

Chapter 9 deals with a construct validity study correlating MAAS-General Practice scores with measurement of medical knowledge, attitude ratings, global ratings of interpersonal and communicative skills and

measurements of medical problem-solving (Crijnen).

Chapter 10 provides the reliability and scalability studies of the MAAS-Primary Mental Health Care in a sample of 40 residents in general practice, each interviewing two simulated patients, one presenting major depression and the other panic disorder (Kraan).

Chapter 11 reports the content validity of the MAAS-Primary Mental Health Care by further elaboration of the reliability on the item level and by comparison of MAAS scores with experts' judgments of interviewing skills (Kraan).

Chapter 12 corresponds to chapter 8 in studying the convergent and divergent validity of the MAAS-Primary Mental Health Care (Kraan).

Chapter 13 is a construct validity study, relating interviewing skills with PSCC variables such as patient-satisfaction, recalled information, insight and intention to comply (Kraan).

In chapter 14, the results of these studies are summarized and their theoretical and practical implications are discussed (Kraan/Crijnen).

Chapter 15 is the translation in Dutch of the previous chapter.

In the appendices, the observers' manual of the MAAS-General Practice and MAAS-Primary Mental Health Care can be found (Kraan/Crijnen).

1.8 Editorial remarks.

- Words frequently encountered in this thesis, such as physician, student, subject, observer, interviewer, etc. and their derivatives, such as his, him, etc. are quite arbitrarily put in the same sex as the authors'.
- To enhance the readability of the text, much tabulated data material has been put in appendices.

REFERENCES

Ajzen I, Fishbein M. Understanding attitudes and predicting social behavior. Prentice Hall, Englewood Cliffs, 1980.

Becker MH, Maiman LA. Socio-behavioral determinants of compliance with medical care and recommendation. Medical Care, 1975; 13: 10-24.

Campbell DT, Fiske DW. Convergent and discriminant validation by the multi-trait multi-method matrix. Psychological Bulletin, 1959; 56: 81-105.

Carroll JG, Monroe J. Teaching clinical interviewing in the health professions; a review of empirical research. Evaluation and the Health Profession, 1980; 3: 21-45.

Cox A, Rutter M, Holbrook D. Psychiatric interviewing techniques V. Experimental study: eliciting factual information. British Journal of Psychiatry, 1981; 139: 29-37.

Dalen J van, Phaff Ch. Arts-patiënt relaties; een uitwisseling van deskundigheden. In: Pierloot R. Arts-patiënt relaties. Stafleu, Alphen a/d Rijn/Brussel, 1981.

DiMatteo MR, DiNicola DD. Achieving patient compliance; the psychology of the medical practitioner's role. Pergamon Press, New York, 1982 (chapter 2).

Elstein AS, Shulman LS, Sprafka SA. Medical problem-solving. An analysis of clinical reasoning. Cambridge, Mass. Harvard University Press, 1978.

Fabb WE, Marshall JR. The assessment of clinical competence in general family practice. MTP Press, Lancaster, 1983.

Goldberg D, Huxley P. Mental illness in the community; the pathway to psychiatric care. Tavistock Publications, London, 1980.

Goldberg D, Steele J, Johnson A, Smith C. Ability of primary care physicians to make accurate ratings of psychiatric symptoms. Archives of General Psychiatry, 1982; 39: 829-833.

Grol R, Eyk J van, Huygen F, Mesker P, Mesker-Niesten J, Mierlo G van, Mokkink H, Smits A. Huisarts en somatische fixatie. Theorie en praktijk van de preventie van somatische fixatie. NUHI, Nijmegen, 1981.

Grol R, Eyk J van, Mokkink H, e.a. Taakopvatting van de huisarts en zijn handelen in de spreekkamer. Gezondheid en Samenleving, 1985; 6: 31-40.

Groot AD de. Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen. Mouton, 's-Gravenhage, 1961.

Ivey AE. Intentional interviewing and counseling. Wadsworth, Belmont, Calif., 1983.

Lipkin M, Quill TE, Napadano RJ. The medical interview: a core curriculum for residencies in the internal medicine. Annals of Internal Medicine, 1984; 100: 277-284.

Mitchell SK. Interobserver agreement, reliability and generalizability of data collected in observational studies. Psychological Bulletin, 1979; 86: 376-390.

Norman GR. Simulated patients. In: Neufeld VR, Norman GR (Eds.). Assessing medical competence. Springer Publ. Cie., New York, 1985.

Parsons T. The social system. Glencoe: The Free Press, 1951.

Pendleton D. Doctor-patient communication: a review. In: Pendleton D, Hasler J (Eds.). Doctor-patient communication. Academic Press, London, 1983.

Pendleton D, Schofield T, Tate P, Havelock P. The consultation; an approach to learning and teaching. Oxford University Press, Oxford, 1984.

Philipsen H. Afwezigheid wegens ziekte. Wolters-Noordhoff, Groningen, 1969.

Rodin J. Patient-practitioner relationships; a process of social influence. In: Johnson AW, Grusky O, Raven BH (Eds.). Contemporary health services; social science perspectives. Auburn House, Boston, 1982.

Rosenstock IL. Historical origins of the Health Belief Model: origins and correlates in psychological theory. Health Education Monographies, 1974; 2: 336-353.

Schouten JAM. Anamnese en advies. Stafleu, Alphen a/d Rijn/Brussel, 1981.

Siegler M, Osmond H. Models of madness, model of medicine. Harper and Row, New York, 1976.

Stoeckle JD, Barsky AJ. Attributions: uses of social science knowledge in the "doctoring" of primary care. In: Eisenberg L, Kleinman A (Eds.). The relevance of social science for medicine. Reidel Publ. Co., New York, 1980.

Swanson DB, Mayewski RJ, Norsen L, Baran G, Mushlin AI. A psychometric study of measures of medical interviewing skills. Proceedings of the 20th Annual Conference on Research in Medical Education, 1981: 3-8.

Verhaak PFM. Interpretatie en behandeling van psychosociale klachten in de huisartsenpraktijk; een onderzoek naar verschillen tussen huisartsen. NIVEL, Utrecht, 1986.

Wright BD, Stone MH. Best test design. Mesa Press, Chicago, 1979.

Zola IK. Pathways to the doctor: from person to patient. Social Science in Medicine, 1973; 7: 677-689.

CHAPTER 2 THE MEDICAL INTERVIEW AND RELATED SKILLS

H.F. Kraan and A.A.M. Crijnen

2.0 Introduction.

For an educational program of interviewing skills, an (ideal-typic) model of an initial medical interview is necessary (Carrol et al., 1980). It should describe, in clear, concrete terms, the interviewing skills to be mastered. The model is to provide students and evaluators with a standard.

In this chapter, we start with an educational model of an initial interview in General Practice and Primary Mental Health Care (2.1.1). Next, we justify this model, describing its main characteristics (2.1.2). Finally, from this model we deduce a matrix of interviewing skills necessary to perform initial interviews in General Practice and Primary Mental Health Care (2.3 and 2.4).

2.1 The educational model of the initial interview.

2.1.1 The three characteristic phases.

In their most simple form, initial medical interviews in general practice, as well as in primary mental health care, have three phases, i.e. exploration of the reasons for encounter, medical history-taking and presenting solutions.

a) Exploring reasons for encounter.

In this phase, the physician asks the patient to describe the complaint(s) in his own way. The physician explores co-existing emotions and concerns. The patient is questioned about his own explanations of the problems and about possible factors influencing them. Moreover, the patient is invited to talk about his attempts so far to solve the problem, about solutions which have proved succesful in the past, and about other comparable situations. Information is also asked concerning the discussion of his complaint(s) within his family or so-called primary group and the decision to seek help.

Very important in this phase is the opportunity for the patient to

express his emotions, fears, worries etc.

Finally, wishes and expectations concerning the desired help from the physician should made clear.

This phase requires a patient-centered, non-directive, explorative, open style of questioning.

The goal of this phase is to clarify in which way the patient desires to be helped with his problem.

b) Medical history-taking.

The questions asked during this phase reflect the classical medical inquiry strategy which is the consequence of the clinical reasoning process. What-, where-, how-, when-questions are posed in order to collect the information needed to generate and to check explanatory hypotheses. Within this clinical reasoning framework, factors are also inquired after which precipitate, maintain and diminish the problem. The physiological systems (as well as past medical history) are reviewed.

In case of psychiatric or psychosocial problems, questions concerning social relationships, biography, important life-events and stress factors are asked. An exploration of psychiatric symptoms is also necessary.

The basic interviewing style during this phase is also non-directive. However, hypotheses-generation may require systematic questioning about major areas of life, whereas in hypotheses testing, directive and closed-ended questions are often necessary.

The goal of this phase is to collect the information necessary for diagnostics and clinical problem-solving.

c) Presenting solutions.

During this phase, information is conveyed by the physician concerning his findings and about his clinical reasoning process. A possible diagnosis may be stated as well as information given about causes, conditions and prognosis. Emotional reactions to this information should be explored with the patient. Supplementary information may be needed. Opportunity for negotiation about the problem definition should be given.

Next, the physician proposes one or more possible solutions: further exploration or examination, treatment possibilities etc. Sometimes,

referral may be needed. Again a negotiation process may follow after the patient is offered alternatives and additional information concerning these alternatives' pros and cons. The final result may be a definitive advice to the patient which should be explored regarding its feasibility. Finally, appointments for follow-up should be made.

In this phase, conveyance of information alternates with periods of exploration of ideas and emotions and of negotiation.

The goal of this phase is the conveyance of information concerning the patient's condition, the proposal for further management and the exploration of the patient's reactions to both of these.

2.1.2 General characteristics of the model.

In this section, we discuss some general notions characterizing this model. Subsequently, the functions of the medical interview in relation to the medical consultation process (2.1.2.1), the background of its phasic structure (2.1.2.2), and the combination of patient and physician-centeredness (2.1.2.3) is discussed.

2.1.2.1 The functions of the medical interview in relation to the consultation process.

In the previous chapter, we discussed Schouten's (1982) three-function model of the medical interview:

1) collection of information for diagnosis and clinical reasoning;
2) conveyance of information on diagnosis, condition, prognosis as well as information on preventive and curative measures;
3) establishment and maintenance of an optimal physician-patient relationship.

By the achievement of these functions, the medical interview is the communicative, expressive dimension of the medical consultation. Lipkin et al. (1984) denote this functional relation by calling the medical interview the "medium" of the medical consultation. It is self-evident that current models of medical consultation determine our educational model of medical interviewing.

The model of initial medical consultation elaborated by Pendleton et al. (1984), has been influential in general practice. The Pendleton model states seven tasks for the physician:

1. To define the reasons for the patient's attendance including:
   - the nature and history of the problems
   - their aetiology
   - the patient's ideas, concerns and expectations
   - the effects of the problems
2. To consider other problems:
   - continuing problems
   - at-risk factors
3. To choose with the patient an appropriate action for each problem.
4. To achieve a shared understanding of the problems with the patient.
5. To involve the patient in the management and encourage him to accept appropriate responsibility.
6. To use time and resources appropriately.
7. To establish or maintain a relationship with the patient, which helps to achieve the other tasks.

In primary mental health care, a model, named "customer's approach to patienthood", extensively described by Lazare et al. (1975), has a strong impact on our educational model. It bears resemblance to the Pendleton model and to the well-known Dutch model of "Methodisch werken" (A General Model of Care-giving for General Practitioners, Holten-Vriesema et al., 1978).

We provide a short description of Lazare's model. It starts with the assumption that the patient has something in mind that he wants, what is called a "need" or "goal". In addition, patients know with "considerable specificity" how they would like the physician to intervene on their behalf. It is the physician's task to elicit the patient's request for help, defined as how the patient would like the physician to help him to achieve the desired goal. In the author's opinion, the elicitation of the request for help "turns out to be a very intimate revelation", requiring an optimal rapport between patient and physician. These requests show a great variety: advice, control, counseling, social intervention, nothing, etc. Complaint(s) (the patient's initial statement of what is bothering him) are invariably elicited. Goals are usually clarified by the physician; requests are often not. The authors further assert the special diagnostic value of

eliciting the patient's request: this is usually exactly what he needs. Moreover, it yields a lot of information about the patient's ideas about what is wrong.

The next step is "negotiation", the core of this consultation model: the patient formulates what he thinks he needs whereas the physician formulates what is medically appropriate. The statement of the request for help itself obliges the physician to consider the legitimacy of the patient's demand and to explain why an alternative formulation might be more valid. The patient has the right to evaluate and, ultimately, to accept or reject any treatment proposal. He may expect additional explanations, alternative treatment plan(s) or a statement that the physician cannot meet his request.

The notion of functionality also makes the relationship between content and process of the medical interview more clear. In the functions of "collection and conveyance of information", the content elements of the interview (medical data, advice, but also facts from the patient's living circumstances, concerns, fears etc.) are of paramount importance.

In the third function: "establishment and maintenance of an optimal physician-patient communicative relationship", the mode of communication is very important. For this function, the process skills or interviewing style is most relevant.

2.1.2.2 The phasic structure of the medical interview.

This phasic structure is a consequence of the phasic course inherent to the medical consultation.

The phasic structure is a universal phenomenon not only in problem-solving, but in all creative processes. In these processes, thinking is always goal-directed. The attainment of a goal is a problem for the acting, problem-solving or creating subject. It is problematic in so far as the means to attain the goal are not immediately available. Characteristic for this process is the freedom to choose a means of achieving the goal and the uncertainty about its adequacy. The tentative solutions will be checked in reality on the grounds of their effectiveness. According to De Groot (1961), the foundation of this phasic process is the empirical cycle with its successive steps in

thinking and action of "observe-guess-predict-check", which is universal to all human thinking and creative processes.

Indeed, the notion of this phasic course is ubiquitous not only in care-giving or "planned change" in its broad sense (Bennis et al., 1962; Romme, 1967; Van Beugen, 1977), but also in the literature on the medical interview (Byrne and Long, 1976; Stiles et al., 1979; Verhaak, 1986).

2.1.2.3 Patient- and physician-centeredness in the medical interview.

The term "patient-centered medicine" was introduced by Balint et al. (1970). It denotes that the physician should enter into the patient's world, to see the illness through the patient's eyes. The term is often contrasted with doctor- or disease-centeredness. In the latter approach, the accent is on the physician ascertaining the patient's complaint(s) and seeking information which will enable him to interpret the patient's illness in terms of his own medical explanatory frame of reference (Levenstein et al., 1986).

Both patient- and physician-centered approaches to the consultation are not sufficient in themselves and should be combined as we have done in the model. Although the interviewing style differs in both approaches, providing a natural boundary between both, their order is not fixed and may be reversed. In the third phase, "presenting solutions", the negotiation concerning problem definition and proposals for preventive and curative measures is a means to combine both approaches.

This combination of patient- and physician-centeredness is indispensable for primary care interviews in which the complaints are undifferentiated, i.e. somatic complaints, mixed with psychosocial and psychiatric problems often resulting from stressful life circumstances or otherwise. In addition to an adequate assessment of the somatic aspects of the complaints, the physician receives an impression of the significance of the complaints and of the medical encounter. When the patient is allowed to express all the reasons for attendance, the physician is better able to tailor an appropriate response. This may prevent unfavorable patient careers, characterized by medicalization of living problems (Illich, 1976) or somatic fixation (Grol et al., 1981).

2.2 From medical interview to interviewing skills.

In the following section, we further elaborate on the educational model by defining more concretely the content elements and the process skills needed for the physician to perform a medical interview. It is, moreover, a further justification of our educational model based on data from the literature.

First we describe the process skills, which, in principle, can be used in every phase (2.3). They are called "interpersonal and communicative skills". We then describe the different phases of initial medical interviews, their content elements and the process skills that are characteristic of each of these phases.

2.3 Process skills of initial medical interviewing: interpersonal and communicative skills.

The skills discussed in this section contribute to the process of the initial medical interview, irrespective of its phase. On reviewing the extensive literature, the reader easily falls into despair: every author uses (when he does at all) his own theoretical framework and definitions of skills, thus making results from empirical studies barely generalizable or comparable. We have therefore made a rather arbitrary selection of which skills to include in our educational model. We have built the systemacy of these skills on two pillars. First, we discuss the important distinction between interpersonal and communicative skills. Second, we turn to the comprehensive taxonomy of Ivey (1983).

The distinction between interpersonal and communicative skills, made by Hess (1969), is based on a more universal distinction, characterising every human performance (art, craft, etc.). An affective, expressive and an instrumental, task-orientated component can always be discerned (Ben Sira, 1976, 1980; Philipsen, 1979).

Interpersonal skills refer to behavior patterns which aid in the establishment of patient rapport, trust and acceptance. They are important in enhancing patient satisfaction and compliance (Korsch et al., 1968; Hulka et al., 1976).

Communicative skills are to promote the information flow between physician and patient: the skills to structure the interview in segments (e.g. opening gambits, data collection, prescription and explanation of therapy and closing), the use of appropriate questioning and of effective techniques to convey information.

The second guideline in our systemacy of these interviewing skills are the microskills defined by Ivey (1983). It is a taxonomy of interviewing skills which are, according to Ivey, common to various mode of task-orientated interviewing, such as medical interview, the Rogerian encounter, business problem-solving, correctional interrogation etc. Generalizability to every kind of intentional interviewing is claimed from this hierarchy of skills.

The importance of Ivey's microskills for our study lies in his clear definitions of the interviewing skills to which we often adhere.

Another argument for the use of the microskills taxonomy is Ivey's claim (1983) that this approach can be taught effectively, witnessed by 150 data-based evaluation studies (Kasdorf et al., 1978).

Our systemacy for the description of the process skills of initial medical interviewing is as follows:

- The interpersonal (2.3.1) and communicative (2.3.2) skills, defined according to Hess (1969), are the two major distinctive classes of process skills.
- Some complex skills, such as confrontation and interpretation, can not easily be classified under the headings of interpersonal or communicative skills. They are discussed separately in 2.3.3.

An overview of the process skills of initial medical interviewing is given in table 2.1.

2.3.1 Interpersonal skills.

By means of these skills, the physician should establish rapport, trust and acceptance in the patient. Hess (1969) does not provide further theoretical insight on how to connect interviewing skills with establishing rapport, trust and acceptance.

Study of the empirical literature reveals that several interviewing skills increase rapport with trust and acceptance in the patient. They are described in the following sections (2.3.1.1 to 2.3.1.6).

2.3.1.1 Non-verbal interviewing skills. Example: active listening.

Non-verbal skills in the physician's (interviewing) behavior have been attributed much importance (Friedman, 1979, 1982). Nevertheless, categorizations and meaningful behavioral descriptions of non-verbal interviewing behavior are scarce. Mehrabian (1972) discerns 3 major areas of non-verbal communication: 1) immediacy: the degree of "closeness" between two persons engaged in an interaction; 2) relaxation: the degree of postural relaxation-tension exhibited by the communicator; 3) responsiveness: the extent of awareness of and reaction to another person.

However, the validity of such distinctions has barely been established (Larsen et al., 1981). Further, "non-verbal cues are only meaningful and interpretable within a situational context", as DiMatteo and DiNicola (1982) contend. They interact with verbal cues and serve as accentuation or addition to verbal messages. A well-known example of non-verbal interviewing skills is "active listening", denoting the physician's openness to important (often patient-centered) information. Moreover, active listening is claimed to convey acceptance and respect to the patient. It further indicates the physician's willingness to share power and control with the patient in the interview (Stone, 1979).

This example shows the intricacies in the operationalization of non-verbal behavior. Active listening is an observation skill of the patient's (non-)verbal behavior as well as a physician's behavior that radiates acceptance and respect.

In summation, we state that great importance has always been attributed to non-verbal behavior, but the effect of well-defined non-verbal behavior on the patient has not yet been established. Empirical data from this complex area are scarce and controversial to date. In any case non-verbal behavior is difficult to operationalize in a teaching program of medical interviewing skills.

2.3.1.2 Facilitative skills.

Closely related to active listening and the physician's non-verbal behavior are the facilitative skills which, according to Baekeland and Lundwall (1975), have the purpose of ensuring that the patient is given

Table 2.1: Overview of the process skills of initial medical interviewing.

2.3.1 INTERPERSONAL SKILLS

2.3.1.1 Non-verbal interviewing skills.
Important example: active listening.
Dimensions:
- immediacy
- relaxation
- responsiveness

2.3.1.2 Facilitation.
- (non) verbal behavior to encourage the patient to talk

2.3.1.3 Reflections.
- make implicit feelings explicit
- make the patient aware of the process of the interview

2.3.1.4 Empathy.
- active listening and understanding
- summarization, reflection of feelings and of meaning

2.3.1.5 Reassurance.
- explicit agreement with emotionally loaded statements
- realistic "prediction" of future events

2.3.1.6 Self-disclosure.
- I-statements in present tense
- own "genuine" experiences close to those of the patient

2.3.2 COMMUNICATIVE SKILLS

2.3.2.1 Appropriate questioning.
- open questioning
- closed questioning
- probing or directive questioning

2.3.2.2 Effective conveyance of information.

cognitive aspects

- simplification
- repetition of important data
- explicit categorization
- specificity of advice
- summarization with feedback function

emotional aspects

- bad news first
- personal approach
- discussion of consequences for daily life
- management of defense mechanisms

2.3.2.3 Structuring the interview by

- agenda
- introducing and terminating the phases of the interview
- announcing and closing important topics

2.3.3 COMPLEX INTERVIEWING SKILLS
(intervention skills)

2.3.3.1 Confrontation

- identifying and working through mixed messages and incongruities

2.3.3.2 Interpretation

- provision of the patient with an alternative frame of reference to reconsider life situations

the opportunity to express himself in the relationship: not only to tell his own version of the history but also to display his own emotions.

Facilitative skills, called "encouragers" by Ivey (1983), are a verbal or non-verbal means which the physician can use to encourage the patient to continue talking, such as nods of the head, "uh-huh"-statements and the simple repetition of keywords the patient has uttered.

"Facilitative skills" is, in a way, a misleading term because much interviewing behavior may have facilitative properties, such as open questions, self-disclosure by the physician, summarizing and directive questions.

2.3.1.3 Reflection.

Reflections of feelings have the purpose of making (partially) implicit feelings underlying the patient's words and behavior explicit and clear to the patient (Ivey, 1983).

Reflections on the process of the interviewing are a kind of meta-communication by the physician, commenting on the course of the interview. In particular, when the flow of communication between physician and patient is hampered by strong defense mechanisms or antagonistic feelings by the patient, it may be helpful to discuss these underlying problems in order to restore communication.

2.3.1.4 Empathy.

This complex skill is part of the Rogerian "trias" empathy, unconditional warmth and positive regard; genuineness (Rogers, 1951). Empathy is the capacity to understand (and, perhaps, to feel) the patient's experiences, needs, sorrow, joy, anxiety, etc., as if they were one's own feelings.

Empathy consists of two types of skills: on the one hand, the ability to listen actively and understand the patient's non-verbal cues of emotion. On the other hand, more verbal skills such as reflections of feelings, summarizing and, what is called by Ivey (1983) "reflection of meaning", are components of empathy.

2.3.1.5 Reassurance.

Psychological factors, such as anxiety, fear and distress, have a profound impact on the patient's physical state (a.o. Pelletier, 1979) and on the outcome of medical treatment, such as surgery (Langer, et al., 1975). In such cases, reassurance may be indicated. It is a generally optimistic and hopeful attitude expressed in specific statements based on data and/or experience designed to allay any exaggerated or unfounded fears of the patient (Leigh et al., 1980). For reassurance to be effective, the physician should know the sources of the patient's fears. The physician, however, should not pacify realistic fears.

Two skills are recommended when used with discretion (Brammer, 1973). First, explicit agreement with emotionally-loaded statements of the patient may be supportive. Second, prediction of future experiences, which should be based on facts and which might be highly probable, can be effective.

Although the definition of reassurance is clear, operationalization on the skills level (by Brammer) remains rather vague.

2.3.1.6 Self-disclosure by the physician.

Self-disclosure is the act of revealing personal information to others (Jourard et al., 1970). It is considered to be a key-element in the formation of trust in patients (Johnson et al., 1975), whereas it may also help to break down some of the interactional barriers between physician and patient (DiMatteo et al., 1982). Self-disclosure by the physician may help the patient to understand the normality of his behavior and to establish a basis of similarity between both partners. In this sense, it is effective when errors in treatment are detected because trying to hide obvious failures will once again raise the barriers to communication.

Concrete guidelines for self-disclosure are given by Ivey (1983): 1) the expression of feelings in I-statements and in the present tense; 2) the experiences disclosed by the physician should be genuine and close to the patient's experience.

2.3.2 Communicative skills.

Communicative skills are necessary for the promotion of the information flow between physicians and patients (Hess, 1969) as well as for the structuring of the interview in segments. We start the next sections with the "promotion of the information flow" and turn then to "structuring the interview".

The exchange of information pertains to factual information as well as to emotions related to these medical facts. The quality of this information exchange is expressed by means of the "theory of effective communication" (Schouten, 1982; citing Fraser, 1976). The communication between two persons is effective when both partners are mutually aware of the meaning that the one attaches to the exchanged messages of the other.

This theory, expressing the mutuality of physician-patient communication, is as important regarding the conveyance of information to the patient as it is for the collection of information from the patient.

In line with the definition of effective communication we define a group of skills that pertain to the control of whether both parties in the interview attach the same meaning to the exchanged messages: appropriate methods of questioning (2.3.2.1) and effective conveyance of information (2.3.2.2).

The skills to structure the interview, such as opening and terminating the interview, introduction and closure of its phases, are ranged by Hess (1969) under "communicative skills". They are discussed in 2.3.2.3.

2.3.2.1 Appropriate questioning.

In this section, we discuss the proper use of open, closed and probing questions:

a) Open questions have no bias against the patient and cannot be answered in a few short words (Ivey, 1983). The patient should feel free to bring up the topic which seems to him the most appropriate answer to the question. Open questions are helpful in facilitating the patient to narrate his experiences, emotions and fears according to his own frame of reference.

b) Closed questions have the advantage of focusing the interview and obtaining specific information (Ivey, 1983). They lead to yes-or-no answers. The proper way of asking closed questions is:
   - not suggestive
   - referring to one subject matter
   - presenting the subject matter clearly in order to provide a yes-or-no alternative for the patient to answer

   Closed questioning is indicated when:
   - the physician needs factual information, useful to the diagnostic and problem-solving process
   - the patient is vague or defensive in discussing a certain topic or deviates from this subject

   They are contra-indicated when restricting answer categories leads to the danger of missing relevant information. In addition, closed questioning is often inappropriate when the physician wishes to explore the patient's frame of reference; in this case, open questions are indicated.

   Although the definitions of open- and closed question look rather sharp, in practice, the distinction may sometimes be difficult. Sometimes, open questions refer to a restricted content domain, leaving few response-possibilities open to the patient whereas, for example, closed questions may hit the nail right on the head and invite the patient to provide a lot of information from his own frame of reference.

c) Probing or directive questions have a position between open- and closed questions. Often used in response to an answer to an open question, the physician focuses on a specific element in this answer and attempts to make this topic more specific, concrete or personal. The well-indicated use of probing questions, especially during "the exploration of the reason for encounter" and "the history-taking" phases, is summarized in the term "concretization". The result of well-indicated concretization(s) leads to a more personal, concrete and specific discussion of the important topics in the interview.

2.3.2.2 Effective conveyance of information.

In the conveyance of information, cognitive and emotional aspects are to be distinguished (Schouten et al., 1982).

a. Cognitive aspects of conveying information.

Several methods have been evolved for the presentation of information to patients in a way which will enhance the probability of its recall. Ley (a.o. 1983) summarizes his own and others' research in the following recommendations by which means the understanding and recall of information provided to the patient might be increased:

- simplification (simple words, short sentences, no jargon)
- repetition of important data
- explicit categorization (segmentation of information into small clusters and explicity notifying these clusters)
- the use of specific rather than general advice statements

In addition to these recommendations, recall and understanding of conveyed information may also be influenced by two psychological factors: the "primacy" and "importancy" effects. The former means that patients best recall that information which is provided first. By "importancy effect" is indicated that statements with a clear significance for the patient are better recalled than other information.

Criticism of Ley's recommendations is given by Schouten et al. (1982), who advise "explorative interventions" after providing important pieces of information in order to avoid "one-way-traffic". In this way, the physician should get feedback from the patient on whether he has understood the conveyed information. If the patient forgets or does not understand something, the physician may add, clarify or modify information.

Summarizing is, in this respect, an appropriate interviewing skill. Summaries feed back to the patient the essence of what has just been said by shortening and clarifying his statements. Besides providing more insight, it may increase in the patient the feeling of being understood and accepted.

The important distinction between the patient's ability to recall information and its comprehension is made by Tuckett (1985).

He shows that patients are able to recall the majority of information conveyed to them but that they do not understand a considerable amount of it. He interprets this finding by claiming that the physician does not attune his information during this phase to the explanatory model of the illness/condition as developed by the patient and as asked for by the physician during the phase "exploring reasons for encounter". Tuckett recommends that the physician should communicate his understanding of the illness/condition and encourage the patient to communicate his "explanatory model". Opportunity must be given to explore questions and doubts on either side and to check and clarify what is being meant and understood. In this sense, a kind of negotiation about the problem definition is taking place (see 2.3.2.3).

b. Emotional aspects of conveying information.

"Bad news" may provoke defense mechanisms in patients that are insufficiently helpful in adapting to the new situation. For acceptance of the new situation, it is necessary that the patient is aware that the news is true and pertains to him. In this sense, the recommendations of Tuckett et al. (1984) that the consequences which the illness/condition has for the daily life of the patient should be discussed, bringing about acceptance. The same holds true for the consequences of treatment and preventive measures. Their acceptance is necessary for compliance: the willingness to cooperate with preventive and therapeutic regimes.

The importance of acceptance of "bad news" is also witnessed by the general recommendation (a.o. Schouten et al., 1982) that the interview should be started with the conveyance of the bad news. The physician may then use much of the available time of the interview for working through the bad news and the fostering of its acceptance. In this respect, Schouten et al. (1982) give guidelines on how to handle defensive reactions of the patient. To a denial, the physician may respond with an emotional reflection, such as "I see you cannot believe this news". To anger, he may react with an empathic understanding of the resistance to the bad news. To states of confusion, the physician may clarify the situation and foster (cognitive) insight which may enhance adaptation to the new situation.

2.3.2.3 Structuring the interview.

The physician should give structure or order to the medical interview (Schouten et al., 1982). Unstructured interviews are often witnessed in student or inexperienced physicians who pay too much attention to the first complaint presented by the patient, whereas other, and even more important, complaints may be put forward later in the interview. A danger of paying too much attention to less important topics may ensue.

Initial medical interviews may be structured on three levels: structuring by means of an agenda, introducing and closing the phases and announcing and closing new topics.

The whole interview may be structured by means of an agenda. It may run as follows. After opening the interview with clarification of roles (if necessary), the physician invites the patient to state why he is visiting the physician. After the initial presentation by the patient of the purposes of the visit (complaints, problems, other requests for help), the physician may draft an agenda in order to make clear to the patient which topics will be discussed and what he expects of the patient. This agenda may provide the advantage of the physician being able to ask at the closure of the medical interview, whether it has met the expectations of the patient in the sense that the topics relevant to the patient have been discussed in their technical-medical as well as in their emotional aspects.

Completing the three phases, "exploration of the reason for encounter", "history-taking" and "presenting solutions", is the most important issue on the physician's agenda. Although the literature provides no clear preference (a.o. Levenstein et al., 1986), in our educational model, "exploration of the reason for encounter" should preceed "history-taking" in general practice as well as in primary mental health care. It is important to attune history-taking to data collected during the exploration of the reasons for encounter. This argument is especially valid for primary mental health care, but for general practice, the request for help and the patient's causal attributions may have great diagnostic value.

Even more important, however, is the introduction and closing of the phases. For example, it is useful to close the phase "exploration of

the reason for encounter" with a summary. The physician shows his understanding in this way and allows the patient to amend or to correct this summary.

Finally, the physician should announce the subject to be discussed before starting to ask questions. Such a statement should help the physician to concentrate on the subject himself and it also helps the patient to stick to the subject and not to expand to another. Concluding a subject which has been sufficiently discussed should be done by summarizing. The physician checks whether he has attached the proper meaning to the words of the patient. A summary is submitted to the patient for judgment, correction and/or completion. Several authors (a.o. Van Dorp, 1977; Ivey, 1983; Holten-Vriesema et al., 1978; Goldberg, 1979) recommend such structuring procedures.

Over-structuring may lead to a tight interview with little opportunity for the patient to put forward his own thoughts or ideas: in other words, the patient's frame of reference may not emerge sufficiently in the interview (Schouten et al., 1982).

Another restrictive observation about structuring concerns the fact that, empirically, its contribution to the positive outcome of the medical interview has not yet clearly been established.

2.3.3 Complex interviewing skills.

The theoretical dividing line between interpersonal and communicative skills becomes more difficult to draw in complex skills where emotional and instrumental elements are closely united. In this "residual" category, we discuss the skills of confrontation (2.3.3.1) and interpretation (2.3.3.2).

2.3.3.1 Confrontation.

Confrontation means pointing out incongruities, discrepancies or mixed messages in behavior, thoughts, feelings or meaning (Ivey, 1983). It involves two major steps: identifying mixed messages and incongruities (between statements, between words and non-verbal behavior, between statements and context) and working toward their resolution (through exploration by means of skills of reflection, summarization, interpretations). The purpose is to promote the solution

of psychological problems connected with these incongruities or discrepancies through increased insight.

2.3.3.2 Interpretation.

Interpretation provides the patient with an alternative frame of reference to reconsider living situations (Ivey, 1983). The purpose of this skill is similar to the skill previously described: resolving psychological problems by increasing insight.

In most instances, this skill consists first in reflections (of feelings) or summaries of the patient's statements to clarify them and then in the presentation of the meaning the physician attaches to them.

It is evident that both skills represent a borderline area between the skills pertinent to initial medical interviews and more "problem-solving", "counseling" or "psychotherapeutic" follow-up interviews.

2.4 Skills specific for the three phases of the initial medical interview.

In the following paragraphs where the phases "exploration of the reason for encounter", "history-taking" and "presenting solutions" are discussed, we first pay attention to the content and then to the process skills.

2.4.1 Exploration of the reasons for encounter.

2.4.1.1 Content.

The information collected during this phase is concentrated around two aspects: patient-centered information pertaining to medical complaints and so-called non-biomedical reasons for the visit (Barsky, 1981).

Patient-centered information was first "discovered" as content relevant to the medical interview by Korsch and co-workers (1968, 1972). When, according to them, the physician neglects information concerning the patient's worries, anxieties and expectations, he causes dissatisfaction entailing non-compliance from the side of the patient.

As important aspects of patient-centered information, are also considered:

- Causal attributions or subjective, personal explanations of complaints.
  These explanations are based on the human need to make sense of the environment and to search for causes of experiences.
  Important in understanding the patient's attributions is the general finding, described by Kelley et al. (1980), that persons attribute unfavorable events to the environment rather than to themselves whereas, in cases of favorable events, the reverse occurs.
  Attributions that patients have about their complaints shape their expectations concerning help in a technical as well as in an emotional sense (Stoeckle et al., 1980). The attributions are also strongly connected with their beliefs about the prognosis of the illness/condition.
  Correction of misattributions results in improved medical outcomes, such as in post-operative pain (Egbert et al., 1964), after child tonsillectomy (Skipper et al., 1968) and in insomnia (Storms et al., 1970).
- Triggers to the decision of the patient to seek help, such as interpersonal crisis, interference of symptoms with social, personal and vocational functioning etc. are important patient-centered information (Zola, 1973). These data give the physician insight into why the patient is seeking medical help for a certain complaint at that very moment (Holten-Vriesema et al., 1978).
- Data concerning the patient's living circumstances give insight into the impact of the complaint(s) on the daily life of the patient. Moreover, they show how the patient has dealt with the complaint(s) in his primary group and how the decision to seek help has been reached.
- The coping of the patient with his complaint(s) has to be explored. How has the patient attempted to gain relief? What solutions have proved successful in the past? etc. These aspects are barely discussed in the literature.

Non-biomedical reasons for why patients consult physicians have been reviewed by Barsky (1981), who assigns them 4 categories.

First, he pays attention to "minor psychiatric disorders". In epidemiological studies, it is found that the incidence (14% of the British population in the course of one year; Shepherd et al., 1966) and the prevalence of these "minor psychiatric syndromes" is high (20-25%) (Goldberg et al., 1980; Shepherd et al., 1982). These "minor psychiatric syndromes" encompass mainly depressive, anxiety-related and psychosomatic symptomatology, often in combination with somatic disorders. Although spontaneous remission is also high, 10-15% of these "minor cases" seem to become more seriously or chronically disabling. Johnstone et al. (1976) have shown that early detection and treatment of "minor cases" results in relief of symptoms and psychological pain without "medicalizing" the illness and with probable cost benefits. However, in other studies (Goldberg et al., 1982; Goldberg et al., 1970), it is shown that about 30% of these cases of patients with "minor psychiatric syndromes" are not detected. In this phase of the medical interview, the physician may be alerted to the fact that some (minor) psychiatric disorder may be present.

Secondly, life stress and emotional distress, caused by major living changes, life events and socio-environmental factors, often lead to adopting the "sick role". Persons often cope in this way with stress situations, which provoke anxiety, frustration and grief but which do not lead to diagnosable psychiatric disorders. This psychological distress causes people to perceive their health status less favorably, so they consider themselves as ill (Tessler et al., 1978).

Thirdly, social isolation may be a motive for a medical encounter (McKinlay, 1980). Persons lacking adequate inter-personal relationships easily turn to physicians for advice, for a feeling of being cared for, and for an opportunity to express their feelings (Balint, 1957).

Fourthly, people distressed by physical symptoms visit physicians more for information (education, explanation or reassurance), than for treatment (Cartwright, 1964).

However, the most important topic drawn from this patient-centered information is the way the patient desires to be helped to attain his needs or goals (Lazare et al., 1975).

The answer to this question is the main purpose of this phase. It is the backbone of the model of the "consumer's approach to patienthood", extensively described in 2.1.2.1. Many of the afore-mentioned topics contribute to this issue.

2.4.1.2 Process skills.

Schouten et al. (1982) operationalizes the patient-centered style of interviewing by means of the following skills: open questioning, summarizing, reflections, active listening and concretizing (see 2.3).

2.4.2 History-taking.

History-taking and clinical reasoning are strongly interwoven: the first is the verbal communicative and the second is the cognitive dimension of the same process.

Metz (1984), summarizing the state of the art, described clinical reasoning as a hypothetico-deductive strategy. It is a three-step mental process, repeated many times throughout the interviews: 1) collecting data (questioning, observation, clinical methods); 2) generating, confirming and refuting hypotheses on the basis of the data collected; 3) employment of clinical methods or strategies to elicit additional data that will generate new hypotheses and confirm or refute old ones.

These clinical reasoning strategies largely determine the content of the medical interview during the history-taking phase. We first consider the content aspects for general practice (2.4.2.1) and then the content aspects for mental health problems (2.4.2.2). Finally, the process skills are given attention (2.4.2.3).

2.4.2.1 The content dimension concerning general practice.

For general medicine, textbooks, like those of Morgan and Engel (1969) in de US or Formijne (1982, 10th ed.) in the Netherlands, describe the content of the medical interview. Morgan and Engel use 5 categories, of which the content is self-evident: present illness, past health, family health, personal and social history, systems review.

As a more detailed illustration of such content descriptions, we show in table 2.2 an overview from the textbook by Formijne o.a. (1982, 10th ed.).

Students have been traditionally taught to collect their information according to this "schedule". Emphasis was on thoroughness and completeness of the information; no item should be overlooked. After the gathering of all these data and after completion of the physical

examination, the student should generate hypotheses about the diagnosis of the complaint and its aetiology (Cutler, 1979).
It is evident that such a "schedule" of the medical interview has a strong impact on the interviewing style: a tightly structured, doctor-centered interview ensues.
This procedure has been critized by several investigators (Elstein et al., 1978; Kassirer et al., 1978) who stated that the problem-solving strategy of the physician during the medical consultation did not bear

Table 2.2 Content overview of initial medical history-taking (adapted from Formijne, 1982).

- personal data (name, place of residence, sex, age, marital state, profession, etc.)
- medical history-taking concerning the complaint(s)/symptoms that are the reason for the medical consultation: the main complaint(s) (local and/or general complaints), their onset, course, intensity, character, localization, irradiation, provoking and releasing factors, accompanying symptoms.
- review of systems
  Systematic checking of the state of the most important organ systems: organic symptoms in the cardiac, circulational, pulmonal, gastric, intestinal, liver and biliary system, kidneys and urinal tracts, bones, joints and spine, blood, endocrine system, genital functions, sensory organs, nervous system, psychological status, social circumstances.
- supplementary data:
  - diseases and illnesses in the past
  - intoxication and substance abuse
  - nutrition
  - heriditary diseases

any resemblance to this linear procedure of checking a standardized list of symptoms and signs.
Therefore, medical teachers now recommend to students the "natural pathway" of interviewing, following the "systemacy" of the clinical reasoning process. Finally, the physician may pose some "screening" questions, especially from the system from which the main complaint originates.

2.4.2.2 The content dimensions concerning primary mental health care.

In this section we describe the content dimension of history-taking in primary mental health care. Within this content dimension, two levels are distinguished:

- "psychiatric examination", which is history-taking on the level of complaint(s) and symptoms (2.4.2.2.1)
- "socio-emotional exploration", which is history-taking on the level of "aetiological conditions": the functional and (often circular) causal relationships with psychosocial conditions (2.4.2.2.2)

Finally, we address the choice of the most appropriate interviewing styles in this phase (2.4.2.2.3).

2.4.2.2.1 Psychiatric examination.

The content dimension of the complaint and symptom level is given by the most elaborated psychiatric classification system, the Diagnostic and Statistical Manual for Classsification of Mental Disorders (DSM-III; 1980).

Nevertheless, some disadvantages of using this system as the content base for medical interviewing should be mentioned. It does not suit the minor psychiatric disorders, common in primary mental health care, because these problems are characterized by complaints mixed with a variety of "problems-of-living". However, axis 4 (psychosocial stressors) and axis 5 (social adaptation) of the DSM-III which are particularly important for the "minor psychiatric syndromes", are not very well elaborated. DSM-III also does not take into account the diagnostic value of the patient's request for help.
Finally, it does not pay sufficient attention to the meaning of psychopathology within family and social systems, which is of paramount importance in social psychiatry.

2.4.2.2.2 Socio-emotional exploration.

For the content dimension pertaining to "aetiological conditions", we refer to a resource which is specially geared to primary mental health care: a classification system for social and psychological problems, proposed by Regier et al. (1982). This is a composite classification system based on Reason for Visit Classification (RVC), CHPPC and DSM-III (Lipkin et al., 1982). A short overview of the items

in this classification is given in table 2.3. This system encompasses the psychiatric examination and the systematic exploration of psychosocial problems.

2.4.2.2.3 Interviewing style during history-taking.

In this section, we discuss the issue of how to collect the relevant information for diagnostics and clinical problem-solving. In the literature, the discussion of interviewing styles of history-taking moves on a continuum between two extreme positions.

On one side, the extreme is an interviewing style according to a structured schedule, such as the Present State Examination (PSE; Wing et al., 1974) or the Diagnostic Interview Schedule (DIS; Robins et al., 1979), which are used for research purposes. These comprehensive schedules entail systematic, screening questioning resulting in diagnostic classifications which describe the clinical picture and the personality structure of the patient.

At the other extreme, is a non-directive, patient-centered style that is sometimes found in time-restricted interviews in general practices. This non-directive style, in which the patient takes the lead, is often criticized by (clinical) psychiatrists. They assert that important information might be missed because it is not spontaneously brought up by the patient. When the questioning is more systematic, the interview yields a more comprehensive coverage of psychiatric pathology (Saghir, 1971). Therefore, in psychiatry, some systematic questioning called "psychiatric examination" is a common procedure (a.o. Kuiper, 1981, 9th ed.).

However, disadvantages of a structured interviewing style with directive and systematic questioning should also be mentioned.
First, important idiosyncratic, biographical data (e.g. life events) might be missed because they are not covered by this way of questioning. Second, the patient's request (Lazare et al., 1975) is not taken into account as a relevant diagnostic. Third, directive interviewing does not allow a "natural" physician-patient relationship which is characteristic of the communicative patterns in which the patient usually engages. In such natural dialogues, idiosyncratic

Table 2.3: Synthesis of systems used in primary mental health care to classify social and psychological problems (Regier et al., 1982).

Social problems

1. housing
2. change in residence
3. financial
4. family (conjugal, parent-child, family disruption, caring for sick persons, other problems of family relationships)
5. non-family interpersonal problems
6. education or learning
7. occupation (incl. house keeping)
8. legal problems
9. personal and environmental circumstances, that impede access to health care or are hazardous to physical health
10. conflict with the practices or belief system of a social or cultural institution
11. other social problems

Psychological problems

1. feeling anxious or nervous
2. feeling depressed
3. disturbances of sleep
4. sexual problems
5. eating problems
6. ideas of suicide
7. feeling angry or irritable
8. psychomotor restlessness
9. trouble with concentration or memory
10. problems with identity
11. social withdrawal
12. disturbing personality traits
13. substance-related problems
14. delusions, hallucinations or incoherence
15. phase-of-life problems
16. age-specific developmental problems (excludes learning problems)
17. "psychophysiological" problems
18. other psychological problems

communication styles may originate which are of great diagnostic value (e.g. transference). Fourth, although systematic questioning with PSE, DIS and the like provides a reliable diagnostic classification, its predictive validity, according to treatment choices or decisions, is not high.

Nevertheless, a majority of authors seems to favor a rather structured interviewing style; for instance, along the guidelines of diagnostic decision-trees (Morgan et al., 1980; Giel, 1981). Cox, et al. (1981) also propose a more directive style with specific probes and requests for detailed descriptions and even systematic questioning. In their experimental studies, comparing several interviewing styles, they found that the advantages of systematic questioning for obtaining factual information were not associated with any disadvantage with respect to the eliciting of emotions and feelings.

We conclude this discussion of the pros and cons of a structured interviewing style by taking the "mid-stream position" which better suits interviewing on the primary mental health care level.

First, the basic interviewing style during this phase is non-directive and patient-centered in which the physician explores issues brought up spontaneously by the patient.

Second, when the physician encounters a salient cue in the patient's story which might have diagnostic value, he then becomes more directive. By means of probing questions, he might test one or more hypotheses about the problem in the framework of his clinical reasoning.

Finally, he might end with some systematic "screening" questioning. In this respect, Goldberg et al. (1980) propose a simple scheme for the assessment of current psychological adjustments in primary mental health care, such as affective, anxiety-related and psychosomatic disorders.

2.4.2.3 Process skills.

Three categories of interviewing skills suit the afore-mentioned styles of history-taking in both general practice and in primary mental health care.

a) Questioning.

It is advised that the questions posed should be as open as possible to avoid suggestive or hidden questions (question-sounding statements that do not demand immediate responses). However,some authors (a.o. Goldberg et al., 1980) claim that the appropriate skill during this phase - in its typical form - is the so-called

open-to-closed-cone questioning: some important cue is raised during the interview; next, the physician proceeds with more probing and directive questions and may end with closed questions to acquire accurate, factual information. This open-to-closed-cone questioning may suit the underlying hypothetico-deductive strategy of the clinical reasoning.

b) Active, unbiased listening.
"Unbiased" listening is a high ideal for physicians who have to select the medical facts indispensable for a medical diagnosis. Nevertheless, the physician should be continuously aware of possible biases.

c) Summarizing.
The physician summarizes the words of the patient in his own words and invites the patient to check his summary. This is a control against "biased" listening.

2.4.3 Presenting solutions.

After data collection (often including physical examination and some simple clinical methods in somatic problems), patients expect conclusive statements about the nature of the presented complaints, their diagnoses, aetiology and prognosis. Of course, advice on how to deal further with their problem is also desired.

The term "presenting solutions" may be misleading, however. Taken as an extreme, it might suggest a one-sided situation in which the physician presents conclusive solutions to a passive, rather ignorant and incapable patient. Fortunately, a negotiation process with the patient generally originates on problem definition and about the proposed preventive and curative measures (Stimson and Webb, 1975; Lazare et al., 1975).

This negotiated consensus model does not fully apply to interviews with patients suffering from severe mental disorders, with disturbed perception or thought patterns and therefore not possessing the necessary interpersonal skills to negotiate.

In the following sections, we discuss the content aspects of this phase (2.4.3.1) and the interviewing skills necessary for negotiation (2.4.3.2). In 2.3.2.2, we have already discussed the skills needed for the conveyance of information.

2.4.3.1 Content.

The content of the conveyed information is a direct consequence of the "negotiated consensus" consultation model. Within the negotiation process, the following "key-points of conveyed information", as Tuckett (1985) calls them, are necessary:

- The "diagnostic-significance" of the problem (what caused it, whether it is life-threatening, progressively deteriorating, likely to recur or self-limiting). Like Katon et al. (1980), we would add to this the information about the "biomedical" model of the illness condition in layman's terms: a problem-definition that is understandable for the patient.
- The "appropriate treatment-action" to deal with a problem (which treatment or other actions are recommended, their value and purpose and how to carry them out). Also included in this item are: referral or no treatment-action at all.
- "Appropriate preventive measures" which may be necessary to forestall or lessen future illness episodes (which preventive measures, their value and purpose, how they relate to the original cause of the problem and how to carry them out).
- "Implications or wider social and emotional consequences" of problems or their treatment (which problems may be experienced and how the patient can be helped to cope). The most important issue in this respect is the prognosis of the problem. The patient's anxiety and expectations should be confronted with the physician's information in order to decrease unrealistic fears.
  Attention should also be paid to expected side-effects of drugs because inappropriate interruptions of or halts in medication are frequently encountered with harmless or even transient side-effects.

These four key-points of provided information also hold for initial psychiatric interviews in primary care.

2.4.3.2 Process skills.

Negotiation is a central issue in "presenting solutions". In 2.1.2.3 we reframed it as a combination of the patient and physician-centered approaches of the problem presented in the consultation process.

Lazare et al. (1975) consider negotiation as the "backbone" of the "negotiated consensus" model of medical consultation. During the consultation, more explicit differences between physician and patient, such as the definition of the problem, the cause of the illness, the goals and priorities of treatment and the methods of treatment or preventive measures, may entail negotiation. "How to negotiate with the patient" is a subject scarcely described in the literature. In contrast with the relatively rich literature on the phenomenon of negotiation in general, the necessary skills are discussed in rather broad terms by a few authors.

Katon et al. (1980) give the most concrete model of negotiation, prescribing the following steps for the physician:

- elicit the patient's explanatory model and illness problems
- present to the patient the medical explanatory model of the disorder, including the treatment recommendations in layman's terms
- when discrepancies in the physician's and patient's expectations of treatment remain: acknowledgement and clarification of the conflict and provision of the patient with the opportunity to present alternatives
- when the conflict cannot be resolved: decide on an acceptable compromise for treatment, based on biomedical knowledge and on the patient's explanatory model (within medical ethical standards)
- when, nevertheless, a stale-mate remains, then the therapeutic alliance should be broken and referral to another physician should be offered
- finally, the negotiation must involve ongoing monitoring of the agreement and of each party's participation.

In this model of negotiation, it is evident that the phase "exploring the reasons for encounter" is a necessary pre-condition. The patient's request for help and his explanatory model of the illness/condition should have been discussed during this phase.

2.5 Concluding remarks.

An educational model showing the ideal course of an initial interview in general practice and in primary mental health care is presented. It is characterized by three phases:

- Exploration of the reason for encounter.
  In this patient-centered phase, the physician and the patient explore the concerns, expectations and causal attributions engendered by the complaint. Further important data on the patient's living circumstances, on events precipitating the visit and on habitual styles of coping are collected.
  In this phase, minor psychiatric disorders or other emotional stressors might also become manifest. This phase is completed when the physician has a view of the way the patient could be helped in fulfilling his needs or goals.
  The interviewing style of this phase is characterized by active listening, open facilitative questioning, reflections, summarizing and concretizing.
- History-taking.
  This physician-centered phase is strongly interwoven with diagnostics and clinical problem-solving. In our educational model, the history-taking phase in general practice corresponds to three distinguishable phases in primary mental health care: history-taking (in sensu strictu), psychiatric examination and socio-emotional exploration.
  Psychiatric examination deals with the description of complaints and symptoms. Socio-emotional exploration pertains to the aetiological, psychosocial conditions of the presented problem.
  The interviewing style in this phase is characterized by an open-to-closed questioning, suiting the underlying hypothetico-deductive strategy of the clinical reasoning. This style may be supplemented by periods of systematically structured questioning based on schedules covering frequently encountered symptoms and aetiological conditions.
- Presenting solutions.
  This phase consists of the conveyance of information (diagnostics, preventive and treatment advice, prognosis) and negotiation about

problem-definition and preventive and treatment advice. This negotiation entails a compromise as regards solution between physician and patient-centered contributions to the consultation.

This model is based on current consultation models, such as the "task" model of Pendleton et al. (1984) and "customers' approach to patienthood" of Lazare et al. (1975). Both models leave much room for patient-centered interviewing. A further feature of this educational model is its phasic character based on the "empirical cycle" (De Groot, 1961), which is, in itself, a general hallmark of all problem-solving activities.

The skills necessary for the process of interviewing are, according to Hess (1969), divided into interpersonal and communicative skills. Interpersonal skills related to the emotional and expressive aspects of the medical interview are intended to establish a rapport of trust and acceptance with the patient. Communicative skills are employed for the exchange of information between physician and patient, the task-orientated aspect of the interview. The skills to structure the interview are also categorized under the heading of communicative skills. In more complex interviewing skills such as confrontation, interpretation this theoretical distinction becomes fussy.

Both interpersonal and communicative skills have been operationalized in behavioral terms, often referring to the definitions of the so-called microskills (Ivey, 1983).

However, interpersonal skills, for a long time considered as the "art of medicine" and related to "personal talents" or "intuition" (DiMatteo et al., 1982), turn out to be difficult to operationalize. Behavioral definitions are provided for active listening, facilitation, reflection, empathy, reassurance and self-disclosure.

Communicative skills are intended to establish a situation in which both physician and patient are mutually aware of the meaning the one attaches to the exchanged messages of the other. The communicative skills involved in this effective communication are divided into questioning skills (open, closed and probing questions), and into skills in conveying information (in its cognitive as well as in its emotional aspects).

REFERENCES

Baekeland F, Lundwall L. Dropping out of treatment: a critical review. Psychological Bulletin, 1975; 82 (5): 738-783.

Balint M. The doctor, his patient and the illness. Tavistock. London, 1957.

Balint M, Hunt J, Joyce D, et al. Treatment or diagnosis: a study of repeat prescriptions in general practice. Toronto; J.B. Lippincott, 1970.

Barsky Aj. Hidden reasons some patients visit doctors. Annals of Internal Medicine, 1981; 94: 492-498.

Bennis WG, Benn KD, Chin R (Ed.). The planning of change. Readings in the applied behavioral sciences. New York, 1962.

Ben-Sira Z. The function of the professional's affective behavior in client satisfaction: a revised approach to social interaction theory. Journal of Health and Social Behavior, 1976; 17: 3-11.

Ben-Sira Z. Affective and instrumental components in the physician-patient relationship: an additional dimension of interaction theory. Journal of Health and Social Behavior, 1980; 21: 170-180.

Beugen M van. Sociale technologie. Van Gorcum, Assen/Amsterdam, 1977.

Brammer LM. The helping relationship. Process and skills. Prentice Hall Inc., Englewood Cliffs, 1973.

Byrne P, Long B. Doctors talking to patients. H.M.S.O. London, 1976.

Carroll JG, Monroe J. Teaching clinical interviewing in the health professions: a review of empirical research. Evaluation and the Health Professions, 1980; 3: 21-45.

Cartwright A. Human relations and hospital care. Routledge and Kegan Paul, London, 1964.

Cox A, Rutter M, Holbrook D. Psychiatric interviewing techniques V. Experimental study: eliciting factual information. British Journal of Psychiatry, 1981; 139: 29-37.

Cutler P. Problem-solving in clinical medicine. From data to diagnosis. Baltimore, The Williams and Wilkins Company, 1979.

DiMatteo MR, DiNicola DD. Achieving patient compliance: the psychology of the medical practitioner's role. New York, Pergamon Press, 1982.

Dorp C van. Luisteren naar patiënten; een analyse van het medisch interview. De Tijdstroom, Lochem, 1977.

Diagnostic and Statistical Manual of Mental Disorders, Third Edition (DSM-III). American Psychiatric Association, 1980.

Egbert LD, Battit BE, Welch CE, Bartlett MK. Reduction of postoperative pain by encouragement and instruction of patients. New England Journal of Medicine, 1964; 270: 825-827.

Elstein AS, Shulman LS, Sprafka SA. Medical problem-solving. An analysis of clinical reasoning. Harvard University Press, Cambridge Mass., 1978.

Giel R. Waarom een psychiatrische diagnose? Samsom/Stafleu, Alphen a/d Rijn/Brussel, 1985 (2e druk).

Formijne P, Mandema E. Leerboek der anamnese en fysische diagnostiek. Bohn, Scheltema, Holkema, Utrecht, 1982 (10e druk).

Fraser C. An analysis of face-to-face communication. In: Bennet AE (Ed). Communication between doctors and patients. Oxford University Press, 1976.

Friedman HS. Nonverbal communication between patients and medical practitioners. Journal of Social Issues, 1979; 35: 82-99.

Friedman HS. Nonverbal communication in medical interaction. In: Friedman HS, DiMatteo RS (Eds). Interpersonal Issues in health care. Academic Press, New York, 1982.

Goldberg D. Training primary care physicians to recognize psychiatric disorder. Institute of Medicine; Publication 79-004. Washington: National Academy of Sciences, 1979.

Goldberg D, Blackwell B. Psychiatric illness in general practice. A detailed study using a new method of case identification. British Medical Journal, 1970; 2: 439-443.

Goldberg D, Huxley P. Mental illness in the communicty; the pathway to psychiatric care. Tavistock Publ., London/New York, 1980.

Grol R, Eyk J van, Huygen F, et al. Huisarts en somatische fixatie. Theorie en praktijk van de preventie van somatische fixatie. NUHI, Nijmegen, 1981.

Groot AD de. Methodologie. Mouton, 's-Gravenhage, 1961.

Hess JW. A comparison of methods for evaluating medical student skills in relating to patients. Journal of Medical Education, 1969; 44: 934-938.

Holten-Vriesema J, Tempot C, Aalderen HJ van, et al. Methodisch werken. Over een algemene methode van hulpverlening en de opbouw van een functionele relatie toegespitst op de huisartspraktijk. Huisarts en Wetenschap, 1978; 21: 322-335.

Hulka BS, Cassel JC, Kupper LL, Burdette JA. Communication, compliance and concordance between physicians and patients with prescribed medications. American Journal of Public Health, 1979; 66: 847-853.

Illich I. Medical nemesis: the expropriation of health. Pantheon, New York, 1976.

Ivey AE. Intentional interviewing and counselling. Wadsworth, Belmond Calif., 1983.

Johnson DW, Matross RP. Attitude modification methods. In: Kanfer H, Goldstein AP (Eds). Helping people change. Pergamon Press, New York, 1975.

Johnstone A, Goldberg D. Psychiatric screening in general practice. The Lancet, 1976; 1: 605-608.

Jourard S, Jaffee P. Influence of an interviewer's behavior on the self-disclosure behavior of interviewees. Journal of Counseling Psychology, 1970; 17: 252-257.

Kasdorf J, Gustafson K. Research related to microtraining. In: Ivey A, Authier J. Microcounseling: innovations in interviewing, counseling, psychotherapy and psycho-education. Charles C Thomas, Springfield Ill., 1978.

Kassirer JP, Gorry GA. Clinical problem solving: a behavioral analysis. Annals of Internal Medicine, 1978; 89: 245-255.

Katon W, Kleinman A. Doctor-patient negotiation and other social science strategies in patient care. In: Eisenberg L, Kleinman A (Eds.). The relevance of social science for medicine. Reidel Publ., Dordrecht, 1980.

Kelley HH, Michela JL. Attribution theory and research. Annual Review of Psychology, 1980; 31: 457-501.

Korsch BM, Gozzi EK, Francis V. Gaps in doctor-patient communication; I. Doctor-patient interaction and patient satisfaction. Pediatrics, 1968; 42: 855-869.

Korsch BM, Negrete VF. Doctor-patient communication. Scientific American, 1972: 66-74.

Kuiper PC. Hoofdsom der psychiatrie. Bijleveld, Utrecht, 1981 (9e druk).

Langer EJ, Janis IL, Wolfer JA. A reduction of psychological stress in surgical patients. Journal of Experimental Social Psychology, 1975; 11: 155-165.

Larsen KM, Smith CK. Assessment of nonverbal communication in the patient-physician interviews. Journal of Family Practice, 1981; 12: 481-488.

Lazare A, Eisenthal S, Wasserman L. The customers approach to patienthood. Attending to patient requests in a walk-in clinic. Archives of General Psychiatry, 1975; 32: 553-558.

Leigh H, Reiser MF. The patient: biological, psychological and social dimensions of medical practice. Plenum Press, New York, 1980.

Levenstein JH, McCraden EC, McWhinney IR, Stewart MA, Brown JB. The patient-centred clinical method. I: A model for the doctor-patient interaction in family medicine. Family Practice, 1986; 3: 24-30.

Ley P. Patients' understanding and recall in clinical communication failure. In: Pendleton P, Hasler J (Eds.). Doctor-patient communication. Academic Press, London, 1983.

Lipkin M Jr, Kupka K. Psychosocial factors affecting health. Praeger Publ., New York, 1982.

Lipkin M, Quill TE, Napadano RJ. The medical interview: a core curriculum for residencies in the internal medicine. Annals of Internal Medicine, 1984; 100: 277-284.

McKinley JB. Social network influences on morbid episodes and the career of help-seeking. In: Eisenberg L, Kleinman AM (Eds.). The relevance of social science to medicine. Reidel, Amsterdam, 1980.

Mehrabian A. Non-verbal communication. Aldine-Atherton, New York, 1972.

Metz JCM. Medische competentie. Een onderzoek naar de betrouwbaarheid en validiteit van het gestructureerd klinisch examen. Diss. Nijmegen, 1984.

Morgan AJ, Morgan MD. Manual of primary mental health care. Lippincott Cie., Philadelphia, 1980.

Morgan WL, Engel GL. The clinical approach to the patient. Saunders, Philadelphia, 1969.

Pelletier KR. Holistic medicine. Delacorte Press, New York, 1979.

Pendleton D, Schofield T, Tate P, Havelock P. The consultation: an approach to learning and teaching. Oxford University Press, Oxford, 1984.

Philipsen H. Het medisch oordeel: besliskunde of speltheorie. Tijdschrift voor Sociale Geneeskunde, 1979; 57: 838-842.

Regier DA, Burke JD, Burns BJ, et al. A proposed classification of social problems and psychological symptoms for inclusion in a classification of health problems. In: Lipkin M Jr, Kupka K. Psychosocial factor affecting health. Praeger Publ., New York, 1982.

Robins LN, Helzer JE, Croughan J, Janet BW, Spitzer RL, NIMH. Diagnostic interview schedule. Version III. Nat. Instit. of Mental Health, Bethesda, Maryland, 1979.

Rogers CR. Client-centred therapy. Houghton Mifflin, Boston, 1951.

Romme MAJ. Doel en middel. Diss. Amsterdam, 1967.

Saghir MT. A comparison of some aspects of structured and instructured psychiatric interviews. American Journal of Psychiatry, 1971; 128: 180-184.

Schouten JAM. Anamnese en advies. Stafleu, Alphen a/d Rijn/Brussel, 1982.

Shepherd M, Cooper B, Brown AC, Kalton GW. Psychiatric illness in general practice. Oxford University Press, London, 1966.

Skipper J, Leonard R. Children, stress and hospitalization: a field experiment. Journal of Health and Social Behavior, 1968; 9: 275-287.

Stiles WB, Putnam SM, Wolf MH, et al. Interaction exchange structure and patient satisfaction with medical interviews. Medical Care, 1979; 17: 667-679.

Stimson G, Webb B. Going to see the doctor: the consultation process in general practice. Routledge and Kegan Paul, London, 1975.

Stoeckle JD, Barsky AJ. Attributions: uses of social sciences knowledge in the "doctoring" of primary care. In: Eisenberg L, Kleinman A (Eds). The relevance of social science for medicine. Reidel Publ., New York, 1980.

Stone GC. Patient compliance and the role of the expert. Journal of Social Issues, 1979; 35 (1): 34-59.

Storms MD, Nisbett RE. Insomnia and the attribution process. Journal of Personality and Social Psychology, 1970; 16: 319-328.

Tessler R, Mechanic D. Psychological distress and perceived health status. Journal of Health and Social Behavior, 1978; 19: 254-262.

Tuckett D. Meetings between experts: an approach to sharing ideas in medical consultations. Tavistock, London, 1985.

Tuckett D, Williams A. An approach to the measurement of explanation and information giving in medical consultations: a review of empirical studies. Social Science in Medicine, 1984; 18: 571-580.

Verhaak PFM. Interpretatie en behandeling van psychosociale klachten in de huisartspraktijk. Diss. NIVEL, Utrecht, 1986.

Wing JK, Cooper JE, Sartorius N. The measurement and classicification of psychiatric symptoms. Cambridge University Press, New York, 1974.

Zola IK. Pathways to the doctor: from person to patient. Social Science in Medicine, 1973; 7: 677-689.

CHAPTER 3 THE MEDICAL INTERVIEW: EFFECTS ON PATIENT AND PHYSICIAN

A.A.M. Crijnen and H.F. Kraan

In this chapter, we discuss the outcomes of a medical interview that are generally considered to be essential (Locker and Dunt, 1978; Ware et al., 1978; Wolf et al., 1978; Lebow, 1982, 1983; Pendleton, 1983).

3.1 Affective satisfaction.

Affective satisfaction is defined as the patient's perception of the quality of physician-patient communication, including feelings of trust and confidence in the physician and his perception of the physician's positive regard and willingness to listen to his concerns (Wolf et al., 1978). Although most patients are moderately to highly satisfied with the communication, 11% of all patients are moderately dissatisfied whereas 15% were even highly dissatisfied (Korsch and Negrete, 1972; Francis et al., 1969). The severest and most common complaint of dissatisfied patients is that physicians show too little interest in their concerns. 26% of all patients did not mention their greatest concern to the physician, because patients had no opportunity provided or were not encouraged to do so. Korsch et al. (1972) frequently observed a breakdown of communication under such circumstances. Some patients were so preoccupied with their dominant concerns that they were unable to respond to the physician's questions and advice. Attention to the patients' worries and concerns was found to correlate highly with success in satisfying them and obtaining their compliance with advice. Korsch recommends the opening of the medical consultation with open-ended questions pertaining to the reasons for the visit in order to elicit the patient's concerns.

Moreover, affective satisfaction has been significantly correlated with the absolute quantity of information conveyed by the patient to the physician. The exchange of patient-centered information can be enhanced by a physician's interview behavior, allowing patients to tell their story in their own words (Stiles et al., 1979; Putnam et al., 1985). Eisenthal and Lazare (1977) found that interview behavior aimed

at helping the patient to put his request into words correlates significantly with measures of satisfaction and the feeling of being helped. They experienced that patients find it difficult to state their request, whilst, at the same time, feeling that verbalizing the request is very important. Patients can be helped in the expression of a request by a structuring activity from the physician or by a specific kind of collaborative involvement. Studies suggest that patients need to be given time to express their concerns and describe their illnesses.

Other studies indicate that the affective dimension of physician-patient communication plays an important part in the evaluation of the consultation because patients lack the knowledge necessary to evaluate the physician's medical competency. The perceived degree of emotional support given and physician's time, interest and devotion are major dimensions in the evaluation of the consultation by the patient (Segal and Burnett, 1980).

Patient-satisfaction is based on the physician's affective behavior rather than on his technical performance in consultations where the presented problem incorporates emotional involvement due to the perceived seriousness of the problem and uncertainty about the consequences, in the situation where the patient is unable to judge the quality of the presented solutions (Ben-Sira, 1976).

3.2 Insight.

Insight refers to the explanations and information given by the physician and to the patient's understanding of diagnosis, etiology prognosis and effects of treatment (Wolf et al., 1978). This issue has been studied several times: some of these studies are focused on the process of information-exchange, whereas others describe the kind of information exchanged.

The relation between explanation and information-giving on several outcomes of the consultation has been studied by Tuckett et al. (1985). At the end of a consultation, 10% of the patients were unable to remember what their physicians had told them about diagnosis, purpose of treatment or prevention. Of all patients who were able to recall important information, only 73% correctly interpreted one of the topics

in the consultation (80% had a correct opinion on diagnosis, 92% about advice on therapy and 75% about advice on prevention). Of those patients who correctly interpreted important information, 25% disagreed with the physician's view on at least one of the important points the physician made. Tuckett (1985) therefore concludes that patients encounter difficulties in processing the information given by the physician; this is not due to a failure of memory but occurs during the interpretation and evaluation of the physician's information. The interpretation given by patients of the information provided by the physician, depends on patients' attributions before the consultation and their scope for using them afterwards. Tuckett recommended that physicians should pay heed to the content of the ideas they communicate. Moreover, they should establish and discuss points of difference between their own and their patients' health beliefs and explanatory models. Earlier studies have also pointed in the same direction: the effectiveness of the information provided depends on the extent to which it addresses the patient's concerns and expectations (Bartlett, 1981).

Other studies have revealed that patients recalled only 40% of information correctly (Anderson et al., 1979), whereas 48% of what patients thought was said, was imagined or misconstrued. The more information given, the more is recalled, although proportionately less of this is correct. Patients remember more information about treatment and medication correctly than about diagnosis.

Ley et al. (1973) attempted to enhance the patient's recall of information by organizing medical information into explicitly labelled categories. The use of explicit categorization increased recall of information, especially in the category on what patients should do about their complaints.

The kind of information that patients expect to acquire during a medical consultation does not differ between patients from different classes or educational backgrounds. All patients expect to receive information about diagnosis, causes, prognosis and about their complaints and a proposal for further treatment (Waitzkin, 1985).

Patients' insight will be increased as a result of the exchange of information. The explanation given correlates significantly with the perceived controllability of the illness. Patients who consider their illness to be more controllable obtain more explanation from the physician and feel more confident about managing their illnesses (Putnam et al., 1985). This is probably achieved by influence over patients' health understanding (Pendleton, 1983). Moreover, patients' rapport and cooperation are stimulated by specific instructions, expressions of trust in patients' ability for self-care, warm concern and individualization of advice.

Unfortunately, many factors impinge on the quality of information exchange. The Korsch study (1972) discloses that nearly a fifth of all patients did not receive a clear statement of what was wrong whereas half of all patients were still wondering what had caused their illness when they left the physician's office. Prognoses were never offered. When the expected information about causation and nature of the illness is not provided, satisfaction and compliance with treatment decreases significantly. Medical jargon leaves many patients uninformed about the nature of their problem. Technical discussions about the patient's condition using inpersonal or institutional expressions are less likely to stimulate patient rapport and cooperation. Reducing the exchange of information by truncating the interaction is negatively associated with insight (Stiles et al., 1979). We conclude, therefore, that patients expect to receive information about their condition, but that the quality of information exchange is easily impaired.

3.3 Compliance.

Compliance or rather, non-compliance, with the medical regimen has been well-studied as an outcome measure of the medical interview. Davis (1966, 1968) classified 37% of patients as non-compliant, whereas Francis et al. (1969) measured 38% as moderately compliant and 11% as non-compliant. In a well-known study, it was reported that 58% of all patients made errors in the taking of medication (Hulka et al., 1976).

Non-compliance appears to be partly related to the quality of physician-patient communication. Davis (1968) describes several types of communication between patient and physician which are associated

with non-compliance. These communication types are a patient with malintegrative behavior, an authoritarian physician, a non-directive physician, or a physician who collects information without giving any feedback. A communication style which is positively related to patient-compliance is characterized by joking, laughing, signs of satisfaction with the physician-patient relationship and tension release. Davis concludes, therefore, that compliance is a function of a delicate balance between providing and obtaining information presented in a manner that is acceptable to the patient.

Furthermore, the clarity of physicians' instructions, followed by the interest in patients' symptoms, the amount of information given about the disease and several dimensions of physicians' medical competence, discriminates most effectively with regard to patients' compliance with physicians' instructions (Vuori, 1972). The lack of congruity between what the patient thinks he is supposed to do and what the physician thinks the patient should do was found furthermore to induce non-compliance (Hulka et al., 1976). Patients tend to be compliant to the best of their knowledge but they are sometimes acting on misinformation which leads to misunderstanding and confusion. Providing knowledge of drug-function and demonstrating which drugs should be taken for which purpose during follow-up visits are recommended as inducing compliance. When patients were informed about what was expected of them, more than 85% complied. Two factors seem particularly important in inducing adherence to the medical regimen. These factors are the extent to which the patient understands the information presented and to which the patient remembers the message (Ley, 1983). Adherence to medication is mediated by patient satisfaction and recall of information which are induced by the effect of physicians' interpersonal skills and the provision of information (Bartlett, 1981).

Eisenthal and Lazare (1977) advocate a customer approach consisting of the elicitation, negotiation and disposition of the patient's request during the initial psychiatric interview in a walk-in clinic. Satisfaction and feelings of having been helped correlate highly with patient participation in the treatment planning. Adherence to the treatment-regimen is significantly related to negotiation.

Negotiation results in the legitimization of patients' ideas about treatment and the sollicitation of their participation in the treatment planning. Even patients who do not get the treatment they want, comply well.

3.4 Anxiety reduction and reassurance.

Anxiety reduction and reassurance are often mentioned as important outcomes of the medical interview, although they are rarely studied.

Sprecher et al. (1983) tried to reduce patients' anxiety by encouraging patients to disclose underlying concerns and then either confirming or disproving these concerns: this proved to be an effective way of reducing patients' anxiety. The exact mechanism by which this was achieved is not yet understood: a physician's interview behavior might communicate that the physician has understood or heard the patient's fears accurately or it might mean that the patient's uncertainties have been further resolved. Reynolds (1978) interviewed patients on a surgical ward in a large hospital and observed that 24 % of the patients wanted more information about investigations and 38% about the results of their investigations. Anxiety and fear were the inevitable consequences of the poor communication. To most hospitalized patients, fear of the unknown is a much heavier burden to bear than full knowledge of their illness. According to DiMatteo and DiNicola (1982), patients who are emotionally upset rarely comprehend information clearly. Learning appears to be most effective when patients experience a moderate level of anxiety. Therefore, explanations given to the patient might do little to change his state of knowledge about his condition unless the patient's anxiety level is also reduced.

3.5 Changes in health status.

Improvement in patients' health status has been recognized and studied only recently as a long-term outcome of the medical consultation.

A significant association was found between the number of patients' utterances during history-taking and the improvement of their symptom status (Putnam et al., 1985).

Greenfield et al. (1985) increased patients' involvement in the medical consultation and found that two months after the intervention, fewer limitations in physical and role-related activities were reported in comparison with a control group. Patients who showed more commitment to the therapeutic regimen and who had an increased sense of self control, were better able to regulate their bloodsugar and showed lowered diastolic blood pressures (Kaplan, 1986).

These studies stress that good physician-patient communication is important for the improvement of patients' health status.

3.6 Diagnosis.

The importance of the medical interview for the physician is mainly studied from the perspective of diagnosis. Hampton et al. (1975) have assessed the relative contributions of history-taking, physical examination and laboratory investigations to diagnosis and management of medical outpatients. The medical history provided enough information to make an initial diagnosis in approximately 80% of the consultations. Others found that history-taking is of major importance in establishing diagnosis and the treatment plan (Sandleb, 1980).

Closely related to diagnosis is the physician's awareness of the patient's request for help. Taylor (1980) observed disagreement between the physician's and the patient's definition of the reason for encounter in one third of all consultations and recommended an early discussion of the primary purpose of the visit to prevent the development of misperceptions. Other researchers have studied the process of clinical problem-solving (Elstein et al. 1978; Kassirer et al., 1978; Neufeld et al., 1981) and have found a strong association between history-taking and clinical problem-solving. History-taking is the verbal (or behavioral) dimension and problem-solving the cognitive dimension of the same process. The act of gathering data by means of asking questions forms input and output for the cognitive process of clinical problem-solving (see chapter 9). The quality of information-exchange is considered to be of importance to the quality of hypothesis-generation and testing and, through that, for the quality of diagnosis. The exchange of information can be distorted or inhibited by means of the use of medical vocabulary, ambiguous questions, two or more questions at the same time, etc.

The effect of patients' mislabelling of symptoms on medical outcome has been studied (Dirks et al. (1982). More than one quarter of 587 chronic asthmatic patients mislabelled one or more airway obstruction symptoms as being an asthma attack. Mislabelling patients were more than 40% more likely to be hospitalized than non-mislabellers. The researchers attributed this effect to the distortion of the patients' reports of their clinical picture.

Furthermore, the number of questions asked during history-taking is not unequivocally related to the quality of diagnosis. Physicians who are specialists in a certain domain ask fewer questions during history-taking, mention the correct diagnosis earlier and are sooner convinced of the diagnosis than are non-specialists (Kassirer et al., 1978). Non-specialists, like students or residents, often revert to a general review of systems when they have no clear idea in which direction to proceed during history-taking. Obviously, physicians can choose between several strategies for achieving the goals of diagnosis.

The relation between the accuracy of detecting minor psychiatric disorders and physicians' medical interview behavior has been studied by Goldberg et al. (1980). Ten types of medical interview behavior were related to the accuracy of detection and they appeared to improve significantly as a result of training: an important finding. These interview behavior patterns are: eye contact, clarification of presented complaints, using open to closed cones, empathic interview style, sensitivity to verbal and non-verbal cues, not reading notes during history-taking, dealing adequately with over-talkative patients and asking fewer questions about past history. In addition to the physician's interview style, his personality and academic ability were also related to the accuracy of detection of psychiatric disorder.

3.7 Concluding remarks.

In this chapter, we have elaborated on the relation between the process of the medical interview and certain outcomes for physician and patient. The studies presented legitimize affective satisfaction, insight, compliance, anxiety reduction and changes in health status as important outcomes for the patient.

Patients appreciate the opportunities they are given to express thoughts and emotions related to their problems. They expect to receive

information concerning their psychological or physical condition and on the treatment proposal which ultimately enhances their health understanding. Understanding, recall and commitment contribute to patients' compliance with treatment and their improved health status. Diagnosis is considered to be an important outcome of the consultation for the physician.

The studies reviewed have increased our understanding of the necessary ingredients of a medical interview. Affective satisfaction is closely related to the patient-centered phase in the interview which allows the patient to express his thoughts and emotions. Diagnosis can only be established by means of history-taking. Patients' insight and compliance-induction are mainly the result of the presentation of solutions in which information is provided and commitment established. Anxiety reduction can be induced by listening to patients' complaints and worries or by providing information. The quality of the relationship appears to be of importance during all phases of the interview.

All outcomes mentioned here can be influenced by a physician's medical interviewing skills. They are therefore ideally suited to form criterion measures for the impact of interviewing skills on the patient. This allows us to establish the contruct validity of our measures of medical interviewing skills and to develop theories on physician-patient communication.

REFERENCES

Anderson JL, Dodman S, Koppelman M, Fleming A. Patient information recall in a rheumatologic clinic. Rheumatology and Rehabilitation, 1979; 18: 18-22.

Bartlett EE. The effects of physician communication skills on patient satisfaction, recall and adherence (dissertation). The Johns Hopkins University , Baltimore, 1981.

Ben-Sira Z. The function of the professional's affective behavior in client satisfaction: a revised approach to social interaction theory. Journal of Health and Social Behavior, 1976; 17: 3-11.

Davis MS. Variations in patients' compliance with doctor's orders: analysis of congruence between survey responses and results of empirical investigations. Journal of Medical Education, 1966; 41: 1037-1048.

Davis MS. Variations in patient's compliance with doctor's advice: an empirical analysis of patterns of communication. American Journal of Public Health, 1968; 58: 274-288.

DiMatteo MR, DiNicola DD. Achieving patient compliance: the psychology of the medical practitioner's role. Pergamon Press, New York, 1982.

Dirks JF, Schraa JC, Robinson SK. Patient mislabelling of symptoms: implications for patient-physician communication and medical outcome. International Journal of Psychiatry in Medicine, 1982; 12: 15-27.

Eisenthal S, Lazare A. Expression of patient's request in the initial interview. Psychological Reports, 1977; 40: 131-138.

Eisental S, Koopman C, Lazare A. Process analysis of two dimensions of the negotiated approach in relation to satisfaction in the initial interview. Journal of Nervous and Mental Diseases, 1983; 171: 49-54.

Elstein AS, Shulman LS, Sprafka SA. Medical problem solving, an analysis of clinical reasoning. Harvard University Press, Cambridge, 1978.

Francis V, Korsch BM, Morris MJ. Gaps in doctor-patient communication: patients' response to medical advice. New England Journal of Medicine, 1969; 280: 535-540.

Hampton JR, Harrison MJG, Mitchell JRA, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination and laboratory investigation to diagnosis and management of medical outpatients. British Medical Journal, 1975; 2: 486-489.

Goldberg D, Huxley P. Mental illness in the community: the pathway to psychiatric care. Tavistock Publ., London, 1980.

Greenfield S, Kaplan S, Ware JE. Expanding patient involvement in care. Annals of Internal Medicine, 1985; 102: 520-528.

Hulka BS, Kupper LL, Kassel JC, Babineau RA. Practice characteristics and quality of primary medical care: the doctor-patient relationship. Medical Care, 1976; 13: 808-820.

Kaplan S. (Personal Communication), 1986.

Kassirer JP, Gorry GA. Clinical problem solving: a behavioral Analysis. Annals of Internal Medicine, 1978; 89: 245-255.

Korsch BM, Negrete VF. Doctor-patient communication. Scientific American, 1972; 227: 66-74.

Lebow J. Consumer satisfaction with mental health treatment. Psychological Bulletin, 1982; 91: 244-259.

Lebow J. Research assessing consumer satisfaction with mental health treatment. Evaluation and Program Planning, 1983; 6: 21-236.

Ley P, Bradshaw PW, Eaves D, Walker CM. A method for increasing patients' recall of information presented by doctors. Psychological Medicine, 1973; 3: 217-220.

Ley P. Patients' understanding and recall in clinical communication failure. In: Pendleton D, Hasler J (Eds.). Doctor-patient communication. Academic Press, London, 1983.

Locker D, Dunt D. Theoretical and methodological issues in sociological studies of consumer satisfaction with medical care. Social Science and Medicine, 1978; 12: 283-292.

Neufeld VR, Norman GR, Feightner JW, Barrows HS. Clinical problem-solving by medical students: a cross-sectional and longitudinal analysis. Medical Education, 1981; 15: 26-32.

Pendleton D. Doctor-patient communication: a review. In: Pendleton D, Hasler J (Eds.). Doctor-patient communication. Academic Press, London, 1983.

Putnam SM, Stiles WB, Jacob MC, James SA. Patient exposition and physician explanation in initial medical interviews and outcomes of clinic visits. Medical Care, 1985; 23: 74-83.

Reynolds M. No news is bad news: patients' view about communication in hospital. British Medical Journal, 1978: 1673-1676.

Sandler G. The importance of the history in the medical clinic and the cost of unnecessary tests. American Heart Journal, 1980; 100: 928-931.

Segal A, Burnett M. Patient evaluation of physician role performance. Social Science and Medicine, 1980; 14A: 269-278.

Sprecher PL, Thomas ER, Huebner LA, Norfleet BE, Jacoby KE. Effects of increased physician-patient communication on patient anxiety. Professional Psychology: Research and Practice, 1983; 14: 251-255.

Stiles WB, Putnam SM, Wolf MH, James SA. Interaction exchange structure and patient satisfaction with medical interviews. Medical Care, 1979; 17: 667-679.

Taylor RB, Burdette JA, Camp L, Edwards J. Purpose of the medical encounter: identification and influence on process and outcome in 200 encounters in a model family practice center. The Journal of Family Practice, 1980; 10: 495-500.

Tuckett D. Meetings between experts: an approach to sharing ideas in medical consultations. Tavistock, London, 1985.

Vuori H, Aako T, Aine E, Erkko R, Johansso, R. The doctor-patient relationship in the light of patients experiences. Social Science and Medicine, 1972; 6 (6): 723-730.

Ware JE, Davies-Yvery A, Stewart AL. The measurement and meaning of patient satisfaction. Health and Medical Care Services Review, 1978; 1: 1-15.

Wolf MH, Putnam SM, James SA, Stiles WB. The medical interview satisfaction scale: development of a scale to measure patient perceptions of physician behavior. Journal of Behavioral Medicine, 1978; 1: 391-401.

CHAPTER 4 THE CONSTRUCTION OF THE MAASTRICHT HISTORY-TAKING AND ADVICE CHECKLIST (MAAS)

H.F. Kraan and A.A.M. Crijnen

4.1 Introduction

In the discussion of our educational model of interviewing skills in chapter 2, we arrived at the following objectives for the construction of the MAAS-Primary Mental Health Care (MAAS-PMHC).

- measurement of interviewing skills derived from our educational model of initial interviews in general practice and in primary mental health care;
- design of a method which measures teachable interviewing skills and which can be used to give feedback in education.

The question "how are medical interviewing skills measured?" now arises. In section 4.2 we address this question and conclude with six criteria, prerequisite to our measurement purposes (4.3).

In 4.4 a search through the literature is carried out in vain for a method which fulfills our 6 pre-set criteria. We subsequently describe the procedure of construction (4.5), the item domain of the MAAS-General Practice (MAAS-GP) (4.6) and of the MAAS-Primary Mental Health Care (4.7).

In 4.8, self- and global rating variants of the MAAS-GP and MAAS PMHC, as well as a content checklist measuring the information obtained from the patient is constructed.
Remarks about the practical use of the constructed instruments can be found in 4.7.

4.2 How are medical interviewing skills measured?

To be more precise, this question is refrased as: which method(s) and which categorizations of interviewing skills are preferable?

As a method for the measurement of interviewing skills, we have chosen behavior assessment by external observers because of its good perspectives with regard to reliability, scalability and validity.

Behavioral assessment has been successfully applied to personality research (a.o. Mischel, 1968) and to research into behavioral modification (a.o. Cone et al., 1977). The superiority of behavioral assessment over other methods, such as self-assessment and global ratings of behavior, has frequently been investigated (see for a discussion a.o. Beekers, 1982; Streiner, 1985). The MAAS, subject matter of this thesis, has therefore been constructed as a behavioral assessment method which is operationalized in rather simple, clearly defined and delimited units of behavior.

In addition to this method of behavioral assessment of interviewing skills several other methods are constructed which are used in forthcoming validity studies.

Firstly, two self-assessment methods of interviewing skills are introduced: an assessment of detailed, circumscribed interviewing behavior and a global rating scale.

Secondly, a global rating scale is constructed which experts use to evaluate the physician's interviewing skills.

Thirdly, an evaluation method used by interviewees (patients), called the Patient Satisfaction with Communication Checklist, (PSCC) has been designed. Chapter 6 is devoted to this method.

With these main types: external observation (behavioral assessment, global rating), self-assessment (behavioral assessment, global rating) and assessment by the interviewees (patients), we cover all possible measurement methods of interviewing skills. We use these other methods, which are strongly related to the MAAS in the theoretical sense, in our validity research.

Let us now return to the categorization of interviewing skills. We restrict ourselves to the categorization within the behavioral assessment method which we prefer for the construction of the MAAS. Five main observable categories of interviewing skills can be distinguished (a.o. Stiles et al., 1986).

- Content categories, pertaining to what is said (the semantic content). Content categories range from particular topics of interest (e.g. topics about a specific medication or specific complaints) to grouped topics (e.g. general categories, such as

"somatic" or "psychosocial" complaints).

- Speech-act categories concerning the acts performed when someone says something as opposed to the content of his words. Example: "How long have you been out of work?", is a question about the patient's employment situation, but the act performed by the speaker, is the asking of a question and the eliciting of an answer from the patient.
- Non-verbal communicative behaviors; for instance: voice tone, gaze, posture, laughter, hesitation, facial expression, etc.
- Ratings of affect and evaluative ratings of complex interviewing skills. The former the rater judges the emotional tone in (a part of) the interview. In the latter, the rater judges how well a complex interviewing skill has been performed.
- "Conjunctive categories" which combine one or more elements of the above-mentioned categories into one single category. For instance: "asking questions about the precipitation of the complaints" is an example of a "conjunctive category", combining content and speech act categories.

After listing observable categories, we take a closer look at how these categories are sampled from medical interviews. Stiles and Putnam (1986) distinguish two types of sampling: coding and rating.

Coding involves the use of nominal scales to categorize bits of content (semantic elements) or bits of behavior (e.g. "open questions"; "closed questions" etc.). This coding can be "complete" when every utterance, sentence, interviewing skill, semantic element (which-ever the category noted) during the interview is scored. Coding is "incomplete" when proceeding from certain criteria for scoring, only pre-defined bits of content or behavior are selected and coded; for instance, the "closed questions" in the interview and not the other types of questions. The distinction between "complete" and "incomplete coding" is, however, always relative because coding can never be "complete" in a philosophical sense. When incomplete coding is used, then "threshold problems in scoring" may arise (Rutter et al., 1981; Stiles et al., 1986). Even when observers acknowledge the coding criteria of a particular type of interviewing behavior or content

element, they may have difficulties in agreeing on whether specific behavior is occurring or not. Example: raters may agree on the criteria by which expression of empathy should be recognized. The problem starts when raters have to agree on whether the "amount of empathy" in a pertinent expression is sufficient in order to be a codable expression of empathy (Rutter et al., 1981).

Rating is an attempt to quantify a quality of interviewing behavior (for instance: affective behavior). Likert-scales are often used for rating.

The first four categories of the Stiles and Putnam list are mainly coded (complete or incomplete), whereas the fifth category is by, definition, rated on (Likert-type) scales.
In our view, Stiles and Putnam's meta-classification provides a useful overview of which measurement categories can be encountered in the literature on measurement of interviewing skills.

We would like however, to include as much as possible in their speech-act category the more complex skills specific to medical interviews. In this way we can thus list complex interviewing skills, such as summarizing, reflection, confrontation etc. in this speech-act category. It is, however, necessary to use rating scales, especially when the quality of complex interviewing skills is to be evaluated. Nevertheless, a measurement-method serving educational objectives should state in precise behavioral terms how these complex skills should be performed.

Those categories appropriate to our purposes are indicated in table 4.1 which presents the matrix of possible measurement categories of interviewing skills.

Measurement of interviewing skills during the phases "exploration of the reasons for encounter", "history-taking" and "presenting solutions" asks for conjunctive categories of content and of speech-acts, whereas the skills for "structuring the interview" as well as "interpersonal and communicative skills" mainly require affective and evaluative ratings.

From the five aforementioned main categories of observable interviewing behavior we do not use non-verbal behavior because of difficulties in operationalization.

Table 4.1: Overview of observable categories of medical interviewing skills.

	speech-acts		medical content elements		constructs of affect during interview	non-verbal inter-viewing behavior
coding or rating	"simple" inter-viewing	complex inter-viewing	specific topics	groups of topics		
complete coding						
incomplete coding	X	X	X	X		
rating	X	X			X	

X Indicates the preferable categories for the construction of the Maastricht History-taking and Advice Checklist.

4.3 Criteria for measurement of interviewing skills according to our requirements.

Based on the previous discussions, the requirements for the measurement of interviewing skills is summarized in a list of 6 criteria. Using these criteria, a literature review of existing methods is carried out.

1. The method should be observational with the following selected categories of interviewing skills:
   - content categories specific for initial medical interviews in general practice and in primary mental health care
   - speech-act categories, pertaining to simple and complex interviewing skills
   - affective and evaluative ratings to complex interviewing skills
   - conjunctive categories of speech-act and medical content elements

   For reasons of practicability, incomplete coding is used.
2. The focus should be on the physician's interviewing skills.
3. The method should measure interviewing skills which can be taught effectively i.e. those which are susceptible to behavior modification. The method should also be suitable as a feedback tool in education.
4. The method should guarantee reliable measurement.
5. Besides content validity, the method should have construct validity.
6. Practicability of the method: it should be expressed in reasonable test length and scoring time, clear definitions and criteria for scoring, training of observers, handy lay-out of instrument and a user-friendly manual.

4.4 Search through the literature methods for measuring interviewing skills according to the pre-set criteria.

Before constructing the MAAS-General Practitioner and the MAAS-Primary Mental Health Care, we reviewed the literature hoping to find a method satisfying the above-mentioned criteria. In 4.4.1, these methods are discussed. We draw conclusions on their testing against our pre-set criteria in 4.4.2.

4.4.1 Discussion of measurement-methods encountered in the literature.

In the literature we found 22 methods which we summarized in table 4.2. at the end of the chapter.

The discussion of the 22 methods, to which we refer by means of the first authors' names takes place according to our pre-set criteria.

1. Are the methods and categorizations suitable?

The majority of methods used in non-evaluative research of the medical interview is based on the Interaction Process Analysis developed by Bales (1950). Bales originally developed the Interaction Process Analysis for assessment of interactions in small groups. It has since been applied to physician-patient communication. The physician's and the patient's interviewing behavior are taken equally into account. As an application of the Interaction Analysis in physician-patient communication, we present Roter's (1977) modification of Bales' system.

Table 4.3: Roter's categories for inter-actional process analysis.

physician	patient
Personal remarks.	Personal remarks.
Shows approval, gives compliment.	Shows approval, gives compliment.
Statement, gives information, opinion.	Statement, gives information, opinion.
Gives direction, instruction.	Request for medication.
Asks questions.	Bids for clarification.
Direct request for questions.	Asks questions.
Shows agreement and/or understanding.	Shows agreement and/or understanding.
Shows disagreement or criticism.	Shows disagreement or criticism.

As shown in table 4.3, eight mutually exclusive inter-actional categories for physicians as well as for patients are stipulated. The differences with Bales' systems are an extension of the speech-

act categories (e.g. personal remarks, given direction, bids for clarification etc.) and the addition of four affective rating scales: anger-irritation; sympathy-kindness; anxiety-nervousness; matter-of-factness-professionalism.
Such "interaction analysis" instruments are widely used in research into the medical interview. Their meticulous, complete coding of "molecular" speech-acts enables the researcher to follow closely the process of information-exchange between physician and patient. They have however several disadvantages:

- They use very time-consuming, cumbersome complete coding [such as the methods of Adler (1966), Hess (1969) and MacDonald (1981)].
- They do not measure complex skills, such as summarizing, confrontation, various types of reflection, self-disclosure, checking of information etc. which are characteristic of medical interviews. The specifics of these complex interviewing skills cannot be retraced by means of these coding systems.
- In the same vein, feedback to interviewers about their interviewing skills is restricted.
- Medical content aspects concerning history-taking, diagnosis, prognosis, treatment etc. are not taken into account.

To compensate for some of these disadvantages, some researchers (e.g. Freemon, 1971) add checklists of content categories in these instances (Sprunger, 1983).

More in accordance with our measurement objectives, the instruments of Hollifield et al. (1957), Van Dorp (1977), Stillman (1980), Hill (1981), Goldberg et al. (1981) use speech-act categories which reflect realistically the character of medical interviewing skills. Moreover, the more practical incomplete coding is used. These methods also have their shortcomings:

- threshold problems which are inherent in the use of incomplete coding
- reliability problems with rating scales
- the absence of medical content aspects

Some of the instruments therefore add medical content elements, for instance, constructing conjunctive categories such as the methods designed by Jarrett (1972), Brockway (1978), Barsky (1980), Rutter (1981), Mumford (1984).

A special position is taken by the "Dutch" instruments of Mokkink (1982), Den Hoed (1982) and Pieters (1982). They all use evaluative, rating scales and pay considerable attention to interviewing skills but, in addition also measure other competency domains of the physician, such as medical problem-solving and perceptual and interpretative skills and attitudes. Although they are comprehensive evaluation methods for initial medical consultations, they are not specific for the competency domain of medical interviewing skills.

2) Is the measurement applicable to initial interviewing in general practice and primary mental health care?

Content of medical interviews can be typified along two axes.

The first axis is the medical discipline of the interviewer; e.g. general practice/primary care, psychiatry, pediatrics, etc.

The second axis becomes clear when one considers the negotiated consensus model of Lazare (1975). Two sub-types of interviews can be derived from this model:

- initial interviews where the patient's request is elucidated and further planning for responding to this request is discussed
- follow-up interviews where the physician actually responds to the patient's request

Examples of measurement-methods of initial interviews in primary care settings are those of Barsky (1980), Mokkink (1982), Den Hoed (1982), Pieters (1982). Freemon's (1971) and Sprunger's (1983) instruments pertain to pediatrics.

Several methods are designed for initial interviews in primary mental health care: Jarrett's (1972), Hill's (1978), Rutter's (1981) and Goldberg's (1981) instruments.

Some evaluative remarks about these methods: Rutter's method is very elaborate, but its content checklist pertains to child psychiatry. Hill's method is especially designed for evaluative counseling sessions. Jarrett's and Goldberg's methods are rather restricted in their variety of interviewing skills to be measured.

3) Is the focus on the physician's interviewing skills?
In general, the "interaction analysis methods" pay the same attention to the physician's as to the patient's communicative behavior.
Methods designed for evaluative purposes obviously focus exclusively on the physician's interviewing skills such as those of Hollifield (1957), Hess (1969), Van Dorp (1977), Barsky et al. (1980), Rutter (1981), Goldberg (1981), Mokkink (1982), Den Hoed (1982), Pieters (1982), Sprunger (1982) and Mumford (1984).

4) Are teachable interviewing skills measured?
Among the instruments we reviewed, there is only a small number that exclusively measure teachable interviewing skills and that are suitable for the provision of immediate feedback to students about their interviewing behavior. This criterion is fulfilled when the methods are operationalized in rather simple, clearly-defined and delimited interviewing behavior. It only concerns the methods of Brockway (1978) and, to a lesser extent, those of Jarrett (1972), Rutter et al. (1981) and Sprunger (1983).
Moreover, in table 4.2, it is indicated which methods have actually been used to measure effects of teaching programs, witnessed by reports in the literature. Not all methods reviewed under this heading have been used in this sense.

5) Are reliability studies available?
In our literature review, the majority of instruments encountered have been tested for reliability: mainly concerns the inter-rater reliability. Although a host of procedures has been used, such as Pearson's correlation between raters, weighted Kappa's, intra-class correlations and percentage agreement, it is a pity that this latter measure, which is the weakest in the methodological sense, has been the most applied by far. Perhaps, therefore this is why the reported reliabilities so frequently vary from sufficient to good.

6) Are validity studies available?

Our literature review reveals a scarcity of validity studies. With the "interaction analysis" instruments (Bales, 1950; Freemon, 1971; Roter, 1977) and the Verbal Response Mode of Stiles (1978), studies have been performed that evidence predictive validity: the physicians interviewing behavior has been proven to explain some variance in outcome variables, such as patient's satisfaction and recall of medical information (Inui et al., 1982).

Stillman et al. (1977) have carried out validity research with their ACIR-scale: convergent validity (the ACIR is able to measure a predicted growth in interviewing skills), divergent validity (the ACIR does not measure medical or scholastic aptitude).

Swanson (1981), has correlated the scores of the ACIR-scale (Stillman, 1980), with a modification of Hess' instrument (1969) and with the History and Physical Exam Checklist (Swanson et al., 1981). He concludes that it is impossible to find evidence for construct validity because the low inter-case reliability does not allow comparison between instruments.

In chapter 5, our desiderata according to validity research is elaborated further.

7) Are the methods practicable?

In most descriptions of the methods used in educational evaluation of interviewing skills, remarks are made about the burden of the scoring and of the training of observers. In this respect the methods of Barbee (1967), Hess (1969), Barsky (1980), Stillman (1980), MacDonald (1981), Mokkink (1982), Den Hoed (1982), Pieters (1983), Sprunger (1983) and Mumford (1984) look feasible.

The "interaction analysis" instruments have, however a rather time-consuming and cumbersome manner of scoring because of their complete coding. Moreover, the interview often has to be transscribed in a verbatim protocol before scoring can take place.

4.4.2 Which methods are applicable to our measurement criteria?

To answer this question we apply our criteria in three successive steps to the reviewed sample. The methods which withstand this procedure, meet our measurement requirements.

First, we select methods by applying the criteria "observation method and preferable categories of interviewing skills", "measurement of teachable interviewing skills" and "practicability". Only those methods remain which measure simple and complex interviewing skills, include content categories and are practicable, i.e. use incomplete coding and/or rating scales. The methods of Barbee, Jarrett, Brockway, Barsky, Rutter, Den Hoed, Sprunger, Pieters and Mumford fulfil these criteria.

Second, several of the remaining instruments fit the requirements of initial interviews in primary mental health care (Jarrett and Rutter) and general practice (Barsky, Den Hoed and Pieters). All these instruments also fit the requirement of "focus on the physician's interviewing behavior".

Third, the scarcity of validity studies, especially in construct validity, is striking, and evidences a low tendency to investigate underlying theories of medical interviewing. Reliability is often investigated, but the research is not very sophisticated. The weak measure of percentage agreement is often used, whereas more robust methods, such as generalizability analysis and probabilistic scale analysis (Thorndike, 1982), have never been reported. None of the methods meets the criterion of "available reliability and validity studies" satisfactorily.

We conclude, therefore that none of the reviewed instruments fits our pre-set criteria precisely.

The consequence of this review entails the construction of a "new" measurement-method for interviewing skills in general practice and also one for mental health problems. Construction of a "new" instrument means, on the one hand, referring back to the theoretical digressions of chapters one and two and on the other hand, learning lessons from the useful experiences of the reviewed authors.

4.5 The construction of the MAAS.

According to Thorndike (1982), it is desirable to draw up a "test plan" when a measurement-method is constructed. These recommendations are briefly described as follows:

1. Initial definition of the competency domain the method is designed to assess.
2. Description of the use of the method (type of subjects, type of decisions on which to base it).
3. Constraints within which the method must operate.
4. Design of a blueprint: assembling content specifications (topics to be covered, skills to be tapped).
5. Specification of the format of the items (nature of stimulus materials, type of response to be made, procedure for scoring).
6. Plan for try-out of the proposed method, for analyzing the try-out data and for selecting items for the final method.
7. Specification of the statistical parameters desired in the finished method.
8. Outlining further data-collection and analysis for further reliability and validity studies.
9. Organization of the test manual and other auxiliaries.

Checking these steps, we must conclude that much preparatory work has already been done.

Steps 1 and 2 are described in chapter 2, especially from 2.2 onwards.

Step 3 is discussed - as far as it is applicable - in 4.2.

Steps 4 to 6 are dealt with below.

We first outline the procedure of how the blueprint and the final version of the MAAS-GP and MAAS-PMHC (4.5.1) have been constructed (4.5.2). We then describe the scales of the MAAS-General Practice (4.6) and of the MAAS-Primary Mental Health Care (4.7).

Steps 7 and 8 are the subject matter of the next chapter.

Step 9, the manual for observers, can be found in appendix A.

4.5.1 Steps involved in the construction of the MAAS-General Practice.

- An initial blueprint was designed by a founding panel consisting of two general practitioners, two psychologists and one social psychiatrist. It was based on the theoretical knowledge of process

and content in initial medical interviews as decribed in chapter 2. After several revisions, this first blueprint resulted in a second 56-item checklist with 4 scales for assessment of the physician's interviewing behavior:

1. exploration of the reasons for encounter
2. structuring the interview
3. quality of basic interviewing skills
4. designing a treatment plan

- This second blueprint was tested by the panel using this method, they rated videotaped initial interviews with patients presenting somatic and mental health problems. This testing resulted in:
  - removal of ambiguously worded items
  - developing definitions of interviewing skills and criteria for scoring
  - clearer item formulation
  - extension and improvement of medical content elements in the items
- A third blueprint was the result of these extensions and improvements. It already revealed the present format of the MAAS with the six scales. It was decided that the method was to reflect the three characteristic phases of initial interviews. The following are the first 3 scales:
  1. exploration of the reasons for encounter
  2. medical history-taking
  3. presenting solutions

  The following 3 scales represent the physician's process skills in initial interviews:
  4. structuring the interview
  5. interpersonal skills
  6. communicative skills

  This third blueprint was pretested by a broader, expert panel of general practitioners, psychiatrists, psychologists and sociologists, all faculty staff charged with education and evaluation of interviewing skills. These pretests consisted of the following procedures:

- the expert panel was invited to comment on the item domain, the item format, the definitions of the interviewing skills and the criteria for scoring described in the manual for observers;
- the expert panel was asked to score 2 test-videotapes of interviews of about 20 minutes: one simulated a patient presenting a somatic problem, another simulated a mental health problem. The panel members were expected to discuss their scores and to attain a consensus in two sessions. During these sessions guided by 2 members of the "founding panel", comments and scoring problems were monitored. These sessions resulted in several adaptations of the third blueprint which finally became MAAS-General Practice (see appendix A).

4.5.2 Steps involved in the construction of the MAAS-Primary Mental Health Care:

- the third blueprint of the MAAS-General Practice was taken as a starting point. The "founding panel" then constructed items mainly based on the content aspects of initial medical interviews in primary mental health care. This content dimension is described in 2.4.2.2. This procedure resulted in the addition of 2 content-specific scales, "socio-emotional exploration" and "psychiatric examination", to the original 6 scales of the MAAS-GP.
- the resulting blueprint was judged by another "broader" panel consisting of about 15 experts in the field of primary mental health care (social psychiatrists, psychologists, social workers, general practitioners). Moreover, these experts are also educators and researchers in their roles as faculty members. Their efforts resulted in adaptations of the item domain, resulting in the final version of the MAAS-PMHC. The items of 8 scales of the MAAS-PMHC are extensively described in the following paragraphs (see also appendix A).

4.6 Description of the items of the MAAS-General Practice.

In this section we describe the 6 scales of this method: "exploration of the reasons for encounter" (4.6.1), "history-taking" (4.6.2), "presenting solutions" (4.6.3), "structuring the interview"

(4.6.4) and "interpersonal and communicative skills" (4.6.5). The theoretical material which furnishes the item domain is derived from chapter 2, where the process and content dimensions of these interviewing skills are described.
The items of the 6 scales and their criteria for scoring are described in appendix B.

4.6.1 Exploration of the reasons for encounter.

In this phase the physician gives the patient the opportunity to describe his complaints and symptoms in his own words. He expands on the causes and consequences of the complaints and the events which triggered the visit to the physician. Further questions may be asked about attempted solutions and about discussions of the complaints in the primary group.
The appropriate process aspects in this phase are open questions, probes into the patient's frame of reference, active listening, emotional reflection, stimulating summarizations. These process skills, summarized in the term "exploration", are assessed in the scales "interpersonal and communicative skills" (see below).
The items in the scale "exploration of the reasons for encounter" belong to a conjunctive category, combining speech-acts and content elements.
The items are edited in a format such as "Asks for ... (content topic)" or "Explores ... (content topic)". Scoring is on a two- point scale: present/absent. The scoring "present" should be given, when - according to the criteria specified in the manual for observers - the pertinent topic is asked or explored.

4.6.2 History-taking.

During this phase of the interview, the physician asks the patient questions from his medical frame of reference in order to collect information for his diagnostic and clinical reasoning process. The process skills used in this phase are mainly closed and directive questions, sometimes in a short series. These process skills are measured on the scales "interpersonal and communicative skills". The content of the items reflects questions about aspects of the

complaints/problems: description of the nature of the complaint, intensity, localization, course through day-time, etc. Psychosocial factors are also operationalized in the items: questions about psychological functioning, quality of interpersonal relationships etc. The 22 items of this scale belong to a conjunctive category of speech-act (mainly types of questions) and content elements. Format and scoring of the items is similar to the previous scale (4.6.1).

4.6.3 Presenting solutions.

This phase follows both previous phases and - if carried out - the physical examination. The physician informs the patient about his condition or problem, causes and prognosis of his disease. He then proceeds with an exploration of the patient's feelings, evoked by this information. A negotiation of the problem definition between physician and patient may ensue. The physician then makes a proposal for follow-up: further exploration or investigations, referral, treatment, preventive advice. Alternative proposals may be given by the physician and again negotiation may follow. Finally, the physician gives concrete advice based on the outcome of the negotiation process. The physician concludes with appointments for follow-up. The 12 items describing the skills during this phase, belong to a conjunctive category of speech-act and content elements and may take various formats: "provides information about (medical information)", "discusses (medical information)", "explains the effects of (medical information)", "explains why (medical information)". The scoring is on a two-point scale: present/absent.

4.6.4 Structuring the interview.

This scale, comprising of 8 items, is intended to measure the skills by which the physician opens and closes the interview, by which he sets an agenda, and by which he links the afore-mentioned three phases. These items consist of conjunctive categories combining a complex interviewing skill with content elements. Example: begins the "presenting solutions" phase with the provision of information on the problem definition or diagnosis. The scoring is on a two-point scale: present/absent.

4.6.5 Interpersonal and communicative skills.

These skills, which are not connected to a specific phase, are not easy to operationalize in concrete behavior and to furnish with criteria, which result in reliable scoring.

Interpersonal skills are operationalized in 8 items. These items in pertain particular to those interviewing skills by means of which the physician approaches the emotional aspects of the interview.

The items of the scale "communicative skills" are rooted in the interviewing skills by means of which the physician starts and maintains the information-flow from and to the patient. The format of these items is a three-point evaluative rating scale. In general, in every item, a number of criteria should be fulfilled. Items are scored with "no", "indifferent" and "yes" in proportion to the number of criteria fulfilled.

These criteria, which can be either qualitative or quantitative yield two types of items.

a. Items with qualitative criteria are scored by application of each criterion to the whole interview. The number of criteria positively fulfilled determines the ultimate score, "yes", "indifferent" or "no".

   For instance, the item on facilitation has 5 global criteria, judging the interview as a whole, as to:

   - quality of the open questions
   - presence of active listening
   - quality of probing within the patient's frame of reference
   - facilitative self-disclosure
   - minor, stimulating remarks

   The quality of each criterion is judged according to the effect on the facilitation of the patient to tell his own medical story. The ultimate scoring is "yes" when 4 or 5 criteria are fulfilled; "indifferent" in the case of 2 or 3 criteria and "no" in the case of fewer than 2. This kind of scale calibration is called "behavioral anchoring". It is claimed that it enhances reliability in global rating scales (Streiner, 1985).

b. Items with quantitative criteria are scored in a manner best illustrated by an example: e.g. the item "Uses closed-ended questions in a proper way". Every closed-ended question is judged against the criteria in the manual, resulting in a count of a proper or improper closed-ended question. The score "yes", "indifferent" or "no" is given when resp. 80% or more, 60-80% or 60% or less of the closed-ended questions are used in a proper way.

4.7 Description of the items of the MAAS-Primary Mental Health Care.

Broadly speaking the MAAS-Primary Mental Health Care consists of the MAAS-General Practice extended by 2 scales: "psychiatric examination" and "socio-emotional exploration". These scales contain items which are content-specific for primary mental health care. The theoretical base of these items is stated in chapter 2. Moreover, some new items are added to the other scales and some items have been re-allocated from one scale to another. Format and scoring of items is similar to those used in the MAAS-GP. We therefore only briefly discuss the 8 scales of the MAAS-PMHC (see for the items and the criteria of scoring appendix B).

4.7.1 Exploration of the reasons for encounter.

In comparison with the MAAS-GP this scale has been extended by 5 items to a total of 13 items.
First, the item "Asks the patient to describe his complaints/problems" has been re-allocated from the scale "history-taking". The reason for this re-allocation lies in the fact that in mental health problems' symptoms, complaints and problems are - in the patient's perspective- strongly interwoven. Separation is often considered as rather artificial. An item about "recent life events" has also been added. This gives an impression of the intensity of stressful "life events".
In addition two items concerning the patient's problem-solving or coping mechanisms and his wishes regarding future changes have been constructed. Finally, an item on the impact of the patient's problem/complaint on members of his primary group has been added.

4.7.2 History-taking.

This scale has been reduced to a total of 13 items. The items with psychosocial content have been relocated to the new scale "socio-emotional exploration". This scale, in combination with the following two scales "psychiatric examination" and "socio-emotional exploration" are for measuring the extent, to which the physician more of less systematically "scans" the psychosocial domain in order to generate explanatory and action hypotheses with respect to the patient's problem.

4.7.3 Psychiatric examination.

This scale has 6 composite items which reflect the physician's interviewing skills pertaining to the collection of information for the psychiatric examination, i.e. the symptoms and signs level. These six composite items cover the important psychiatric diagnostic groups in primary mental health care: affective disorders, anxiety related disorders, disorders characterized by disturbances in thought and sensory perception and disturbances in memory. The scoring of these items are elucidated by the following example:

"Explores anxiety"

a) character and intensity	yes/no
b) anxiety (fear) of objects	yes/no
c) anxiety provoking or releasing factors	yes/no
d) consequences of anxiety	yes/no

A total score of the "sub-items" a) to d) represents a measure of the depth to which the symptom anxiety has been explored.

4.7.4 Socio-emotional exploration.

This scale has 20 items. Its content reflects a broad area of socio-emotional functioning. The items pertain to the emotional functioning of the patient, norms and values in taking responsibility, exploration of relationships in family or primary groups, social support, functioning profession and education, cultural conflicts, financial and housing situation, substance (ab)use, developmental issues until adolescence. The items (scored on a 2-point scale yes/no) require a certain exploration in depth - according to the criteria stated in the manual - in order to be scored with "yes".

4.7.5 Presenting solutions.

This scale has 15 items and is an extended version of the similar scale in the MAAS-GP. Four new items have been added: concerning the degree of responsibility the patient will take for his treatment; concerning the patient's opinion of the proposed help; concerning the impact of "important others" on the proposed help; concerning the opportunity the physician has provided for making a choice between proposed alternative solutions. One item has been removed from the similar scale of the MAAS-GP, because its content is covered by the newly included items. This extention arises from the concept of "psycho-role" (Siegler et al., 1976).

4.7.6 Structuring the interview.
Interpersonal and communicative skills.

These scales are the same as those used in the MAAS-GP. Their similarity is based on the discussion in 2.1.

4.8 Variants and extensions of the MAAS-GP and MAAS-PMHC.

Starting from the item domain of the MAAS-GP and MAAS-PMHC, we also constructed methods which can be used by the interviewer as a self-evaluation method. The objectives of their construction are threefold.

First, such methods may serve to provide feedback during education. Second, they can be used for evaluative purposes. Third in validity research, we would like to answer the question whether the same underlying concepts of interviewing skills are measured by both the interviewers themselves and observers alike. Two self-evaluation methods are presented: one for behavioral self-assessment, the MAAS-SELF (4.8.1.1), and one global rating scale, the Global Self-rating Scale (4.8.1.2).

The Global Expert-Rating Scale (4.8.1.3) of medical interviewing skills was then developed.

Moreover, the MAAS-GP and MAAS-PMHC have been extended by checklists which register the patient's responses to the physician's questions or topics raised on the patient's initiative. They are described in 4.8.2.

We start, however in 4.8.1 with theoretical considerations of self-evaluation and global rating scales.

4.8.1 Theoretical considerations of self-evaluation and global rating scales.

The capacity for self-evaluation has long been considered as a hallmark of professionals, its main objective being changed behavior of the self-raters (Arnold, 1982). In our case, an improvement in medical interviewing skills may be expected. These expectations are based on data from literature on behavioral modification (Mahoney et al., 1973; Kazdin, 1974).

Self-evaluation is considered as measuring non-cognitive abilities such as interviewing skills, attitudes, rapport with the patient, rather than cognitive aspects such as medical knowledge (Arnold et al., 1985).

The reliability of self-evaluation methods seems to be acceptable according to a test-retest method (Linn et al., 1975). Validity research reveals, that self-evaluations of medical students and residents are generally lower than the ratings which they receive from faculty and peers (Morton et al., 1977; Stuart et al., 1980), but controversial findings are also reported. Sclabassi et al. (1984) report, for example, that medical students over-estimated as well as under-estimated their own knowledge and skills after an anaesthesia clerkship. In addition, self-evaluation ratings show greater variations within competence dimensions than faculty ratings (Stuart et al., 1980). Therefore, in factoring self-evaluation ratings, a "picture" which is differentiated in more competence aspects arises in comparison with faculty ratings (Kolm et al., 1985). Finally, students' self-evaluations correlated with concurrent grades, faculty assessments of the students and peer ratings at a modest yet significant level (Morton et al., 1977).

It is suggested that the accuracy of self-evaluation can be enhanced by practice and experience with self-evaluation and by the use of unambiguous (behavioral) criteria for assessment (Stuart et al., 1980).

The second topic of this section are the global rating scales which are widely applied to the measurement of medical competency because of some major advantages. Their ability to tap "soft" areas, their

unobtrusiveness, their low cost of development and application as well as their potential for feedback are frequently mentioned (a.o. Streiner, 1985). Reliability and validity studies of these scales, though relatively scarce, also show disadvantages of this method. Inter-rater reliability is often low, because raters observe different behavior or experiences, rate on different criteria and have difficulties in agreeing on the meaning of particular numerical scores (Levine et al., 1975). In addition, most global rating scales consist of multiple scales, each measuring a separate dimension, but each having a low reliability (Dielman et al., 1980). Finally, global rating scales suffer from halo-effects which hamper a multi-faceted differentiated judgment of subjects. It therefore seems un-likely that raters can accurately assess more than two dimensions of performance (Streiner, 1985).

The validity of global rating scales is generally hampered by this lack of reliability which always sets its upper limit to validity. However, concurrent and predictive validity compared with other measures of clinical proficiency is generally disappointing (a.o. Donnelly et al., 1978; Streiner, 1985).

In his review article Streiner (1985) makes several recommendations for the improvement of the reliability and validity of global rating scales. First, rating scales should be provided with behavioral anchors: points on the continuum of the scale ought to be linked to behavioral criteria. Second, the fineness of the scale should be chosen on the basis of the rater's ability to discriminate between levels of performance. Third, the raters should be trained in the use of the scale, the definitions of the terms and different points of the continuum on the scale. Fourth, Ebel (1951) recommends, on statistical grounds, an average rating of more raters in order to improve reliability.

We take these recommendations into account in the construction of the Global Self-Rating Scale (4.8.1.2) and the Global Expert-Rating Scale (4.8.1.3).

4.8.1.1 The MAAS-Self-evaluation in General Practice and in Primary Mental Health Care.

This method has the same format as the MAAS-GP and the MAAS-PMHC. Its items are re-edited in an "I-format". For instance, the item "Asks the patient what attempts he has made to solve the problem" has been re-edited into "I asked the patient what attempts he has made to solve the problem". These methods are used in the validity studies. The complete methods can be found in the appendices C and D.

4.8.1.2 and 4.8.1.3 The Global Self-Rating Scale and Global Expert-Rating Scale.

The items of these rating scales corresponds to the scales of the MAAS-GP and MAAS-PMHC. They have the following format, e.g. "The physician (or I) adequately performed the exploration of the reason for encounter", e.g. the complaints and their meaning for the patient have been elucidated". Both methods are used in the convergent and divergent validity study with the MAAS described in chapters 8 and 12. They are also expanded on in the theoretical content of the restricted number (7-8) items of these global rating scales. They are given in the appendices E and F.

4.8.2 The checklists "obtained information".

Checklists, reflecting the content of the patient's utterances during initial medical interviews, have been constructed. In these checklists, items are constructed that reflect the content of the questions stated in the "original" items of the MAAS-GP and the MAAS-PMHC.

As an illustrative example we take from the scale "exploration of the reasons for encounter" the item "Asks the patient what attempts he has made to solve the problem". This item corresponds in the checklist "obtained information" with an item "own attempts to solve the problem".

These checklists "obtained information" have been constructed parallel to the scales "exploration of the reason for encounter", "history-taking", "socio-emotional exploration".

For scoring of these items, a 2-point scale is used (present/absent). The criteria for scoring are dependent on the research or evaluation situation in which the MAAS is used.

When used in field settings, these items should be rather open, specifying the content in general terms. In this thesis, however, the MAAS is used with simulated patients whose roles are programmed in detail. This circumstance allows for a more circumscript definition of the content aspect of the items. Referring to the above described example this item of the patient's side may be formulated - according to the pertinent role of the simulated patient - : "I took my sleeping pills and drank a glass of hot milk" (in a case of sleep-disturbance). In this example, both points, "sleeping pills" and "a glass of hot milk" should be mentioned by the patient in order to be scored "present". When the conjunction "or" is used between the two objects, then the stating of one of the two objects is sufficient to score "present".

The purpose of these checklists "obtained information" is to measure:

- directly: the quantity and quality of the information obtained by the physician
- indirectly: it may yield information about initiatives taken by the patient, talkativeness of the patient, etc.

These checklists are used in a content validity study of the MAAS-PMHC, described in chapter 11.

4.9 Use of the MAAS.

The MAAS-GP and MAAS-PMHC can be used in educational evaluation (4.9.1) and theoretically orientated research into the properties of medical interviews and medical interviewing skills (4.9.2).

We end this section with a description of the training of observers, a necessary condition for use of the MAAS (4.9.3).

4.9.1 Use of the MAAS for evaluation in educational settings.

The MAAS is used in formative evaluation in order to give an assessement to the students of their interviewing skills and to recommend improvements where needed. These evaluations are made by peers, experts or by the student himself (and do not lead to decisions

concerning study progress). In the last instance, the MAAS-SELF may be used. Students can use the MAAS serving as a taxonomy of the skills which will be trained during the medical curriculum.
The MAAS is also used in summative evaluations. These examinations have consequences for the students' progress through the medical curriculum. In such instances, the MAAS with its extensions "obtained information checklist" can be used as a medical interviewing test. For test purposes, the MAAS-GP can also be used in abbreviated versions. First, a selection of items fitting the Rasch model (see chapter 7 and 10) can be taken. Second, items may be randomly selected from the whole item domain of the MAAS-GP. That selections are previously unknown to the student is vital, otherwise the student could arrange his interviews in such a manner, that they would display behavior leading to artificially increasing his scores on the MAAS, at the cost of validity.
As no internal criterion of an insufficient or sufficient interview is available, we had recourse to group-referenced measurement, taking the student's own peer group as reference (Wijnen, 1971). The critical threshold is the mean of the summed scores of the group's members minus one standard deviation. Scores below this threshold are considered as insufficient.

4.9.2 Use of the MAAS for research purposes.

In addition to reliability and validity reseach with simulated patients, as extensively treated in this thesis, the MAAS is also used in field settings, for example in general practice (see also chapter 5). The mode of use of such naturalistic studies is similar to that with simulated patients. Such studies are, however, beyond the scope of this thesis.

4.9.3 Training of observers.

To obtain reliable scores it is necessary for observers to undergo initial and, later "a refresher" training.
The initial training takes two 3-hour sessions and encompasses the following activities.

- Preparatory reading of the manual for observers which consists of a short introduction of the MAAS (construction, theoretical background, measurement rationale), a list of definitions of interviewing skills and a list of criteria for scoring per item.

- Discussion of this documentation during the first session.
- Use of the MAAS in scoring a videotaped interview, whilst this videotape is played for periods of 1 or 2 minutes. After every period, possible scores are discussed.
- During the second 3-hour session, one videotaped interview is scored in periods of 5 minutes and discussed as to the possible scores.
- Finally, videotaped interviews are viewed in their entity and scored afterwards. Results are compared and discussed after scoring.

"Refresher" training is needed when the observer has not used the MAAS for a period of two months. A quick re-reading of the list of definitions and of the criteria for scoring of each item as well as the scoring of one test videotape with discussion afterwards is necessary. The required time does not exceed one hour's preparation.

The scoring for the MAAS-GP and MAAS-PMHC does not take more than 20 minutes of a trained observer. About the half of the items can be scored by an observer, during the interview, and the second half after the interview. Thus the entire time spent by the observer in scoring consists of the duration of the interview extended by approximately 10 minutes.

Table 4.2 Twenty two methods measuring medical interviewing skills reviewed according to pre-set criteria.

AUTHORS OF MEASUREMENT METHODS FOR INTERVIEWING SKILLS	METHODS AND CATEGORIES OF MEASUREMENT	FOCUS OF MEASURE-MENT	TYPE OF MEDICAL INTERVIEW	TEACHABLE SKILLS MEASURED IN USED IN EDUCATIONAL RESEARCH	RELIABILITY STUDIES	VALIDITY STUDIES	PRACTICA-BILITY
Bales (1950)	- complete coding in 12 speech act categories	physician and patient	all	not/not used	reasonable inter-rater reliability	predictive	low
Hollifield e.a. (1957)	- 8 evaluative and affective rating scales	physician	initial clinical interviews	hardly to moderate/ not used	good inter-rater reliability	---	high
Adler e.a. (1966)	- incomplete coding of 27 "general psycho-therapeutic interventions" and patient responses ("complex speech-acts")	physician and patient	psycho-therapeutic and psychia-tric inter-views	moderately/ used	---	---	low
Barbee e.a. (1967)	- 9 evaluative and affective rating scales - 12 content cate-gories	physician and patient	initial esp. for educa-tional settings	hardly/used	reasonable inter-rater reliability	---	high

Hess (1969)	- rating/complete coding of 11 speech-act cat. (not Bales') - 14 evaluative rating scales	physician (student)	interview in the "general medical clinic"	moderately/ not used	moderate inter-rater reliability	---	moderate to low
Freeman e.a. (1971)	- Bales' system - various content categories	physician and patient	initial, pediatric interviews	moderately/ not used	reasonable inter-rater reliability	predictive	low
Jarrett e.a. (1972)	- evaluative rating of 23 conjunctive (speech-act and content) categories	physician	initial, psychiatric interviews	moderately to very well/not used	good inter-rater reliability	descriminant	high
Roter (1977)	- complete coding of 10 speech-act cat. (modification of Bales') - 4 affective rating scales	physician and patient	all medical interviews	hardly/not used	reasonable inter-rater reliability	predictable	low
Van Dorp (1977)	- incomplete coding of 20 speech-act cat. and more complex interviewing skills - 16 evaluative and affective rating scales		all medical interviews	moderately to well/not used	reasonable inter-rater reliability and generalizability coefficients	content, convergent and predictive	high

Table 4.2: (continued)

AUTHORS OF MEASUREMENT METHODS FOR INTERVIEWING SKILLS	METHODS AND CATEGORIES OF MEASUREMENT	FOCUS OF MEASURE-MENT	TYPE OF MEDICAL INTERVIEW	TEACHABLE SKILLS MEASURED IN USED IN EDUCATIONAL RESEARCH	RELIABILITY STUDIES	VALIDITY STUDIES	PRACTICA-BILITY
Brockway (1978)	- evaluative rating of 21 complex interviewing skills and - 35 conjunctive (speech-act and content) cat. and - 4 non-verbals	physician	initial medical interviews	very well/ not used	reasonable inter-rater agreement	convergent/ divergent	moderate to high
Stiles (1978)	- complete coding of 64 speech-act cat.	physician and patient	all medical interviews	not/not used	reasonable inter-rater reliability	predictive	low
Hill (1978)	- incomplete coding of 14 counseling skills	physician (counsellor)	counseling	moderately to well/ used	inter-rater reliability moderate to high	---	low

Barsky e.a. (1980)	- 19 evaluatative and affective rating scales of conjunctive (content and complex interviewing skills) categories	physician	initial interviews in primary health care	moderately/ not used	substantial inter-rater reliability	---	high
Stillman (1980)	- 16 evaluative and affective rating scales	physician	all initial medical interviews	moderately/ used	reasonable inter-rater reliability and internal consistency	construct validity	high
MacDonald e.a. (1981)	- 29 "Bales'-like" speech-act categories	physician and patient	all medical interviews especially educational settings	hardly to moderately/ used	reasonable to good inter-rater reliability	---	high with micro-computer program
Rutter e.a. (1981)	- mixed system of incomplete coding and activity counts of interviewer activity, directiveness, questioning, handling feelings, non-verbals - evaluative ratings of content and ratings of 5 affective dimensions	physician and patient	initial interviews in child psychiatry	moderately to well/ not used	reasonable to good inter-rater reliability	---	moderate to high

Table 4.2: (continued)

AUTHORS OF MEASUREMENT METHODS FOR INTERVIEWING SKILLS	METHODS AND CATEGORIES OF MEASUREMENT	FOCUS OF MEASURE-MENT	TYPE OF MEDICAL INTERVIEW	TEACHABLE SKILLS MEASURED IN USED IN EDUCATIONAL RESEARCH	RELIABILITY STUDIES	VALIDITY STUDIES	PRACTICA-BILITY
Goldberg e.a. (1981)	- 14 simple and complex speech-act categories	physician	initial interviews in primary mental health care	hardly/ used	good inter-rater reliability	---	high
Mokkink e.a. (1982)	- 50 evaluative rating scales of various aspects of patient management in general practice		initial medical interviews	moderately/ used	moderate to reasonable test-retest reliability	---	high
Den Hoed e.a. (1982)	- 22 evaluative rating scales of complex inter-viewing skills - 4 affective rating scales - coding and rating of various descrip-tive variables of the consultation process		initial medical interviews in general practice	moderately/ used	good inter-rater reliability	convergent validity with expert judgements	high

Sprunger (1983)	- complete coding of 4 speech-act cat. - complete coding in 8 major content cat.	physician	prenatal interviews in pediatrics	moderately/ not used	good inter-rater agreement	---	moderate rate
Pieters e.a. (1983)	- 12 evaluative rating scales of interviewing skills - 12 evaluative scales of patient management		initial medical interviews in general practice	moderately/ used	good inter-rater reliability	---	high
Mumford e.a. (1984)	- 12 evaluative rating scales of complex interviewing skills - 6 content rating of grouped topics	physician	initial medical interviews especially in educational settings	moderately/ used	reasonable inter-rater reliability and internal reliability	predictive	high

REFERENCES

Adler LM, Enelow AJ. An instrument to measure skill in diagnostic interviewing: a teaching and evaluation tool. Journal of Medical Education, 1966; 41: 281-288.

Arnold L. Self-evaluation in undergraduate and graduate medical education. Proceedings of the 20th Annual Conference on Research in Medical Education. Washington DC, 1981.

Arnold L, Willoughby TL, Calkins EV. Self-evaluation in undergraduate medical education: a longitudinal perspective. Journal of Medical Education, 1985; 60: 21-28.

Bales RF. Interaction Process Analysis. Addison Wesley, Cambridge, 1950.

Barbee RA, Feldman S, Chosy LW. Quantitative evaluation of student performance in the medical interview. Journal of Medical Education, 1967; 42: 238-243.

Barsky AJ, Kazis LE, Freiden RB, Goroll AH, Hatem CJ, Lawrence RS. Evaluation of the interview in primary care medicine. Social Science in Medicine, 1980; 14A: 653-658.

Beekers M. Interpersoonlijke vaardigheidstherapieën voor kansarmen. Diss. Swets en Zeitlinger, Lisse, 1982.

Brockway BS. Evaluation of the physicians competency: what difference does it make? Evaluation and Program Planning, 1978; 1: 211-220.

Cone JD, Hawkins RP (Eds.). Behavioral Assessment. Brunner and Mazel, New York, 1977.

Dielman TE, Hull AL, Davis WK. Psychometric properties of clinical performance rating scales. Evaluation and the Health Professions, 1980; 3: 103-117.

Donnelly MB, Gallagher RE. A study of the predictive validity of patient management problems, multiple choice tests and rating scales. Proceedings, 17th Annual Conference on Research in Medical Education, Washington DC, 1978.

Dorp C van. Luisteren naar patiënten; een analyse van het medisch interview. De Tijdstroom, Lochem, 1977.

Ebel RL. Estimation of the reliability of ratings. Psychometrika, 1951; 16: 407-424.

Freemon B, Negrete VR, Davis M, Korsch BM. Gaps in doctor-patient communication: doctor-patient interaction analysis. Pediatric Research, 1971; 5: 298-311.

Goldberg D, Steele JL, Smith C, Spivey L. Training family practice residents to recognize psychiatric disturbances. Final Report. Dept. of Psychiatry, Biometrics and Family Practice, Medical University of South Carolina, 1980.

Hess JW. Methods for evaluating medical students skills in relating to patients. Journal of Medical Education, 1969; 44: 934-938.

Hill CE. Counselor verbal response category system. Journal of Counseling Psychology, 1978; 25: 461-468.

Hoed FE den, Sluys EM. Het meten van "Methodisch Werken". NHI Utrecht, 1982.

Hollifield G, Rousell CT, Bachrach AJ, Pattishall EG. A method of evaluating student-patient interviews. Journal of Medical Education, 1957; 32: 853-857.

Inui TS, Carter WB, Kukull WA, Haigh VH. Outcome-based doctor-patient interaction analysis. I. Comparison of techniques. Medical Care, 1982; 22: 537-549.

Jarrett FJ, Waldron JJ, Borra P, Handforth JR. The Queen's University Interviewer Rating Scale (QUIRS). Canadian Psychiatric Association Journal, 1972; 17: 183-188.

Kazdin AE. Reactive self-monitoring: the effects of response desirability, goal setting and feedback. Journal of Consulting and Clinical Psychology, 1974; 42: 704-716.

Kolm P, Verhulst J. Comparing self and supervisor evaluation. A different view. Proceedings of the 25th Annual Conference on Research in Medical Education, Washington DC, 1985.

Lazare A, Eisenthal s, Wasserman L. The customer approach to patienthood. Attending to patients requests in a walk-in clinic. Archives of General Psychiatry, 1975; 32: 553-558.

Levine HG, Gustavson LP, Emergy JLR. The effectiveness of various assessment clerkship. Proceedings, 14th Annual Conference on Research in Medical Education, Washington DC, 1975.

Linn BS, Arostegui M, Zeppa R. Performances rating scale for peer and self assessment. British Journal of Medical Education, 1975; 9: 98-101.

MacDonald M, Templeton B. Interpersonal skills assessment technique (ISEE-81). National Board of Medical Examiners. Philadelphia, 1981.

Mahoney MJ, Moura HGM, Wade TC. Relative efficacy of self-reward, self-punishment and self-monitoring techniques for weight loss. Journal of Consulting and Clinical Psychology, 1973; 40: 404-407.

Mischel W. Personality and assessment. Wiley, New York, 1968.

Mokkink H, Smits A, Grol A. Prevara: een observatie-instrument voor het handelen van de huisarts in the kader van processen van somatische fixatie. Nederlands Tijdschrift voor Psychologie, 1982; 37: 35-50.

Morton JB, MacBeth WAAG. Correlations between staff, peer and self assessments of fourth-year students in surgery. Medical Education, 1977; 11: 167-170.

Mumford E, Anderson D, Cuerdon T, Scully D. Performance-based evaluation of medical students' interviewing skills. Journal of Medical Education, 1984; 59: 133-135.

Pieters HM, Jacobs HM. Hulpverlening van huisartsen in opleiding getoetst; een gedetailleerde consult observatie. Medisch Contact, 1983; 38: 1539-1542.

Roter DL. Patient participation in the patient-provider interaction: the effects of patient question asking on the quality of interaction, satisfaction and compliance. Health Education Monographics, 1977; 5: 281-315.

Rutter M, Cox A. Psychiatric interviewing techniques: I. Methods and measures. British Journal of Psychiatry, 1981; 138: 273-282.

Sclabassi SE. Development of self-assessment skills in medcial students. Medical Education, 1984; 18: 226-231.

Siegler M, Osmond H. Models of madness, model of medicine. Harper and Row, New York, 1976.

Sprunger LW. Analysis of physician-parent communication in pediatric prenatal interviews. Clinic Pediatrics, 1983; 22: 553-558.

Stiles WB. Verbal response modes and dimensions of interpersonal roles: a method of discourse analysis. Journal of Personality and Social Psychology, 1978; 36: 693-703.

Stiles WB, Putnam SM. Classification of medical interview coding systems. Paper presented at the International Conference on Doctor-patient Communication. 16-18 Sept. 1986, London, Ontario.

Stillman PL. Arizona Clinical Interview Rating Scale. Medical Teacher, 1980; 2: 248-251.

Streiner DL. Global rating scales. In: Neufeld VR, Norman GR (Eds). Assessing clinical competence. Springer Publ. Cie., New York, 1985.

Stuart MR. Goldstein HS, Snope IC. Self-evaluation by resident on family medicine. Journal of Family Practice, 1980; 10: 639-642.

Swanson DB, Mayewski RJ, Norsen L, Baran G, Mushlin AI. A psychometric study of measures of medical interviewing skills. Proceedings of the 20th Annual Conference on Research in Medical Education, 1981: 308.

Thorndike RL. Applied psychometrics. Houghton Mifflin Cie., Boston, 1982.

Wijnen WHF. Onder of boven de maat. Een methode voor het bepalen van de grens voldoende/onvoldoende bij studietoetsen. Diss. Swets en Zeitlinger, 1971.

CHAPTER 5 ASSESSING INSTRUMENTAL UTILITY: ISSUES OF VALIDITY, RELIABILITY AND SCALABILITY

A.A.M. Crijnen and H.F.Kraan

5.1 Instrumental utility.

After constructing a measurement method of physicians' medical interviewing skills the question is raised: what is the significance of these measurements? Up to this point, we combined personal experience, opinions from interested physicians and psychologists and research evidence to construct the Maastricht History-taking and Advice Checklist. This intensive and continuing exchange of ideas formed a prerequisite for test development, but was not expected to lead automatically to an appropriate measurement of the concepts under study. We therefore wished to study the instrumental utility of our measurement methods of medical interviewing skills to see how well the operationally-defined, empirical measurement methods assess the intended concept (de Groot, 1972).

Instrumental utility can be expressed in four parameters: validity, reliability, scalability and practicability. A group of variables constituting a measurement of the concept under study is assumed to represent the intended concept acceptably and adequately, and to measure it with considerable precision and efficiency. The study of validity addresses the question of the quantitative and empirical relationship between the concept under study and the variables which represent this concept (de Groot, 1972). Reliability and scalability contribute to the quality of measurement in terms of precision and efficiency and, subsequently, to the assessment of validity. Practicability refers to the convenience with which the instrument can be applied in research or educational settings. In this thesis, the validity, reliability and scalability of MAAS-GP and MAAS-PMHC is studied. The practicability of our measurements will not be analysed but we report some of our experiences below.

The question is now raised of how to establish the instrumental utility of our measurements of medical interviewing skills in terms of validity, reliability and scalability. In order to develop adequate

research designs, we reviewed methodological and psychometric theories of validity, reliability and scalability which are described in the following paragraphs. Subsequently, an overview of research settings that were used to answer the formulated questions is given.

5.2 Validity.

Recently, Van Berkel (1984) drew up an inventory of 77 distinct types of validity and classified them into four major categories (Cronbach et al, 1955; Cronbach, 1970; de Groot, 1972; Cook et al, 1979; Thorndike, 1982). These four categories of validity are

1. criterion-orientated validity, which correlates results of a test with a criterion outside the test situation;
2. content validity, which refers to how adequately the content of the test represents the universe that the test intends to measure;
3. construct validity, which analyses the meaning of test scores in terms of psychological constructs;
4. experimental validity, which studies the generalizability of conclusions derived from experiments to situations outside the experimental setting.

Inferences from the first three validity types are based on what a subject achieves on the pertinent and related tests, whereas inferences from the last validity type are based on a critical appraisal of the design of the test setting. Philipsen (1984) approached the issue of validity slightly differently by differentiating between two dimensions in validity studies. Firstly, he recognized the goals researchers try to achieve ranging in hierarchical order from face-validity, content validity to construct validity. Secondly, he differentiated between the procedures which can be applied as predictive validity, discriminant validity or concurrent validity. By combining both dimensions, nine types of validity are discerned. Philipsen's contribution emphasizes the depth of analysis with regard to each procedure. As textbooks are organized around the four major types of validity mentioned by van Berkel, they are elaborated on in the following paragraphs, but we keep in mind that the depth of analysis can vary for each procedure.

5.2.1 Criterion-orientated validity.

In criterion-orientated validity, the question is studied of how well test scores are able to predict criterion performance. Criterion-orientated validity is sometimes called "concurrent validity" when no time has elapsed between the measurements, or "predictive validity" when a criterion is predicted for the future.

Criterion-orientated validity is primarily applied to tests which are used to select or classify subjects such as students, patients or employees; to tests which are used to decide what treatment should be given to subjects, and to tests which are used as a substitute for a more cumbersome assessment procedure (Cronbach, 1970). Criterion-orientated validity is operationalized by the correlation between test performance and future criterion performance. High or modest correlations confirm the criterion-orientated validity of a test. Criterion-orientated validity only provides firm evidence of validity when the measurements are intended as predictors in a specific research setting with a specified criterion which is measured validly in itself (de Groot, 1972). The construction of criterion measurements forms the greatest problem for predictive validity since it is difficult, but vital, to obtain suitable, valid measurements of the criterion. Difficulties may arise when the criterion behavior is multi-dimensional, vague or equivocal. Theoretical considerations of the relationship between predictor and criterion are essentially unimportant in the determination of criterion-orientated validity.

Although we acknowledge the importance of criterion-orientated validity, especially for selection or treatment purposes, we do not apply it to the MAAS. An adequate research design for the determination of the criterion-orientated validity or, more precisely, the predictive validity, of the MAAS, would be to record students'/future physicians' medical interviewing skills by means of the MAAS at this moment and then correlate them with measurements of their interviewing skills recorded after the physicians have had several years of experience in daily practice (Crijnen et al, 1984). The strength of the correlations would indicate MAAS's criterion-orientated validity. Since the time-lag necessary to obtain the future criterion measurements is beyond the scope of our project, we have been unable to study the predictive

validity of the MAAS. Determination of the MAAS's criterion-orientated validity should certainly be carried out in the future as the MAAS is already used to classify and select medical students.

The establishment of criterion-orientated validity in terms of concurrent validity is elaborated in the sections on construct validity since theoretical considerations are strongly taken into account.

5.2.2 Content validity.

Content validity is studied when researchers evaluate whether the items of the test adequately represent the universe the test intends to measure. Determination of content validity is especially required when tests are designed to measure the degree of mastery of some domain of knowledge or skill. Unfortunately, objective determination of a test's content validity is difficult since few attempts are made to develop quantitative indices of content validity. Content validity is determined by judgment of experts on how adequately the domain of interest is represented in the test. A prerequisite for the assessment of content validity is a clear, detailed and explicit definition of the universe a researcher wishes to measure (see also chapter 2, 3 and 4). This definition ought to cover the kind of tasks and situations covered by the universe; the kinds of responses the observer wishes to count; and the instruction to the subjects (Cronbach, 1970).

During the construction of the Maastricht History-taking and Advice Checklist, a large group of physicians and psychologists continually scrutinized the content of the MAAS and thereby enhancend content validity. In this thesis, we have attempted to define as clearly as possible the context in which we are interested (initial medical consultations), the behavior we have tried to measure (six different categories of medical interview behavior) and the task given to the subjects (perform a medical interview). A second contribution to content validity is provided by the theoretical considerations of medical interviewing skills and empirical evidence which were used to construct the scales. Since content validity was secured during test construction, no systematic efforts have been made or will be made to objectify content validity of the MAAS.

5.2.3 Construct validity.

In 1955, Cronbach and Meehl recommended that the construct validity of new tests should be established in addition to the criterion-orientated validation procedure which was used at that time but was severely criticized. They defined "construct validity" as the analysis of the meaning of test scores in terms of psychological concepts or constructs. In interpreting test scores, researchers have to face the question: what constructs account for variance in test performance? Constructs were seen as 'some postulated attributes of people assumed to be reflected in test performance' (Cronbach et al, 1955). The concept of constructs was developed to describe or account for certain recurring characteristics of a subject's behavior (Thorndike, 1982).

Although constructs cannot be assessed directly, researchers have developed a theory of the construct to a certain level of sophistication. They know how a construct will express itself, what sub-groups in the population possess a high or low degree, what conditions favor or inhibit expression of the construct, what test-tasks elicit the construct, etc. The theoretical considerations form an essential part of construct validity, since they suggest kinds of evidence that are relevant for assessing how well a measurement depends upon the construct. Cronbach (1970) described a general outline for establishing construct validity that was based on these theoretical considerations. First of all, researchers have to suggest what constructs might account for test performance. Secondly, testable hypotheses are derived from the theory surrounding the construct. Finally, researchers carry out studies to test the hypotheses empirically. Theoretical reflections on the behavior of the construct under study underlie all procedures for the investigation of construct validity.

Over the course of time, several procedures have been elaborated to evidence a measure's construct validity. We confine ourself here to the four procedures described by Thorndike (1982) which rely heavily on the original work of Cronbach and Meehl (1955).

5.2.3.1 Comparison of test tasks with conception of the attribute.

The first question to ask about a method of measurement is: do the items and the test task appear to call for the construct in question? Is the content reasonable for eliciting the construct we wish to measure. Congruence between the assumed construct and item content forms a first indicator of the essential nature of our method of measurement, but is in no way conclusive. Unfortunately, no precise methods are available for properly outlining the item or variable domain of a construct (Nunnally, 1967). This matter is left entirely to the researcher's understanding of the construct. This procedure comes close to establishing content validity (de Groot, 1972).

5.2.3.2 Correlational evidence of construct validity.

This procedure comes closest to the meaning of construct validity. It states very generally that measurements should show substantial correlations with different measurements of the same construct, as well as with measurements of theoretically related constructs, whereas low correlations with measurements of other, theoretically unrelated attributes are expected. This type of construct validity is elaborated in two distinct directions: firstly, the assessment of the nomological network and, secondly, the assessment of convergent and divergent validity (Cronbach et al, 1955; Campbell et al, 1959; Thorndike, 1982).

Construct validity of a test is underscored when the relations in the nomological network, defined as the interlocking system of laws which constitute a theory, are supported by empirical evidence. The nomological network of a theory contains a theoretical model, related hypotheses and predictions which include empirical references, and empirical evidence stemming from previous validity studies (de Groot, 1972). Based on the nomological network, researchers develop hypotheses about the nature and strength of relations between the constructs under study and other constructs. They make judgments about the nature of certain activities and the skills required to perform them succesfully. In construct validity, these judgments are tested. When the predicted relations appear empirically, the construct validity of the measurements of the construct is supported. The relations predicted by the nomological network should be able to explain the strength of the

correlations. An additional advantage of this procedure is that the researcher's understanding of the coherence in daily-life is increased. When the relations fail to appear, the nomological network and/or the construct validity of the measurements of the construct are questioned. The uncertainty about the interpretation of negative results forced Nunnally (1967) to discredit the idea that sufficient evidence for construct validity is brought forward when the supposed measurements of a construct behave as expected. He stated that all that can be tested is the correlation between measurements of constructs, whereas researchers came to conclusions about both the theory which surrounds the test and the construct validity of the measurements. Studies of construct validity are only safe, according to Nunnally, when firstly a supposed measurement of a construct is related to a particular observable variable of which the domain is well defined and, secondly, when the assumption of the relationship between the two constructs is unarguable. Moreover, Nunnally warned researchers from assuming that constructs have objective reality. He proposed that a construct's name could act as a useful way of labelling a particular set of observable variables. Validity would then be indicated by the extent to which the name accurately communicates the kind of observables that are being studied.

The relation between a physician's interview behavior as measured by MAAS-GP and MAAS-PMHC and several outcomes of the interview such as patient-satisfaction and the quality of diagnosis and treatment plan (see chapters 6 and 14) are studied here to determine the nomological net of both instruments. Studies are carried out in a simulated consultation hour and in consultation hours in which general practitioners interview real patients.

The assessment of a test's convergent and divergent validity by means of the multitrait-multimethod matrix has been recommended by several authors as an appropriate way of assessing the identifiability of a proposed construct (Campbell et al, 1959; Fiske, 1971; Cronbach, 1972; Kerlinger, 1981; Thorndike, 1982). Convergent validity refers to the assessment of the same construct by means of different methods, whereas divergent validity refers to the assessment of distinct constructs by means of the same and/or other methods. Campbell and

Fiske (1959) approached the assessment of a test's convergent and divergent validity systematically by applying each of several methods of measurement to each of several constructs. They proposed the examination of the resulting matrix of correlations according to four criteria which refer to convergence of the methods with regard to a pertinent construct, to the divergence respectively of constructs and methods and to a general pattern of correlations among the constructs. This classical approach to convergent and divergent validity brings evidence to bear on the quality of the representation of the construct by the content of the test and it brings to light systematic variance introduced by the method of measurement.

We study the convergent and divergent validity of MAAS-GP and MAAS-PMHC by constructing a traditional multitrait- multimethod matrix (see chapters 8 and 12).

In addition to the classical approach, several methodologists have described a less elaborated procedure to determine convergent and divergent validity by stating that the theory of a construct should be able to explain what other variables are correlated or uncorrelated with the measurements of the construct (Kerlinger, 1981; Thorndike, 1982). This procedure fails to provide information about the influence of shared method variance, but enables researchers to describe the content of the constructs more effectively. In this way, it is closely related to the assessment of the nomological net.

We apply this procedure to establish the relations between measurements of medical interviewing skills (MAAS-GP) and other dimensions of medical competency (see chapter 9).

5.2.3.3 Group differences as evidence of construct validity.

If the understanding of a construct leads researchers to expect that distinct groups of subjects will respond differently to their measurements, this hypothesis can be tested. Evidence of construct validity is obtained when the hypothesis that the groups differ on the specific issue is proved by the data (Cronbach et al, 1955; Thorndike, 1982). When researchers apply this kind of construct validity, they have to be aware that they test simultaneously their understanding of and theory about the differences between the groups and the construct

validity of their measurements. Positive results affirm both; negative results may stem from a shortcoming in one or both of them.

5.2.3.4 Treatment effects as evidence of construct validity.

Any experimentally introduced intervention or any naturally occuring change in conditions that might be expected to influence the construct under study, can be used to study construct validity (Cronbach et al, 1955; Thorndike, 1982). Construct validity is supported when scores are in the predicted direction. When two measurements are affected similarly by a variety of treatments, the suggestion is raised that they are measuring much the same trait which is a slightly different way of assessing convergent validity (Nunnally, 1967). Whether the degree of stability is encouraging or discouraging for the proposed interpretation depends upon the theory defining the construct (Cronbach et al, 1955). Furthermore, Thorndike (1982) remarked that measurements of states, as contrasted with measurements of traits, are especially sensitive to interventions. Traits are expected to be relatively insensitive to manipulations of conditions. The impact of an intervention on a pertinent construct provides useful information about the construct.

We have studied the growth of medical students' interviewing skills during medical school. Results supported the construct validity of MAAS-GP measurements of interviewing skills. Results are not presented in this thesis (Kraan et al, 1986).

5.2.3.5 Conclusions about construct validity.

It is evident that Cronbach and Meehl's thoughts on construct validity form a fruitful contribution to the study of validity. Construct validity is essentially based on two important notions. Firstly, researchers formulate a nomological network from which testable hypotheses may be derived on the relation between the construct under study and other constructs. Secondly, researchers confirm or refute the hypotheses based upon empirical evidence stemming from a variety of test situations. Moreover, construct validity forces researchers to be very explicit about the theory which surrounds their constructs.

The approaches to construct validity provided a useful frame of reference for the development of procedures for determining the construct validity of the Maastricht History-taking and Advice Checklist in General Practice and Primary Mental health Care. A more extensive description of how the theoretical notions of construct validity were used to develop research settings, procedures and additional instruments will be provided further in this chapter.

5.2.4 Validity from experiments.

A different type of validity from those addressed in the preceding paragraphs is the fourth type here described because it points to the justification of generalizations drawn from results of experiments as related to situations outside the experiment. With regard to this issue, Campbell and Stanley (1966) and Cook and Campbell (1979) invoked two types of validity called "internal" and "external" validity. "Internal" validity refers to the inferences made by researchers that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause. The concepts of covariation, time sequence and confounding variables are important in internal validity. "External validity" refers to the validity from which researchers infer that the presumed causal relationship can be generalized to and across different types of persons, settings and times. Matters of internal and external validity are of importance to the study of the validity of the MAAS and are discussed in the pertinent chapters.

With regard to the external validity of the MAAS-GP and MAAS-PMHC, influences of different physicians, different simulated-patients, real patients, different cases, different groups of subjects, etc, have to be taken into account in our measurements of interviewing skills. Some of these issues will be elaborated on in this thesis, such as case-influences or the influence of simulated-patients, whereas other influences have to be reserved for future studies. To enhance the external validity of the MAAS, the study of physicians' interviewing skills while they are talking with real patients during their daily practice is very important. This study is currently being carried out, but is not yet analysed.

5.3 Scalability.

A central question in the construction of scales is how researchers go from responses on items by subjects to indices on the dimension of interest. Scales consist of groups of items which are all intended to measure various aspects of one common property or dimension. The number of items that are scored positively is usually taken as a measurement of the underlying dimension. Scaling models formalize the relationship between responses to a group of items and indices which represent the underlying dimension, also called "the latent trait". In the literature, several scaling models are described which are based on either classical test theory or latent trait models.

The place of scalability is not always clear since it is positioned somewhere between reliability and validity. The study of a test's scalability provides information about the concurrent validity of the items and it determines whether all the items are measuring only one dimension. Scalability therefore combines issues of reliability and validity.

5.3.1 Scaling models in classical test theory.

A hallmark of classical test theory is the decomposition of a subject's test-score into a "true-score"-component and an "error-score"-component which is uncorrelated with the "true-score". This decomposition permits assessment of the reliability of a test by means of correlating equivalent forms of a test, by correlating results from two testings and by studying the consistency of performance over items (internal consistency). The process of item selection and the subsequent scale construction is primarily based on these indices of reliability.

Classical test theory has been criticized over the last 15 years for its use of a linear model for the "number of items correct"- score, and for the restricted generalizability of the test conclusion. According to the critics, classical test theory neglects the discrete character of each item and its score by using the sums of "zero" or "one" scores as indices of a subject's position on the dimension of interest. Important test parameters, such as item-difficulties, item-test-correlations and internal consistency, depend heavily on the sample of

subjects tested and the specific items used. Arbitrary elements therefore impinge on the final conclusions about persons and items. Although classical test theory assumes that the scaling problem is solved, critics state that the scaling model is dependent on the particular test and the tested population.

5.3.2 Latent trait models.

An alternative class of scaling models, called "latent trait models" or "item-response models", has recently gained in popularity (Wright, 1977; Thorndike, 1982). The latent trait models specify the relationship between observable test performance and the unobservable traits or dimensions that are assumed to underlie test performance. Latent trait models are characterized by three fundamental notions which are briefly described below (Hambleton et al, 1977).

The first notion refers to the unidimensionality of the latent space, which assumes that all the items in a test are homogeneous in the sense of measuring only one single ability or latent trait. The unidimensionality can be secured either by the formulation of a sound theory or through a factor-analysis of the test items.

The second notion is the assumption of local independence which states that the test item responses of a given subject are statistically independent. This means that a subject's performance on one item does not affect his performance on other items in the test. In effect, no other ability besides the ability under study is common to the items.

The third notion refers to the item-characteristic curve which is a mathematical function that relates the probability of success on an item to the ability of a subject measured by the test. The number of parameters required to describe an item characteristic curve depends on the particular latent trait model. The Rasch model, for example, is the most demanding latent trait model and requires only one parameter to be estimated.

Latent trait models have several advantages, the most important being that it is possible to estimate a subject's ability on the same ability scale from any subset of items in the domain of items that have been fitted in the model. This enables tailored testing and the

construction of tests measuring a similar dimension with different items. A second advantage of latent trait models is that the shape of the item characteristic curve is invariant across subgroups of subjects chosen from the studied population. As a result, scales can be constructed independent of the specific sample of subjects on which the data were obtained. Furthermore, statistical models and computer programs are available that estimate the fit between the latent trait model and subjects' responses to the items (Gustafsson, 1977; Molenaar, 1981). These programs enable researchers to select and eliminate items until the fit between the data and the model is optimal.

In conclusion, we observe that the latent trait models overcome the disadvantages of the scaling model in classical test theory. In the present study, the Rasch model is used because of its attractive though demanding features as methodology for constructing the scales of the MAAS-GP, the MAAS-PMHC, and the scales of the Patient Satisfaction with Communication Checklist (see chapters 6, 7 and 11).

5.4 Reliability.

Some error is involved in any type of measurement whether the subject of measurement is a person's blood pressure or medical interviewing skills. Reliability concerns the extent to which measurements are stable over a variety of conditions in which essentially the same results should be obtained (Nunnally, 1982). The need for reliable measuring instruments is generally recognized and researchers are remarkably uniform about the definition of reliability. A reliable instrument is one with small errors of measurement, one that shows stability, consistency and dependability of scores for individuals on the trait, characteristic, or behavior being assessed (Mitchell, 1979).

There are at least three approaches to the reliability of observational data: observer agreement, the classical psychometric theory of reliability and generalizability theory (Mitchell, 1979).

5.4.1 Observer agreement.

The percentage agreement between observers is a commonly used index of the quality of data collected in observational studies but it is not recommended as denoting the quality of observational data because it has several shortcomings. The shortcomings are insensitivity to degrees of agreement; some degree of agreement can be expected on the basis of chance alone; behavior with very high and low frequencies will have extremely high chance-levels of agreement.

Several alternative coefficients have been developed to overcome these shortcomings. Cohen (1960) proposed as a coefficient of agreement for nominal scales the proportion of agreement corrected for chance, the so-called "kappa". Kappa is almost never reported in scientific communications about medical interviewing skills, despite the fact that its use is recommended (Sanson-Fisher et al, 1981). Moreover, intra-class correlation, which is based on analysis of variance, forms a different type of agreement among observers (Shrout et al, 1979; Guilford et al, 1981). In intra-class correlation, the ratio of the variance of interest over the sum of the variance of interest plus error is determined and interpreted as the correlation between observers, which is taken as an indication of the reliability of the observations. Moreover, intra-class correlation enables determination of the reliability of the mean of several observers' ratings which is of importance because averaging reduces the relative importance of errors of measurement, leaving the relationships enhanced.

Mitchell recommends the assessment of inter-observer agreement as a necessary part of the development and use of observational measures.

5.4.2 Classical psychometric theory of reliability.

The classical test theory views a test score as consisting of two components, the true score and the error-score. The true reflects the presence or extent of some trait or behavior attributed to stable differences among individuals. The error-score is independent of the true-score and includes real error due to random fluctuations and influences of other sources of variation. Determination of the reliability of a test is operationalized by the correlation between two scores on the test, which enables assessment of the size of the

true-score and error-score components (Mitchell, 1979; Thorndike, 1982).

In classical test theory, three procedures are commonly used to determine reliability. Firstly, intra-observer or inter-observer reliability is obtained by the correlation between separate scorings of the same instrument by one or two observers. The true score reflects real differences between subjects, whereas the error-score reflects either inconsistencies in the observer or differences between observers in their use of the instrument along with random error. Secondly, split-half or alternate-forms reliability is obtained by scores on two parts of the same instrument or on two very similar instruments. The true-score reflects consistent individual differences among subjects, whereas the error-score includes random fluctuations and real differences in subject behavior between the observed subdivisions. Thirdly, test-retest reliability is obtained from scores on two separate administrations of the same instrument. The true-score is assumed to reflect a stable trait or behavior. The error-score consists of changes in behavior that occur between the two test administrations in addition to random fluctuations.

It is evident that the constituents of true-score and error-score are dependent on the research setting in which reliability is determined. Mitchell (1979) concluded, therefore, that there is no perfect reliability coefficient, nor is there one that can be generally regarded as the best. She recommended a more inclusive theory to assess reliability; this is described in the following section.

5.4.3 Generalizability theory.

The generalizability theory assumes that test-scores are the result of a number of sources of variation such as subjects, observers, items or conditions. These sources of variation are called "facets" and a particular combination of facets makes up the universe about which test scores may be generalized. By analysis of variance, estimations are made about the contribution of each facet to the overall variation and the size of the different variance components can be established (Cronbach et al, 1972; Mitchell, 1979). The size of the components can be compared and subsequently combined to form a ratio, called

"generalizability coefficient", that represents the proportion of variance attributable to individual differences for a particular universe of conditions.

The ideal data-gathering and analysis design for determination of the variance components seems to be a completely crossed multidimensional analysis of variance (Cronbach et al, 1972). Researchers thus have to identify the different facets that are likely to be sources of variation. After data-collection, an analysis of variance design provides independent estimations of the contribution of each facet to the overall variation in the test scores. The conventional F-statistic establishes whether each facet makes a significant contribution to the scores. More importantly, the size of the variance components can be computed and compared. Finally, different generalizability coefficients can be computed depending on the universe about which researchers wish to generalize. The cost of the extra information provided by each facet is that the number of observations required is multiplied by the number of conditions sampled in the facet.

Generalizability studies are recommended by Mitchell (1979) and Thorndike (1982) because the procedures are conservative and the coefficients are considered as reflecting the lower limits of the true dependability of the observational data, because the generalizability study provides much additional information and because the attention of the researcher is focused on the influence of other factors on behavior.

5.5 The practicability of MAAS-GP and MAAS-PMHC.

The issue of practicability refers to the tension between the costs of using an instrument and the goals that can be achieved. Since no procedures are available for determining the practicability of an instrument, we describe below some of our experiences with regard to the practicability of the MAAS-GP in particular.

The costs of developing instruments such as the MAAS-GP and the MAAS-PMHC are high. Much manpower is required of researchers-physicians who have both research experience and clinical experience during the process of scale construction. Moreover, psychologists and

psychometricians have often to be consulted to provide additional expertise on methodology and psychometrics. Once developed, the costs of applying this type of instrument are low. Observers are capable of scoring the MAAS-GP after two to three training sessions of 3 hours in which they discuss the content of the items and observe several demonstration videotapes under supervision. Faculty, including non-medical professionals, and medical students appear to be able to observe a consultation and to score a physician's interviewing skills on the MAAS-GP. During their first observations, observers are generally overwhelmed by the number of items, but this diminishes after they became acquainted with the content of the items and the position of the items in the checklist. Observers are able to score approximately three interviews of 15 minutes' duration with simulated patients in one hour and 18 to 20 interviews in one day. Almost no evidence is available on the observation of consultations in general practice.

It is also our impression that the use of the MAAS is not limited to the Dutch situation. Our colleagues from the Ben Gurion University of the Negev, Beer Sheva, Israel, have applied the MAAS to videotaped consultations of 4th year medical students (Maoz and Katz, personal communication). We have, moreover, observed and assessed several consultations of North-American physicians without experiencing difficulties.

Several goals can be achieved by observing medical consultations with the MAAS-GP or MAAS-PMHC. Firstly, in medical education, detailed feedback can be given to students on their interview behavior during the consultation. Secondly, the quality of students'/physicians' medical interviewing skills can be assessed. The MAAS-GP has been applied for both reasons to the undergraduate medical curriculum at Maastricht Medical School. Thirdly, research into the relationship between the process of a consultation and the outcomes can be carried out (see chapter 13).

5.6 Instrumental utility - a summary.

In the preceding paragraphs, we have addressed the question of how to assess the instrumental utility of the Maastricht History-taking and Advice Checklist. Instrumental utility has been separated into three elements: validity, scalability and reliability.

Validity can be differentiated into four types:content validity, criterion-orientated validity, construct validity and the validity of experiments. To establish the validity of a test, several research settings and procedures have been developed and described in the literature.

The scalability of measurements can be studied by two scaling models which formalize the relationship between responses of subjects to a group of items and indices that represent the latent trait. The latent trait models have attractive measurement properties when compared to the classical test theory, although they are very demanding.

Determination of the reliability of observational data can be approached in three different ways: namely, by means of observer agreement, by the classical theory of reliability, and by the generalizability theory. Several methodologists recommend the application of generalizability theory because the coefficients do not inflate reliability and do provide much additional information.

The question of the instrumental utility of the Maastricht History-taking and Advice Checklist can be answered by separating the question into the afore-mentioned aspects. Each aspect provides the conceptual framework for the development and application of specific research settings, procedures and statistical analyses. The research settings are elaborated on below.

5.7 From theory to practice.

After reviewing the constituents of instrumental utility, the question is raised of what research has to be conducted to assess the instrumental utility of the Maastricht History-taking and Advice Checklist. The review has made clear that one research setting will not suffice to answer all these questions. It also suggested the research settings that should be developed and carried through to assess the usefulness of our measurements of medical interviewing skills. In the

following paragraphs, we describe the research settings and accompanying extensions that were originally conceived for the purpose of answering questions pertaining to the validity, reliability and scalability of the MAAS (Crijnen et al, 1984; Kraan et al, 1984). Subsequently, an overview is provided of the studies that have been carried out, the chapters in which they are described and the studies which have to be carried out in the near future.

5.7.1 Research setting 1: simulated consultation hours with 40 residents in family medicine.

This research setting, together with some small extensions, forms the core of our study because it enabled us to answer the major research questions. A simulated consultation hour was organized in which 40 residents in Family Medicine interviewed four simulated patients who presented different cases (Crijnen et al, 1986). Each resident was asked to behave as if he had taken charge of a colleague's practice and had to perform a complete medical consultation with each simulated patient. Two cases represented difficult but frequently occurring somatic problems (myocardial infarction, inception of diabetes mellitus), whereas two other cases represented psychological problems (major depression, anxiety states).

The following research questions were studied:

1. Reliability was determined by means of a generalizability study by analyzing observations of a group of 6 observers who observed videotaped interviews of 20 somatic case presentations for MAAS-GP and 20 psychological case presentations for MAAS-PMHC. The studies are described in chapters 7 and 11.
2. Scalability of the scales in MAAS-GP and MAAS-PMHC was studied by means of Rasch analyses on 100 interviews of physicians/ medical students with patients who presented the myocardial infarction case (MAAS-GP) and on 100 interviews of physicians/medical students with patients who presented a dysthymic disorder (MAAS-PMHC). The number of observations was achieved by extending the 40 observations obtained during the simulated- consultation hour to a total of 100 observations. The studies are described in chapters 7 and 11.

3. Convergent and divergent validity of the MAAS-GP and the MAAS-PMHC were studied as part of construct validity by means of two multitrait-multimethod matrices that were constructed by correlating four different methods of measurement of the physician's medical interviewing skills. Data were obtained during the simulated-consultation hour. Studies are described in chapters 8 and 12.
4. The nomological network of medical interviewing skills was studied as part of construct validity by measuring the physician's medical interviewing skills by means of MAAS-GP and MAAS-PMHC, by the assessment of diagnosis and treatment plan and by the assessment of the patient's satisfaction with the quality of the communication. To facilitate determination of the nomological net, the Patient Satisfaction with Communication Checklist was constructed (see chapter 6). Data were obtained during the simulated consultation hours. The nomological net of MAAS-GP has been analyzed although not yet described. The study of the nomological net of MAAS-PMHC has been analyzed and is described in chapter 13.
5. The construction of the Patient Satisfaction with Communication Checklist was accomplished by extending the 160 checklists which were filled in during the simulated consultation hour with 117 checklists filled in by real patients after they had consulted their own physician. This procedure anchored the PSCC in reality and enabled us to apply Rasch analyses during scale construction. Scale construction of this checklist is described in chapter 6.

5.7.2 Research setting 2: Interviewing skills and medical competence.

This research setting enabled us to establish the MAAS-GP's convergent and divergent validity as elaborated by Kerlinger (1981) and Thorndike (1982). In this setting, physicians' medical interviewing skills during a consultation with a simulated patient were measured together with their medical knowledge, interpersonal skills, care and concern for the patient, and medical problem-solving skills. The study is described in chapter 9.

5.7.3 Research setting 3: Growth of interviewing skills during medical school.

In this research setting, treatment effects were considered to provide evidence of construct validity. By assessing the quality of interviewing skills of all 563 undergraduate students at Maastricht Medical School at one point in time, we had the opportunity to reveal the patterns of growth of interviewing skills during medical school. In addition to the evidence on construct validity, this study provided information on the susceptibility of interviewing skills to influences from the medical curriculum. The study is not described in this thesis (Kraan et al, 1986).

5.7.4 Research setting 4: Medical interviewing skills during consultation hours of general practitioners.

The purpose of this study is to enhance external validity and to provide information about the nomological net of the MAAS-GP and MAAS-PMHC. Medical consultations of 30 General Practitioners with 600 patients will be recorded and observed with MAAS-GP and MAAS-PMHC. Moreover, physicians' perception of the quality of the communication and the established diagnosis will be recorded. Patients will be asked to fill in the Patient Satisfaction with Communication Checklist and the General Health Questionnaire which determines whether patients' experience psychological problems. This study has not yet been carried out.

It is clear that the study of the instrumental utility of the MAAS-GP and MAAS-PMHC can be separated into many smaller studies. The interested reader will surely be able to conceive of additional studies to test the instrumental utility of our measurements of medical interviewing skills. We have confined ourselves here to studies that were originally developed to assess reliability, scalability and validity of both instruments. A summary of studies on the instrumental utility of MAAS-GP and MAAS-PMHC is given in "time-tables" 5.1 and 5.2.

5.8 The validity of simulated patients.

Since the studies presented in this thesis rely heavily on the use of simulated patients, we reviewed the - non-too-abundant - literature about the validity of simulated patients and the communication between physicians and simulated patients. Simulated patients have several attractive characteristics which make them a valuable tool in medical education and research into the medical interview. The characteristics are that they appear to be available at times appropriate to the programme; that they can be trained to present a wide array of problems; that they can be interviewed repeatedly which yields the advantage of standardizing the clinical situation; that they are able to provide critical feedback.

Several criteria were formulated for the use of simulated patients in educational settings that also hold true for research situations (Norman et al, 1982). - credibility or face-validity: several studies reveal that simulated patients are almost indistinguishable from real patients. - comprehensiveness: in comparison with written types of simulation or computer problems, simulated patients can be used to simulate virtually all aspects of the physician-patient encounter. - precision: well-trained simulated patients appear to provide a consistent clinical picture from one encounter to the next. - validity: the differences in interviewing styles of physicians who talked with real and simulated patients have been studied several times. No differences were observed in the number of questions on history-taking and physical examination between interviews with real and simulated patients. (Norman et al, 1982). No differences were observed in the level of empathy expressed by medical students who interacted with real and simulated patients presenting psychological problems (Sanson-Fisher et al, 1980). Our own study revealed that simulated patients presented their roles very naturally according to the judgments made by general practitioners and residents in primary care (Crijnen et al, 1986).

We conclude, therefore, that simulated patients form an accurate and valid representation of real patients because their performance resembles the behavior of real patients. Futhermore, physicians do not seem to have different interview styles when talking to real or simulated patients.

"Time table" 5.1: Establishing the instrumental utility of the Maastricht History-taking and Advice Checklist in General Practice (MAAS-GP).

Issue	Method	Research setting/ population	Intended Studied Analyzed Described	Chapter or described in
inter-observer reliability	generalizability study	20 residents in general practice during simulated consultation hours	S, A, D	7
scalability	Rasch analyses	100 medical students, residents, general practitioners during simulated consultation hours	S, A, D	7
content validity	literature study discussion with colleagues	-	S, D	2,3,4
predictive validity	correlation with future criterion	medical students/ future physicians	I	-
construct-validity	1. convergent and divergent validity by MTMM-matrix	40 residents in General Practice during simulated consultation hours	S, A, D	8
	2. correlation with measures of medical competence	28 physicians talking with one simulated patient	S, A, D	9
	3. nomological net of medical interviewing skills	40 residents in General Practice during simulated consultation hours	S, A	-
	4. nomological net of medical interviewing skills	30 general practitioners talking with 600 patients during real consultation hours	I	-
	5. treatment effects	563 medical students talking with simulated patients	S, A, D	Kraan et al, 1986

"Time table" 5.2: Establishing the instrumental utility of the Maastricht History-taking and Advice Checklist in Primary Mental Health Care (MAAS-PMHC).

Issue	Method	Research setting/ population	Intended to study Studied Analyzed Described	Chapter
inter-observer reliability	generalizability study	20 residents in general practice during simulated consultation hours	S, A, D	10,11
scalability	Rasch analyses	60 medical students and 40 residents in General Practice during simulated consultation hours	S, A, D	10
content validity	literature, discussion with colleagues, content analyses	40 residents in general practice during simulated consultation hours	S, A, D	2,3,4, 11
construct-validity	1. convergent and divergent validity by MTMM-matrix	40 residents in general practice during simulated consultation hours	S, A, D	12
	2. nomological net of medical interviewing skills	40 residents in general practice during simulated consultation hours	S, A, D	13
	3. nomological net of medical interviewing skills	30 general practitioners talking with 600 patients during real skills consultation hours/detection of minor psychiatric disorder by means of General Health Questionaire	I	-

REFERENCES

Berkel HJM van. De diagnose van toetsvragen (The diagnosis of test items - dissertation). Universiteit van Amsterdam, Amsterdam, 1984.

Campbell DT, Fiske DW. Convergent and discriminant validation by the multi-trait multi-method matrix. Psychological Bulletin, 1959; 56: 81-105.

Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Rand McNally, Chicago, 1966.

Cohen J. A coeffient of agreement for nominal scales. Educational and Psychological Measurement, 1960; 20: 37-46.

Cook TD, Campbell DT. Quasi-experimentation. Rand McNally, Chicago, 1979.

Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin, 1955; 52: 281-302.

Cronbach LJ. Essentials of psychological testing. Harper and Row, New York, 1970.

Cronbach LJ, Gieser GC, Nanda H, Rajaratham N. The dependability of behavioral measurements: theory of generalizability for scores and profiles. John Wiley and Sons, New York, 1972.

Crijnen AAM, Kraan HF. Reliability and validity of the Maastricht History-taking and Advice Checklist in General Practice (research proposal). Department of Social Psychiatry, Rijksuniversiteit Limburg, Maastricht, 1984.

Crijnen AAM, Thiel J van, Kraan HF. Evaluatie van consultvoering: een spreekuur nagebootst (Evaluation of a medical consultation: simulating consultation hours). Huisarts en Wetenschap, 1986; 29: 316-318.

Fiske DW. Measuring the concepts of personality. Aldine Publishing Company, Chicago, 1971.

Groot AD de. Methodologie (Methodology). Mouton, 's-Gravenhage, 1972.

Guilford JP, Fruchter B. Fundamental statistics in psychology and education. McGraw-Hill, London, 1981.

Gustafsson JE. The Rasch-model for dichotomous items: theory, applications and a computer program. Reports from the Institute of Education, No. 64, University of Göteborg, Sweden, 1977.

Hambleton RL, Cook LL. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977; 14: 75-96.

Kerlinger FN. Foundations of behavioral research. Holt, Rinehart and Winston Inc., New York, 1981.

Kraan HF, Crijnen AAM, DeVries M, Zuidweg J, Imbos T. Are medical interviewing skills teachable? Perspectief, 1986; 4: 29-51.

Kraan HF, Crijnen AAM. Reliability and validity of the Maastricht History-taking and Advice Checklist in Primary Mental Health Care (research proposal). Department of Social Psychiatry, Rijksuniversiteit Limburg, Maastricht, 1984.

Mitchell SK. Interobserver agreement, reliability and generalizability of data collected in observational studies. Psychological Bulletin, 1979; 86: 376-390.

Molenaar IW. Programma beschrijving van PML voor het Rasch model (Description of the PML-program for the Rasch model - version 3.1). Heymans Bulletin, Vakgroep Statistiek en Meettheorie, Universiteit van Groningen, Groningen, 1981.

Nunnally JC. Psychometric theory. McGraw-Hill, London, 1967.

Nunnally JC. Reliability of measurement. In: The encyclopedia of educational research. McMillan and Free Press, New York, 1982.

Norman GR, Tugwell P, Feightner JW. A comparison of resident performance on real and simulated patients. Journal of Medical Education, 1982; 57: 708-715.

Philipsen H. Onderzoek als datamatrix (Research as datamatrix-samenvatting college). Rijksuniversiteit Limburg, Maastricht, 1984.

Sanson-Fisher R, Fairbairn S, Maguire P. Teaching skills in communication to medical students - a critical review of the methodology. Medical Education 1981; 15: 33-37.

Sanson-Fisher RW, Poole AD. Simulated patients and the assessment of medical students' interpersonal skills. Medical Education, 1980; 14: 249-253.

Shrout PE, Fleiss JL. Intraclass correlation: uses in assessing rater reliability. Psychological Bulletin, 1979; 86: 420-428.

Thorndike RL. Applied psychometrics. Houghton Mifflin Cie., Boston, 1982.

Wright BD. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977; 14: 97-116.

CHAPTER 6 MEASURING PATIENT SATISFACTION WITH THE COMMUNICATION

A.A.M. Crijnen and H.F. Kraan

6.1 Introduction.

In this chapter, we describe the construction of the Patient Satisfaction with Communication Checklist which measures patients' evaluation of distinct dimensions in physician-patient communication. This checklist was originally developed to study the construct validity of the Maastricht History-taking and Advice Checklist by simultaneously measuring process and outcomes of an initial medical interview. The checklist also appears to enhance our understanding of the constituents of a good medical interview as judged by the patient. Three elements are considered to be of prime importance: the patient's desire to express concerns and emotions, the patient's expectation of receiving information about complaints, their aetiology and prognosis, and the patient's intention to comply with the advice and medical regimen. To measure adequately the patient's evaluation of physician-patient communication, demanding psychometric statistics are applied in the process of scale construction.

6.2 Measuring patient satisfaction.

In patient satisfaction, which reflects the patient's perception of the quality of the delivered care, three dimensions are generally discerned: namely, an affective dimension, a cognitive dimension and the intention to comply, although the quantity and diversity of scientific communications on this issue lead one to believe that much differentiation is observed (Korsch et al, 1972; Wolf et al, 1978; Pendleton, 1983). Research regarding patient satisfaction is, unfortunately, often characterised by a lack of sound methodology because the dimensions within the patient's satisfaction with health care are ill defined and because many studies pay insufficient attention to the psychometric quality of measurement (Ware et al, 1978; Locker et al, 1978; Lebow, 1982; Lebow, 1983).

Affective satisfaction refers to the patient's feelings of trust and confidence in the physician and his perception of the physician's positive regard and willingness to listen to the patient's concerns. Affective satisfaction is likely to be increased when the physician deals with the patient's concerns and expectations and when the physician enables the patient to express thoughts and emotions.
Cognitive satisfaction points to explanations and information given by the physician and to the patient's understanding of diagnosis, etiology, prognosis and (side-)effects of treatment. Cognitive satisfaction is increased when the physician volunteers a lot of information and explains the patient's condition comprehensibly.
The intention to comply refers to the patient's determination to follow-up the proposed treatment plan. Compliance is affected by the quality of information-exchange from physician to patient and by the patient's participation in the negotiation on the request for help. The patient's opinion of the quality of the communication is usually evaluated in terms of these three dimensions.

Furthermore, research on patient satisfaction is characterized by two shortcomings: measurements of satisfaction are often ill-defined and insufficient attention is given to the psychometric properties of the scales. The first shortcoming is due to the fact that the term "patient satisfaction"' points to the patient's opinion of the provided care in general and covers a wide array of dimensions, ranging from a global impression of satisfaction immediately after a consultation to opinions on more specific topics such as the physician's technical qualities or interpersonal manner, etc.. Moreover, researchers appear to be very creative in the development of instruments, each with a specific content (Lebow, 1983). As a result, confusion arises over the constituents of satisfaction and it is difficult to compare different studies and to determine what patients actually value in physician-patient communication.

The second shortcoming pertains to the quality of measurement because the process of scale construction is often insufficiently guided by sound psychometric procedures and statistics (Ware et al, 1978; Locker et al, 1978; Thorndike, 1982). Evidence of reliability, such as coefficients of internal consistency used to support item

selection, is usually not provided. Many studies fail to achieve variability in scores due to social desirability or halo effects (Ware et al, 1978; Lebow, 1982). Ultimately, differences between patients with regard to the studied dimensions are inadequately measured and it remains unclear to which degree the scales are measuring the concepts validly.

To overcome both shortcomings, Ware et al (1983) suggested the construction of multi-item scales based on factored homogeneous item dimensions, groups of items which have a similar content theoretically and which share a substantial amount of variance empirically. They constructed the Patient Satisfaction Questionaire, which measures six different characteristics of providers of medical care services that are considered in general as influencing patient satisfaction. These dimensions are interpersonal manner, technical quality, accessibility/convenience, finances, efficacy/outcomes, continuity, physical environment, availability. The dimensions, technical quality and interpersonal skills, appeared to be most influential in determining patient satisfaction but these dimensions were more difficult to measure when compared with the other dimensions. To enhance the quality of measurement, both scales had more items added. In addition to the scale construction procedure, the multitrait-multimethod validation procedure was applied to these measurements (see also chapter 5). Difficulties were experienced with regard to divergent validity which suggests that characteristics of the method of measurement interfered with the assessment of the dimension under study. The conclusion can therefore be drawn that even the application of solid scale construction procedures does not guarantee the quality of measurement, especially with regard to the evaluation of the patient's perception of the physician's competence.

In the study presented here, the construction of the Patient Satisfaction with Communication Checklist is described. To overcome the shortcomings mentioned above, the dimensions contributing to patient satisfaction were determined both theoretically and empirically and, secondly, the quality of measurement of the scales was enhanced by applying demanding psychometric statistics, such as the Rasch-model, in the process of scale construction.

6.3 Method.

The method section comprises the construction of the original Patient Satisfaction with Communication Checklist, a justification for the choice of the scaling model, a description of the subjects and, finally, the analyses.

6.3.1 Construction of the PSCC.

Formulation of the 54 statements which constituted the original PSCC was based on empirical evidence reported in the literature and theoretical reflections on the impact of physician-patient communication on the patient. Nine dimensions were selected to provide a framework for item formulation. Brief definitions of the dimensions are presented below. 1. Facilitation: The patient's opinion of the physician's interview behavior that encouraged him to voice his concern and to ask any questions. 2. Understanding: The patient's perception of whether the physician understood his concerns and complaint(s). 3. Directivity: The patient's opinion on the degree to which the physician controlled the content and course of the consultation. 4. Disrupted communication: The physician's behavior generally considered as hampering the communication. 5. Providing information: The quality and amount of information provided by the physician. 6. Insight: The patient's judgment of whether his comprehension of his complaint(s) and problems has been increased by the end of the consultation. 7. Intention to comply: The patient's intention to comply with the proposed advice and treatment. 8. Empathy: The patient's perception of the physician's understanding of his inner world. 9. Satisfaction: The patient's opinion on the degree to which the consultation was satisfactory.

Within each dimension, several statements have been formulated pertaining to either the quality of the physician's communicative behavior with regard to this dimension or to the impact of the physician's behavior on the patient. All statements referred to a specific, well-described topic. Moreover, they were worded in brief sentences in easy to understand Dutch. During the formulation of items, a methodological suggestion made by Locker and Dunt (1978) was adopted to measure patient satisfaction with items referring to specific topics.

Patients were requested to recall what they had experienced during the foregoing consultation and to indicate on Likert-type, 5-point scales whether they agreed or disagreed with the statements.

6.3.2 Choice of the scaling model.

The central issue during scale construction is the assembly of a set of items which measure the dimension of interest to a satisfactory degree and which collectively reflect different levels of achievement on this dimension. The process of scale construction is supported scientifically by scaling models that formalize the relationship between subjects' responses on items and indices which represent the level of achievement. Two well-known scaling models are based on either the classical test theory or latent trait models.

Classical test theory has been criticized because of its use of a linear model for the 'number of items correct"-score and because of the restricted generalizability of the test conclusion. Critics state that the discrete character of each item and its score is neglected by using the sums of "zero" and "one" scores to position a subject on the dimension of interest. Moreover, test parameters, such as internal consistency, item difficulties and item-test correlations, depend heavily on the sample of persons and the group of items tested. Arbitrary elements impinge strongly on the final conclusions about persons.

Latent trait models have recently gained in popularity (Hambleton et al, 1977; Wright, 1977; Meerling,1981). These models specify the relation between observable test performance and the unobservable traits by means of a mathematical function. The Rasch-model is the most demanding but also the most attractive latent trait model. In the Rasch-model, all items are assumed to have equal discriminating power and to vary only in terms of their difficulty. The advantages of the Rasch-model are that item-parameters are invariant across samples of subjects chosen from the population of interest and that a subject's ability can be estimated from any subset of items that appeared to fit in the model. The Rasch-model therefore allows the construction of scales independent of the specific sample of subjects on which the data for scale construction were obtained and independent of the specific items that are used.

In the present study, the Rasch-model was applied to secure optimal measurement properties of the scales in the Patient Satisfaction with Communication Checklist.

6.3.3 Subjects.

Since the PSCC is intended to be employed as an outcome measurement of physician-patient communication in general practice care and as an outcome measurement of the physician/student-simulated patient communication in medical education, the sample of subjects was drawn from both populations (see table 6.1).

Table 6.1: Participating subjects.

Setting	Physicians	Patients	Number of Checklists with missing values	without missing values
1. Simulated consultation hours	40 residents in general practice	9 simulated patients	160	151
2. Consultation hours	7 general practitioners	117 real patients	117	95
		Total	277	246

The PSCC was filled out by 117 consecutive patients during consultation hours of seven general practitioners. After leaving the consultation-room, the physician's secretary or a research-assistant asked the patients to fill in the checklist. Approximately 30 patients did not take part in the study, mostly because they were in a hurry or because they had left the physician's office before they were asked to participate. After reading the checklist, no patient refused to answer it.

In addition, 160 checklists were filled in by 9 different simulated patients who were interviewed by 40 residents in primary care as part of the validity study of the Maastricht History-taking and Advice

Checklist, a measurement of physicians' medical interviewing skills (Crijnen et al, 1985; see this thesis). During simulated consultation-hours, residents were asked to behave as if they were taking charge of a colleague's practice and to interview 4 simulated patients for 15 minutes. Simulated patients presented complaints accompanying myocardial infarction, inception of diabetes mellitus, panic disorder or major depression. At the end of each interview, they were asked to give their personal opinion on the quality of the communication on each of the 54 PSCC items. The participating simulated patients were recruited from a larger pool of lay-people trained by the Skills laboratory at Maastricht Medical School to simulate complaints in the undergraduate medical curriculum. Simulated patients are able to offer feedback on a physician's performance and to provide suggestions for the improvement of interviewing skills (Stillman et al, 1983).

Due to missing values on one or more of the 54 items of the original checklist, only 95 of the 117 checklists filled in by real patients and 151 of the 160 checklists filled in by simulated patients were available for computations.

6.3.4 Analyses.

A sequence of analyses was carried out to explore the dimensions in the checklist and to construct the scales. The PSCC was first factor-analyzed by means of a Varimax rotation for 9 factors since 9 theoretical dimensions were considered to underlie the construction of the original checklist (SPSS - Nie et al, 1975). Items with a factorloading greater than .50 for the first five and .35 for the remaining factors were selected to constitute the dimensions. Responses on these items formed input for Rasch analyses. Since the theoretical considerations and empirical evidence with regard to the dimensions in patient satisfaction appeared to be only partly in accordance, and because the number of dimensions was considered to be too large, additional analyses were carried out. A second Varimax-rotated factor analysis for 4 factors, followed by Rasch analyses, was applied. Results were still unsatisfactory because the disagreement between theory and empirical evidence continued. In the final step, items stemming from both 9 and 4 factor factor analyses which pertained to identical dimensions formed input for Rasch analyses. The sequence of these analyses is depicted in table 6.2.

Table 6.2: Sequence of analysis during the construction of the PSCC.

hypothesized dimensions	9 - factor interrelation	fit to Rasch-model	4 - factor interpretation	fit to Rasch-model	fit to Rasch model after combining 9 and 4 factor interpretation	final scale
1. facilitation	+ (with satis-faction)	± (with satis-faction)	+	±	± (satisfaction excluded)	facilitation
2. providing information	+ (with insight)	+	+ (with insight)	+	+ (with insight)	(see insight)
3. disruption of communi-cation	+	±	+	±	±	disruption of communication
4. directivity	+	+	-	-		directivity
5. insight	+ (with providing information)	+	+ (with providing information)	+	+ (with providing information)	insight
6. intention to comply	+	+	+	±	±	intention to comply

7. physician's understanding	-	-	-	-		-
8. empathy	+	-	-	-		-
9. satisfaction	+ (with facilitation)	±	+ (with satisfaction)	±	- (excluded)	-
10. -	doctors' expertise	-	-	-		-
11. -	(not interpretable)	-	-	-		-

+ = can be interpreted or fits immediately in Rasch-model
± = fits in Rasch-model after deleting several items
- = can not be interpreted or does not fit to Rasch-model

Rasch analyses were carried out by means of the PML-program (Gustaffson, 1977; Molenaar, 1981). By applying the Binomial test and Allerup's Graphical test, items were selected to fit in the Rasch-model. Subsequently, the scalability of item groups was analyzed by means of the Martin Löf chi-square test which determines the fit between the observed proportion positive answers for a pertinent item on each ability level and the estimated probability of a positive answer according to the assumptions of the Rasch-model. Items which diminished the fit were eliminated. In addition, the unidimensionality of the ultimate scales was tested by a second Martin Löf test which determines differences of estimated person parameters between pairs of scales.

Finally, item parameters and confidence intervals were computed separately for the sample of simulated patients and for the sample of real patients to determine the occurrence of item-bias. Results are presented in tables 6.3, 6.4, 6.5 and 6.6.

Table 6.3: Rasch homogeneous scales in PSCC.

Scale	Number of items	Martin Löf test	DF	Probability*
1. Facilitation	3	0.666	2	0.73
2. Insight	5	8.166	12	0.77
3. Intention to comply	4	8.143	6	0.23
4. Disruption of communication	4	4.443	6	0.62
5. Directivity	3	5.462	2	0.07

*) A high level of probability indicates a good fit of the item group to the assumptions of the Rasch-model.

Table 6.4: Measurement properties of items in PSCC.

item	score group		N persons	N positive answers	observed proportion positive answers	probability of positive answer according to the Rasch-model	item difficulty
	rawscore	ability					
facilitation							
1.	0	- ∞	9	0	0.00	0.00	-1.68
	1	-.97	10	8	0.80	0.82	
	2	.99	76	74	0.97	0.97	
	3	+ ∞	151	151	1.00	1.00	
2.	0	- ∞	9	0	0.00	0.00	-1.07
	1	-.97	10	2	0.20	0.14	
	2	.99	76	62	0.82	0.82	
	3	+ ∞	151	151	1.00	1.00	
3.	0	- ∞	9	0	0.00	0.00	0.59
	1	-.97	10	0	0.00	0.03	
	2	.99	76	16	0.21	0.21	
	3	+ ∞	151	151	1.00	1.00	
insight							
4.	0	- ∞	103	0	0.00	0.00	-0.63
	1	-1.42	50	15	0.30	0.35	
	2	-0.42	31	19	0.61	0.59	
	3	0.42	24	21	0.87	0.77	
	4	1.42	23	20	0.87	0.90	
	5	+ ∞	15	15	1.00	1.00	

Table 6.4: (continued)

item	score group		N persons	N positive answers	observed proportion positive answers	probability of positive answer according to the Rasch-model	item difficulty
	rawscore	ability					
5.	0	- ∞	103	0	0.00	0.00	-0.01
	1	-1.42	50	8	0.16	0.19	
	2	-0.42	31	14	0.45	0.40	
	3	0.42	24	16	0.67	0.61	
	4	1.42	23	17	0.74	0.81	
	5	+ ∞	15	15	1.00	1.00	
6.	0	- ∞	103	0	0.00	0.00	0.06
	1	-1.42	50	11	0.22	0.18	
	2	-0.42	31	12	0.39	0.38	
	3	0.42	24	13	0.54	0.59	
	4	1.42	23	17	0.74	0.80	
	5	+ ∞	15	15	1.00	1.00	
7.	0	- ∞	103	0	0.00	0.00	0.13
	1	-1.42	50	10	0.20	0.17	
	2	-0.42	31	9	0.29	0.36	
	3	0.42	24	11	0.46	0.57	
	4	1.42	23	21	0.91	0.79	
	5	+ ∞	15	15	1.00	1.00	
8.	0	- ∞	103	0	0.00	0.00	0.44
	1	-1.42	50	6	0.12	0.12	
	2	-0.42	31	8	0.26	0.28	
	3	0.42	24	11	0.46	0.47	
	4	1.42	23	17	0.74	0.71	
	5	+ ∞	15	15	1.00	1.00	

Table 6.4: (continued) *

item	score group		N persons	N positive answers	observed proportion positive answers	probability of positive answer according to the Rasch-model	item difficulty
	rawscore	ability					
intention to comply							
9.	0	- ∞	1	0	0.00	0.00	-2.33
	1	-1.85	14	11	0.79	0.75	
	2	-0.02	51	49	0.96	0.96	
	3	1.85	91	90	0.99	0.99	
	4	+ ∞	89	89	1.00	1.00	
10.	0	- ∞	1	0	0.00	0.00	-1.07
	1	-1.85	14	3	0.21	0.21	
	2	-0.02	51	42	0.82	0.87	
	3	1.85	91	91	1.00	0.98	
	4	+ ∞	89	89	1.00	1.00	
11.	0	- ∞	1	0	0.00	0.00	1.00
	1	-1.85	14	0	0.00	0.03	
	2	-0.02	51	11	0.22	0.14	
	3	1.85	91	70	0.77	0.81	
	4	+ ∞	89	89	1.00	1.00	
12.	0	- ∞	1	0	0.00	0.00	2.40
	1	-1.85	14	0	0.00	0.01	
	2	-0.02	51	0	0.00	0.03	
	3	1.85	91	22	0.24	0.22	
	4	+ ∞	89	89	1.00	0.00	

Table 6.4: (continued)

item	score group		N persons	N positive answers	observed proportion positive answers	probability of positive answer according to the Rasch-model	item difficulty
	rawscore	ability					
disruption of communication							
13.	0	- ∞	11	0	0.00	0.00	-0.30
	1	-1.15	20	4	0.20	0.31	
	2	-0.01	21	13	0.52	0.59	
	3	1.14	40	35	0.88	0.83	
	4	+ ∞	154	154	1.00	1.00	
14.	0	- ∞	11	0	0.00	0.00	-0.25
	1	-1.15	20	6	0.30	0.30	
	2	-0.01	21	14	0.67	0.58	
	3	1.14	40	31	0.78	0.82	
	4	+ ∞	154	154	1.00	1.00	
15.	0	- ∞	11	0	0.00	0.00	-0.20
	1	-1.15	20	8	0.40	0.28	
	2	-0.01	21	9	0.43	0.56	
	3	1.14	40	33	0.83	0.82	
	4	+ ∞	154	154	1.00	1.00	
16.	0	- ∞	11	9	0.00	0.00	0.74
	1	-1.15	20	2	0.10	0.11	
	2	-0.01	21	6	0.29	0.27	
	3	1.14	40	21	0.53	0.52	
	4	+ ∞	154	154	1.00	1.00	

Table 6.4: (continued)

item	score group		N persons	N positive answers	observed proportion positive answers	probability of positive answer according to the Rasch-model	item difficulty
	rawscore	ability					
directivity							
17.	0	- ∞	58	0	0.00	0.00	-0.16
	1	-.70	34	16	0.47	0.39	
	2	.70	50	33	0.66	0.71	
	3	+ ∞	104	104	1.00	1.00	
18.	0	- ∞	58	0	0.00	0.00	0.03
	1	-.70	34	13	0.38	0.32	
	2	.70	50	31	0.62	0.66	
	3	+ ∞	104	104	1.00	1.00	
19.	0	- ∞	58	0	0.00	0.00	0.13
	1	-.70	34	5	0.15	0.29	
	2	.70	50	36	0.72	0.62	
	3	+ ∞	104	104	1.00	1.00	

Table 6.5: Results of the Martin Löf test for dimensionality of the scales in the Patient Satisfaction with Communication Checklist.

Scales	Chi-square	DF	Probability*	Pearson product-moment correlation
DIR with ITC	105	11	.00	.12
DCO with DIR	63	11	.00	.46
DCO with ITC	107	15	.00	.12
DCO with FAC	105	11	.00	.07
FAC with ITC	49	11	.00	.21
DIR with FAC	50	8	.00	.31
INS with DIR	171	14	.00	.12
INS with DCO	180	19	.00	.00
INS with ITC	31	19	.03	.37
INS with FAC	50	14	.00	.27

*) A low level of probability on the Martin LöF test for dimensionality indicates that scales are measuring distinct dimensions.

To facilitate statistical analyses, data were prepared in several ways. Eighteen negatively-formulated items were recoded and the 5-point scale was split into one and zero scores because Rasch analysis requires dichotomized data. One is a positive answer e.g. "strongly agree" and "agree", whereas zero is a negative answer e.g. "no opinion", "disagree" and "strongly disagree".

6.4 Results.

The 9 factor factor analysis produced 7 factors related to 8 hypothesized dimensions (see table 6.2). One new dimension emerged called "physician's expertise". One factor could not be interpreted. Six factors fitted in the Rasch-model, although the content of the factors was not always clearly related to the theoretical dimensions. Facilitation and satisfaction clustered in one factor. Two factors contained items of both the dimensions providing information and insight. Disrupted communication, directivity and intention to comply pertained clearly to their hypothesized dimensions.

In the 4 factor factor analysis, satisfaction and facilitation formed the first factor, providing information and insight the second, disrupted communication the third and the intention to comply the fourth. All item-groups fitted well in the Rasch-model.

Tabel 6.6: Estimates of item parameters and confidence intervals for real and simulated patients of the Patient Satisfaction with Communication Checklist.

item	simulated patients		real patients	
	parameter	confidence-interval	parameter	confidence-interval
1.	-1.84	-2.68 - -1.00	-0.83	-2.58 - 0.92
2.	0.13	-0.37 - 0.63	-0.13	-1.58 - 1.31
3.	1.71	1.21 - 2.21	0.96	-0.35 - 2.28
4.	-0.56	-1.01 - 0.09	-0.83	-1.41 - -0.25
5.	0.36	-0.25 - 0.97	-0.04	-0.27 - 0.50
6.	-0.02	-0.60 - 0.55	0.02	-0.52 - 0.67
7.	0.28	-0.33 - 0.88	0.58	0.03 - 1.13
8.	-0.16	-0.72 - 0.41	0.27	-0.27 - 0.82
9.	-3.73	-4.91 - -2.55	-0.89	-1.81 - 0.04
10.	-0.66	-1.21 - -0.12	-2.56	-4.08 - -1.04
11.	3.53	2.95 - 4.11	1.81	1.19 - 2.43
12.	0.86	0.40 - 1.32	1.63	1.01 - 2.26
13.	0.62	-0.83 - 2.08	-0.22	-1.35 - 0.92
14.	-0.08	-1.57 - 1.41	0.01	-1.11 - 1.12
15.	-0.54	-2.14 - 1.05	0.21	-0.89 - 1.31
16.	-	- - -	-	- - -
17.	0.41	-0.49 - 1.30	-0.34	-1.36 - 0.67
18.	-1.22	-2.29 - -0.14	0.17	-0.77 - 1.11
19.	0.81	-0.10 - 1.72	0.17	-0.77 - 1.11

*) Item 16 was not computable due to a lack of variance in simulated patients.

Finally, after Rasch-analyzing items stemming from a combination of the 9 and 4 factor solutions, the dimension "satisfaction" disappeared from the dimension "facilitation". "Insight" and "providing information" formed a second dimension, whereas "disrupted communication" and the "intention to comply" formed two further dimensions. Since "directivity" emerged in the first analyses as a Rasch homogeneous scale, and because it formed a theoretically well-defined dimension, it became part of the final PSCC together with "facilitation", "insight", "intention to comply" and "disrupted communication".

Statistics of the Martin Löf-test presented in table 6.3, indicate a good fit of the item-groups to the assumptions of the Rasch-model. The hypothesis that the item-groups form Rasch-homogeneous scales cannot be rejected according to the goodness of fit-test used, as is evidenced by the high probabilities (Löf, 1973). Directivity fits least well.

The measurement properties of each item are presented in more detail in table 6.4. Estimates of a subject's ability on the dimensions of interest are indicated by raw scores and ability scores. The item parameter estimates, which reflect the strength of an item in measuring the dimension of interest, are indicated by the item difficulties. Low item difficulties require only low ability scores to answer an item positively. Moreover, the observed proportion of positive answers is compared to the estimated probability of a positive answer according to the Rasch-model. Strong similarities between observed and estimated proportions positive answers can be seen. Sometimes violations of the Rasch-model occur when the range of item difficulties is relatively narrow, e.g. disrupted communication.

The unidimensionality of the scales was tested by Martin-Löf chi-square tests for pairs of scales. The low probability levels presented in table 6.5 indicate that each scale measures only one dimension which forms additional evidence of divergent validity.

Finally, item-bias was examined by separately comparing the confidence intervals of the item parameters estimated over both samples of real and simulated patients. The results, displayed in table 6.6, reveal that only items 9 and 12 and none of the other items are influenced by item-bias. The items in the final scales are presented in table 6.7.

6.5 Discussion.

For patient satisfaction with the medical consultation, the quality of physician-patient communication appears to be of paramount importance. Notwithstanding a considerable amount of research, the dimensions in satisfaction distinguished by patients are not clearly defined. Moreover, shortcomings in the quality of measurement are often observed. In the present study, the construction of the Patient Satisfaction with Communication Checklist is described. The study

Table 6.7: Patient Satisfaction with Communication Checklist.

Facilitation.

1. The doctor gave me the opportunity to tell my own story.
2. The doctor enabled me to talk frankly.
3. I could ask anything I wanted to.

Insight.

4. I know the pros and cons of my treatment.
5. The doctor clarified the meaning of my conduct.
6. The doctor explained side-effects of my medication.
7. The doctor gave me new insight into my problem, enabling me to cope with it.
8. The impact of my problem on people close to me became clear.

Intention to comply.

9. I'm able to recall the agreements which I reached with the doctor.
10. I will certainly cooperate with the proposed treatment.
11. The doctor checked whether I had understood the advice and further information.
12. The doctor gave me much responsibility in the choice of my treatment.

Disruption of communication.

13. The doctor used incomprehensible jargon.
14. The doctor switched too fast from one subject to another.
15. I was made anxious and uneasy by the questions the doctor asked.
16. Sometimes the doctor reexpressed my statements differently from my original meaning.

Directivity.

17. Sometimes the doctor invited me to talk about subjects which I would rather not have discussed at that moment.
18. Sometimes the doctor would have liked to take a decision in my place.
19. On some aspects of my life, the doctor wanted me to adopt a view that was different from my own.

attempts to obviate the shortcomings by careful examination of the dimensions discerned in patient satisfaction and by the application of adequate psychometric statistics in the process of scale construction. Ultimately, five dimensions are distinguished in patient satisfaction with communication; these are measured by Rasch homogeneous scales.

The scale "facilitation" measures the patient's opinion of the opportunity provided for talking about his complaint(s), for voicing concerns and for asking any questions. Facilitation is important since many patients strongly desire to mention their greatest concerns to the physician, but either do not often have the opportunity to mention them or they are not encouraged to do so (Korsch et al, 1972). The influence of facilitation on the patient is considerable since the physician's willingness to listen to the patient is related to feelings of trust and acceptance, and it enhances compliance with the prescribed regimen. Moreover, a strong relation between facilitation and satisfaction is observed which emphasizes the importance of a patient-centered phase during the medical consultation (Byrne et al, 1976; Wolf et al, 1978; Mishler, 1984).

The scale "insight" measures the impact of the exchange of information on the patient's comprehension of his condition and treatment. The original theoretical dimensions of providing information and insight collapsed into one dimension because patients do not make a distinction between them. This is understandable since the dimensions are theoretically related, although the findings are in conflict with Tuckett's observations (1985) that information which is conveyed effectively does not automatically entail comprehension by the patient. With regard to the content of the exchanged information, patients expect to receive a comprehensible explanation of the nature, the course and the prognosis of their complaint(s) or disease (Korsch et al, 1972). Moreover, they expect to be informed about the ins and outs of the proposed treatment. Explanations of the patient's psychological functioning represent the upper part of the scale. In general, insight seems to be a legitimate outcome of the consultation because the provision of information can change attributions, is able to resolve unfounded fears and enhances the patient's sense of mastery of his situation. Especially in ambulatory care, information appears to be of importance because patients are expected to deal with their complaint(s) themselves after they have left the consultation room (Vuori et al, 1972). Both the content of exchanged information and the quality of the physician's interviewing skills during the presentation of solutions are involved in increasing the patient's insight (Stiles et al, 1979; Tuckett et al, 1985).

The scale "intention to comply" measures the patient's determination to cooperate with the proposed advice and treatment. The pioneer of research into the medical interview, Davis, observed the relation between the quality of physicians' interviewing skills and patients' adherence to the medical regimen (Davis, 1966; Davis, 1968). This finding was later supported by other studies (Korsch et al, 1972; Ley, 1980; Ley, 1983). The scale "intention to comply" is composed of three statements pertaining to variables which can be influenced by communication: namely, recall, understanding and responsibility, and one statement measuring the patient's commitment to compliance. Ley (1983) elaborated the positive relation between recall and comprehension to compliance, while Eisenthal and Lazare (1976) described the confirmative impact of responsibility when patients are given the opportunity to negotiate about their request for help.

The scale "disrupted communication" measures the patient's opinion of the occurrence of interview behavior that is generally considered to impair communication and to enhance feelings of confusion and anxiety. Medical jargon, hurried interviews and the lack of verbal tracking (the changing rapidly from one subject to another by the physician) are well-known examples (DiMatteo et al, 1983; Ivey, 1983; Mishler, 1984).

The scale "directivity" was originally constructed to measure a specific quality of individual psychotherapeutic relations. It refers to an authoritarian communication style in which the physician acts from his own frame of reference. The dimension was operationalized as paying no respect to the patient's initiative or pace in acknowledging problems or solutions (Lietaer, 1976). Moreover, attempts to change the patient's emotions or behavior were considered to form expressions of directivity. Unfortunately, the significance of high or low levels of directivity for the quality of physician-patient communication has not yet been established.

Several hypothesized dimensions were not retraced after the subsequent factor analyses and Rasch analyses. These dimensions were obviously not distinguished by patients or not well measured by the items in the checklist. Although the original checklist contained several items pertaining to satisfaction, no distinct dimension "satisfaction" came forth. Items referring to satisfaction emerged in

the factor solution facilitation, but after Rasch analysis, they were eliminated from the scale "facilitation". Apparently, a strong connotation of facilitation with satisfaction exists. This finding calls into question the validity of global satisfaction measurements, because patients do not seem to evaluate the quality of the communication in terms of satisfaction. Moreover, the hypothesized dimension "physician's understanding", was not retraced in the analyses. In retrospect, it is clear that the dimension "physician's understanding of the patient's situation" should be put rather to the physician than to the patient. In close relation to the previous observation is the notable absence of the theoretical dimension "empathy". Although the items in this checklist were chosen from the scales of Truax and Carkhuff (1967) and carefully adapted to the Dutch situation by Lietaer (1976), our subjects failed to distinguish empathy as an important dimension in the communication. The finding supports Hogan's (1969) assertions about the lack of validity in measurements of empathy.

This study reveals that within patient satisfaction with communication, several dimensions can be discerned. Patient satisfaction on its own is not recognized as a separate dimension. We therefore recommend the use in future studies of these measurements of the patient's opinion as they enhance our understanding of the relation between process and well-defined outcomes of a medical interview. Furthermore, we consider it to be a shortcoming of our study that no items on anxiety-reduction and reassurance were included in the original checklist. Additional research into these dimensions seems to be necessary.

The measurement properties of the scales in the PSCC are very attractive notwithstanding the dramatic loss of items. Firstly, each scale measures only one dimension: this is an important achievement as it increases our understanding of the impact of the medical interview on the patient. Secondly, all item groups but one fit in well with the assumptions of the Rasch-model. Directivity fits only marginally. The results of the Rasch analyses enable researchers to determine the patient's opinion of a pertinent dimension with considerable precision. All items show a strong similarity between the observed proportion

positive answers and the probability of a positive answer according to the estimations of the Rasch-model. Moreover, the observed proportion positive answers on an item increases considerably when patients' ability scores exceed the item difficulties. One of the problems encountered during the operationalization of several dimensions is the formulation of items with very low or very high item difficulties. In the present study, this problem arose with regard to the dimension "directivity" and, to some degree, with "insight" and "negative communication". As a result, violations of the Rasch-model may take place.

A feature of the Rasch-model is that conclusions on the measurement properties of the scales are dependent on the population studied but independent of the specific sample of subjects. The construction of the Patient Satisfaction with Communication Checklist was based on a combined sample of subjects from two populations: namely, real patients who visited their own general practitioners and simulated patients who were interviewed by residents in general practice. Since many researchers may question this procedure, we studied the influence of group membership on item responses, generally known as "item-bias" (Mellenbergh, 1985). Item-bias was assessed by computing the confidence intervals of the estimates of item parameters over both samples and by comparing the overlap of the intervals. The results show that group membership is not associated with the responses on the items with the exception of items 9 and 12. Real patients and simulated patients approach the dimensions similarly. With regard to items 9 and 12, simulated patients report earlier that they are able to recall the provided information and they agree less easily that they had responsibility for the choice of their treatment. The former is easily understood because simulated patients were informed approximately 20 times about the treatment of "their" disease; this enhances recall of information. The latter is explained by the fact that simulated patients are not really in need of treatment because they are making no real request for help. This makes them less involved in decision-taking. In conclusion, we observe that the Patient Satisfaction with Communication Checklist can be applied to both populations because real patients and simulated patients discern the

same dimensions in their satisfaction with the communication and approach the dimensions similarly.

The advantage of this scale construction procedure is that identical checklists can be given to subjects of both populations which allows comparisons between and additional studies of the validity of simulated patients. Moreover, checklists filled in by simulated patients provide detailed feedback to medical students on dimensions which are also valued by real patients. Medical education can thus be attuned to the reality of daily practice.

In summary, facilitation, insight, intention to comply, disrupted communication and directivity are dimensions in physician-patient communication which are discerned and evaluated by patients. The formulation of theoretical dimensions and the application of adequate procedures in the process of scale construction form important tools for the development of scales that measure only one well-defined dimension. Hopefully, when they are applied in future research, the scales will enhance our understanding of the impact of physician-patient communication on the patient.

REFERENCES

Byrne PS, Long BEL. Doctors talking to patients: a study of the verbal behavior of general practitioners consulting in their surgeries. HSMO, London, 1976.

Crijnen AAM, Dalen J van, Kraan HF, Zuidweg J. Medische interviewvaardigheden gemeten: de Maastrichtse Anamnese en Advies Scoringslijst (Measuring medical interviewing skills: the Maastricht History-taking and Advice Checklist). Medisch Contact, 1986; 41: 114-116.

Davis MS. Variations in patients' compliance with doctor's orders: analysis of congruence between survey responses and results of empirical investigations. Journal of Medical Education, 1966; 41: 1037-1048.

Davis MS. Variations in patient's compliance with doctor's advice: an empirical analysis of patterns of communication. American Journal of Public Health, 1968; 58: 274-288.

DiMatteo MR, DiNicola DD. Achieving patient compliance: the psychology of the medical practitioner's role. Pergamon Press, New York, 1982.

Eisenthal S, Lazare A. Expression of patient's request in the initial interview. Psychological Reports, 1977; 40: 131-138.

Eisenthal S, Koopman C, Lazare A. Process analysis of two dimensions of the negotiated approach in relation to satisfaction in the initial interview. Journal of Nervous and Mental Disease, 1983; 171: 49-54.

Gustafsson JE. The Rasch-model for dichotomous items: theory, applications and a computer program. Reports from the institute of education, No. 64, University of Göteborg, Sweden, 1977.

Hambleton RK, Cook LL. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977; 14: 75-96.

Hogan R. Development of an empathy scale. Journal of Consulting Clinical Psychology, 1969; 33: 307-316.

Ivey AE. Intentional interviewing and counseling. Wadsworth, Belmont, California, 1983.

Korsch BM, Negrete VF. Doctor-patient communication. Scientific American, 1972; 227: 66-74.

Lebow J. Consumer satisfaction with mental health treatment. Psychological Bulletin, 1982; 91: 244-259.

Lebow J. Research assessing consumer satisfaction with mental health treatment. Evaluation and Program Planning, 1983; 6: 211-236.

Ley P. Giving information to patients. In: Eisser JR (Ed.). Social psychology and behavioral medicine. Wiley, London, 1980.

Ley P. Patients' understanding and recall in clinical communication failure. In: Pendleton D, Hasler J (Eds.). Doctor-patient communication. Academic Press, London, 1983.

Lietaer G. Nederlandstalige revisie van Barrett-Lennard's Relationship Inventory voor individueel-therapeutische relaties (The relationship inventory of Barrett-Lennard: Dutch revision for therapeutic relationships). Psychologica Belgica, 1976; 15: 73-94.

Locker D, Dunt D. Theoretical and methodological issues in sociological studies of consumer satisfaction with medical care. Social Science and Medicine, 1978; 12: 283-292.

Martin-Löf P. Statistical models. Notes from seminars, 1969-1970, by Rolf Sunberg (2nd ed.). Institute för försäkrings matematika och matematik statistik vid Stockholms Universitet, Stockholm, 1973.

Meerling. Methoden en technieken van psychologisch onderzoek (Methods and techniques of psychological research - part 1 and part 2). Boom, Meppel, 1981.

Mellenbergh GJ. Vraag-onzuiverheid: definitie en onderzoek (Item bias: definition and research). Nederlands Tijdschrift voor de Psychologie, 1985; 40: 425-435.

Mishler EG. The discourse of medicine: dialectics of medical interviews. Ablex, Norwood, 1982.

Molenaar IW. Programma beschrijving van PML voor het Rasch-model (Description of the PML-program for the Rasch-model, version 3.1). Heymans Bulletin, Vakgroep Statistiek en Meettheorie, Universiteit van Groningen, Groningen, 1981.

Nie N, Hull CH, Jenkins JG, Steinbrenner K, Bent DH. Statistical package for the social sciences. McGraw-Hill, New York, 1975.

Pendleton D. Doctor-patient communication: a review. In: Pendleton D, Hasler J (Eds.). Doctor-patient communication. Academic Press, London, 1983.

Stiles WB, Putnam SM, Wolf MH, James SA. Interaction exchange structure and patient satisfaction with medical interviews. Medical Care, 1979; 17: 667-679.

Stillman PL, Burpeau-Di Gregorio MY, Nicholson GI, Sabers DL, Stillman AE. Six years of experience using patient instructors to teach interviewing skills. Journal of Medical Education, 1983; 58: 941-946.

Thorndike RL. Applied psychometrics. Houghton Mifflin Company, Boston, 1982.

Truax CB, Carkhuff RR. Toward effective counseling and psychotherapy. Aldine, Chicago, 1967.

Tuckett D, Williams A. An approach to the measurement of explanation and information giving in medical consultations: a review of empirical studies. Social Sciencec and Medicine, 1984; 18: 571-580.

Vuori H, Aako T, Aine E, Erkko R, Johansso R. The doctor-patient relationship in the light of patients' experiences. Social Science and Medicine, 1972; 6 (6): 723-730.

Ware JE, Davies-Yvery, A, Stewart AL. The measurement and meaning of patient satisfaction. Health and Medical Care Services Review, 1978; 1: 1-15.

Ware JE, Snyder MK, Wright WR, Davies AR. Defining and measuring patient satisfaction with medical care. Evaluation and Program Planning, 1983; 6: 247-263.

Wolf MH, Putnam SM, James SA, Stiles WB. The medical interview satisfaction scale: development of a scale to measure patient perceptions of physician behavior. Journal of Behavioral Medicine, 1978; 1: 391-401.

CHAPTER 7 SCALABILITY AND RELIABILITY OF THE MAASTRICHT HISTORY-TAKING AND ADVICE CHECKLIST IN GENERAL PRACTICE

A.A.M. Crijnen and H.F. Kraan

7.1 Introduction.

The instrumental utility of methods of measurement is usually expressed in three parameters: scalability, reliability and validity. "Instrumental utility" concerns the degree to which the empirical measurements reflect the theoretical concepts. "Validity" refers to the relation between the concept under study and the variables which represent this concept. "Scalability" and "reliability" contribute to the quality of measurement.

In this chapter, we elaborate on the issues of scalability and reliability with regard to the Maastricht History-taking and Advice Checklist in General Practice (MAAS-GP). We decided to study first the scalability of our measurements and then the reliability because the groups of items constituting the final scales formed the starting point for reliability studies and further validity studies (see chapter 8). In section 7.2, we describe the research setting and the two extensions developed to study reliability and scalability. In section 7.3, the scalability of the MAAS-GP is studied by means of Rasch analyses. In section 7.4, the inter-observer reliability of each of the 68 MAAS-GP items is assessed by means of generalizability studies. In section 7.5, the inter-observer reliability on the level of MAAS-GP-scales is examined by means of generalizability studies. Finally, in section 7.6, we summarize our findings and delineate the utility of the MAAS in terms of scalability and reliability.

7.2 Research setting.

Scalability and reliability were studied in the same research setting described more extensively in chapter 8. During simulated consultation hours, 40 residents in General Practice interviewed several simulated patients (Crijnen et al, 1986). Residents were asked to behave as if they had taken charge of a colleague's practice and to perform a complete medical consultation.

Simulated patients presented two difficult but frequently occurring somatic problems (myocardial infarction and inception of diabetes mellitus). Reliability and scalability studies are based on MAAS-GP-recordings of these (videotaped) interviews.

The first extension was formed by independent observations of 20 randomly chosen medical interviews (10 diabetes mellitus cases and 10 myocardial infarction cases) by 6 observers which provided the data for assessing the reliability of the MAAS. The second extension was formed by increasing the number of 40 original interviews between residents and patients simulating the myocardial infarction case to a total of 100 interviews which enabled us to examine the scalability of the MAAS-GP by means of Rasch analyses. Since Rasch analyses are independent of the sample of subjects studied but dependent on the population, we decided to compile a heterogenous group of physicians and medical students. The number of 100 observations was achieved by observing 40 videotaped consultations of residents in General Practice who interviewed the patient as part of the simulated consultation hours, 29 residents in General Practice who interviewed the patient as part of the residency educational program, 7 general practitioners from the staff of the residency program who took part in the simulated consultation hours, and 24 third year medical students who interviewed the simulated-patient during examination of their medical skills at Maastricht Medical School. All interviews were observed by experienced observers: four of them had been actively participated in the process of MAAS-GP-construction.

7.3 Scalability of the MAAS-GP.

The central issue during the process of scale construction is to bring together a set of items all of which measure to a satisfactory degree the trait of interest and which collectively reflect different levels of possession of this trait. The construction of the six scales in the MAAS-GP and the theoretical considerations of their content are described in chapter 4. The study of the scalability of the six MAAS-scales is elaborated below.

Scale construction is guided scientifically by scaling models that formalize the relation between a subject's responses on items and indices representing a subject's ability on the intended dimension. In the literature, two scaling models are known that are based on either the latent trait models or the classical test theory. In this thesis, we rely heavily on the Rasch-model, a one-parameter logistic model, which is the most demanding but also the most attractive latent trait model. The advantages of the Rasch-model are that item parameters are invariant across samples of subjects chosen from the population of interest, and that a subject's ability can be estimated on the same ability scale from any subset of items that have been fitted in the model. The Rasch-model assumes that all items have equal discriminating power and that the items vary only in terms of their difficulty. Moreover, it assumes that each item in the scale measures only one latent trait, and that a subject's performance on one item will not affect performance on other items.

7.3.1 Method.

To provide data for Rasch analyses, 100 interviews between physicians/medical students and simulated patients who presented complaints accompanying a myocardial infarction were observed and scored on the MAAS-GP-scales by experienced observers. Since Rasch analyses require dichotomized data variables in the scales, "interpersonal skills" and "communicative skills" were dichotomized according to predetermined criteria.

Rasch analyses were carried out using the PML-programm (Gustafsson, 1977). A sequence of analyses was carried out. First of all, items that did not fit in the Rasch-model were selected by means of the Binomial test and Allerup's Graphical test (Molenaar, 1981). Secondly, the scalability of item groups was analyzed by means of the Martin Löf chi-square-test. The Martin Löf chi-square test assesses the fit between the observed proportion positive answers for a specific item on each ability level and the estimated probability of a positive answer according to the assumptions of the Rasch-model. Construction of Rasch homogeneous scales was secured by eliminating items which appeared to diminish the fit between the estimated probability of a positive answer

according to the Rasch-model and the empirical pattern of responses to the group of items which constituted each scale. Thirdly, the unidimensionality of each MAAS-scale was tested by determining whether pairs of Rasch homogeneous scales, which were considered to measure distinct traits, could be positioned on one scale. This occurred by means of a second type of Martin Löf chi-square test which establishes unidimensionality by estimating the differences of person parameters for pairs of scales.

7.3.2 Results.

Statistics from the Martin Löf chi-square tests for the six MAAS-GP scales, such as chi-square, degrees of freedom, probability and number of items, are presented in table 7.1. A high level of probability indicates a good fit of the itemgroup to the assumptions of the Rasch-model.

Table 7.1.: Martin Löf Chi-square-test which tests the fit of the item group to the assumptions of the Rasch-model.

MAAS-GP scales	N of items	Chi-Square	D.F.	Proba-bility**	K.R.-20
Exploring reasons for encounter	8 / 7	22.19	24	.57	.27
History-taking	23 / 11	69.98	70	.48	.48
Presenting solutions	12 / 11	75.88	90	.86	.86
Structuring	8 / 6	23.46	20	.27	.60
Interpersonal skills	10 / 8	43.07	42	.42	.52
Communication skills	7 / 6	30.19	20	.07	.66

*) Number of items before and after Rasch analyses.
**) A high level of probability indicates a good fit of the item group to the assumptions of the Rasch-model.

The item-difficulties of items that appeared to fit in the Rasch homogeneous scales are presented in the right-hand column of table 7.2 as a summarizing statistic. Item-difficulties reflect the point on the ability-scale where subjects have a 50% chance of scoring an item positively.

Statistics from the Martin Löf chi-square test for unidimensionality, such as chi-square, degrees of freedom, probability and Pearson product-moment correlation between pairs of scales are presented in table 7.3. A low level of probability on the Martin Löf chi-square test for unidimensionality indicates that scales are measuring distinct dimensions.

Table 7.3: Results of the Martin Löf chi-square test for unidimensionality of the scales in the MAAS-GP.

Pairs of scales	chi-square	D.F.	Probability*	Pearson product-moment correlation
EE with HT	63	76	.86	.21
EE with PS	69	76	.71	.11
EE with STR	42	41	.43	.38
EE with IPS	45	69	.99	.16
EE with COM	54	48	.26	.27
HT with PS	119	120	.50	.05
HT with STR	76	65	.16	.28
HT with IPS	78	109	.99	.20
HT with COM	96	96	.06	.21
PS with STR	25	21	.24	.13
PS with IPS	80	99	.92	.16
PS with COM	88	69	.06	.14
STR with IPS	32	59	.99	.50
STR with COM	42	41	.43	.46
IPS with COM	33	69	.99	.50

*) A low level of probability on the Martin Löf chi-square test for unidimensionality indicates that scales are measuring distinct dimensions.

Table 7.2.: Generalizability coefficients of MAAS-GP-items for one, two and six observers and item difficulties computed during Rasch analyses.

MAAS-GP items (abbreviated)	Generalizability			Item-difficulty
	1 obs	2 obs	6 obs	
EXPLORING REASONS FOR ENCOUNTER				
01 Asks about the reason for encounter	.34	.51	.76	-1.23
02 Explores the emotional impact	.14	.25	.50	-0.84
03 Asks to clarify why at this moment	.48	.65	.85	+0.56
04 Asks opinion about cause	.60	.75	.90	-0.87
05 How complaint is discussed in family	.67	.80	.92	
06 What help desired	.59	.74	.90	-0.28
07 How patient solved the problem himself	.61	.76	.91	-0.04
08 Influence on daily life	.32	.49	.74	+2.75
HISTORY-TAKING				
09 Asks to describe the complaint	.53	.69	.87	-1.52
10 Explores the intensity	.39	.56	.79	+0.24
11 Asks about localization	.62	.77	.91	-1.52
12 Asks about shifts/radiations	.89	.94	.98	-1.85
13 Asks about the course during day	.64	.78	.91	
14 Asks about history	.57	.73	.89	
15 Asks which factors triggered	.35	.51	.76	
16 Asks which factors increased	n.c.	n.c.	n.c.	+1.33
17 Asks which factors maintained	.00	.00	.00	
18 Asks which factors decreased	.11	.20	.43	+1.33
19 Asks about accompanying circumstances	.35	.52	.71	
20 Explores gains of the complaints	n.c.	n.c.	n.c.	
21 Explores both soma and psyche	.06	.11	.26	+0.40
22 Explores relationships in family	.77	.87	.95	
23 Explores professional functioning	.93	.96	.99	
24 Explores leisure-time functioning	.54	.70	.88	+0.58
25 Explores vulnerability factors	.18	.31	.57	
26 Asks about past illnesses	.71	.83	.94	+0.53
27 Asks about professional treatment	.52	.68	.86	+1.79
28 Asks about present consultations	.48	.65	.85	
29 Asks about medication (ab)use	.61	.76	.91	-0.25
30 Asks about hereditary aspects	.77	.87	.95	
31 Reviews the system of the complaint	.27	.42	.69	

PRESENTING SOLUTIONS				
32 Explains diagnosis understandably	.14	.24	.49	-1.80
33 Explains causes	.43	.60	.82	-0.14
34 Gives information about prognosis	.54	.60	.82	+2.20
35 Explores expectations	.44	.61	.82	+1.25
36 Proposes solutions	.24	.38	.65	-3.46
37 Explains appropriateness of solution	.19	.31	.58	-0.96
38 Discusses pros and cons of solutions	.28	.43	.70	+1.71
39 Explores different points of view	.35	.52	.77	
40 Asks for the intention to comply	.05	.09	.23	+1.14
41 Expl. how advice should be carried out	.26	.41	.67	-0.24
42 Checks patient's understanding	.05	.10	.25	+1.03
43 Makes appointments for follow-up	.43	.60	.82	-0.74
STRUCTURING				
44 Introduces himself	.62	.77	.91	-0.09
45 Offers agenda	.33	.49	.74	+0.69
46 Summarizes the reasons for encounter	.27	.42	.69	+0.47
47 Concludes history-taking with ordering	.08	.15	.35	+0.97
48 EE precedes HT	.20	.33	.60	-1.05
49 Explores EE and HT sufficiently	.15	.28	.52	-0.99
50 Begins PS with diagnosis	.10	.18	.40	
51 Main complaint discussed satisfactorily	.37	.53	.78	
INTERPERSONAL SKILLS				
52 Facilitates communication	.17	.29	.55	+0.55
53 Reflects emotions properly	.33	.50	.75	-0.29
54 Reacts properly to emotions	.20	.33	.60	
55 Asks about feelings during interview	.06	.11	.32	+1.93
56 Makes metacommunicative comments	.16	.27	.53	+3.77
57 Performs history-taking properly	.14	.25	.50	-0.14
58 Puts the patient at ease	.06	.11	.28	+0.35
59 Sets proper pace during the interview	.08	.15	.35	-1.44
60 Appropriate non-verbal behavior	.00	.00	.01	-4.75
61 Makes proper eye-contact	.00	.00	.00	
COMMUNICATION SKILLS				
62 Uses closed-ended questions properly	.00	.00	.00	
63 Concretizes properly	.02	.04	.10	+1.18
64 Makes proper summaries	.29	.44	.71	-1.66
65 Provides information in small amounts	.00	.00	.00	-0.60
66 Checks patient's understanding	.00	.00	.00	-0.00
67 Makes proper confrontations	.03	.06	.17	+3.20
68 Uses comprehensible language	.11	.19	.42	-1.66

n.c. : not computable

7.3.3 Discussion.

In the following paragraphs, we discuss the scalability of the six MAAS-GP-scales determined by means of Rasch analyses.

The scale "exploring reasons for encounter" fits in the Rasch-model after elimination of one item. The items occupy a broad range of item difficulties. The scale thus has adequate measurement properties meaning that it is able to distinguish levels of ability in the skill to explore reasons for encounter. The eliminated item is, unfortunately, very important, as it addresses the issue of a patient's interpersonal relations and social support system. The fact that it is eliminated may suggest that it just does not fit in the Rasch-model or that it measures a different trait. Since we made no theoretical distinction between this item and the remaining items, we do not think that it is measuring a different trait. The test for unidimensionality reveals furthermore that the skill to explore reasons for encounter is not different from any of the other medical interviewing skills; this suggests that they are all measuring interviewing skills.

The scale "history-taking" fits in the Rasch-model and the items use a broad range of item difficulties. The scale thus displays adequate measurement properties. To increase the fit of the scale in the Rasch-model, the number of items was reduced significantly from 23 to 11. Since both the included and excluded items refer to the physician's questioning behavior about medical topics, and since we made no theoretical distinctions between the items (see also chapter 2 and 4), it is unlikely that the excluded items measure a different trait. The test for unidimensionality discloses, furthermore, that history-taking is not different from any of the other types of interviewing skills, although it is differentiated slightly from communication skills.

The scale "presenting solutions" fits in the model after exclusion of one item. The items display a wide range of item difficulties. The scale thus has attractive measurement properties. The item referring to the negotiation between physician and patient about problem-definition and treatment was eliminated from the original scale during the process of scale construction. Since the remaining items pertain to the exchange of information, we presume that the item on negotiation measures a different trait. Furthermore, the ability to present

solutions is essentially not different from other interviewing skills.

The scale "structuring the interview" corresponds well to the Rasch model, although two items had to be eliminated. Since the eliminated items are theoretically not different from the remaining items in the scale, we do not expect them to measure a different trait. The items in this scale utilize only a limited range on the continuum of item difficulties which means that the confidence intervals of the estimates of item parameters will show some overlap. The measurement properties of this scale are thus considered to be less optimal when compared to the other scales and violations of the Rasch-model are to be expected. The skill to structure the medical interview is identical to other medical interviewing skills according to the test for unidimensionality.

The scale "interpersonal skills" fits well in the Rasch-model and the items appear to display a broad range of item difficulties which enhances their measurement properties. During the process of scale construction, two items were excluded pertaining to the physician's ability to handle negative emotions directed at himself and to the quality of eye-contact. Since no theoretical distinction was made between these items and the other components of interpersonal skills, we do not expect them to measure a distinct trait. Moreover, observers often experienced difficulties in scoring the quality of eye-contact because we were not able to define strict criteria. The "interpersonal skills" scale, which measures the physician's ability to establish a warm human relationship with the patient, is not different from any of the other scales measuring medical interviewing skills.

The scale "communication skills" fits only marginally in the Rasch model and it is interesting to note that the item referring to the quality of closed-ended questions, which is generally regarded as essential, is eliminated from this scale. Despite the fact that the items display a wide range of item difficulties, the scale has therefore no optimal measurement properties. The fact that communicative skills are not really different from "exploring reasons for encounter", "structuring" and "interpersonal skills", whereas rather low probabilities are obtained for "history-taking" and "presenting solutions" on the test for unidimensionality, is attributed

for the moment to the low reliability of this scale. Further research with reliable instruments is necessary to establish the dimensionality for this type of skills.

Latent trait models have strong advantages when compared to classical test models (Hambleton and Cook, 1977). Since the MAAS-GP-scales fit in the Rasch-model, they can take advantage of the characteristics of this demanding latent trait model. One advantage is that it is possible to estimate the quality of a physician's medical interviewing skills on each of the MAAS-GP-scales from any subset of items that proved to confirm to the model. The estimates of item difficulty are free from the influence of the ability distribution in the calibrating sample of subjects. This forms an important advantage for the construction of tests used in examination or evaluation situations because small tests can be constructed which are less demanding for the observers, because parallel forms of tests can be developed and compared, and because scales which are endorsed by subjects from different classes, institutions or even different nations can be compared.

A second advantage is that the estimates of item parameters are independent of the sample of students/physicians used for the statistical analyses, but dependent on the population. In our situation, data for computations were obtained on a rather heterogeneous sample of subjects, ranging from inexperienced third year medical students to residents in general practice and even to general practitioners with years of experience in daily practice who all interviewed simulated patients. Item and scale characteristics are therefore assumed to be stable for a population of medical students, residents in general practice and experienced general practitioners while they are interviewing simulated patients. The limits of the population are less well defined. We are unable to say whether the same results hold for surgeons, psychiatrists or cardiologists, etc. Moreover, we are unable to defend the hypothesis that the fit of the scales in the Rasch-model holds true for interviews between physicians and real patients in daily practice, although the general practitioners and residents considered the simulated consultation hours to be an appropriate simulation of real life (Crijnen et al, 1986).

A third advantage of the Rasch-model is that it treats each of these six types of medical interview behavior as a hierarchy of skills: physicians who are able to perform the more difficult skills well are also likely to perform the less difficult skills well, whereas physicians who experience difficulties with the less difficult skills are not able to perform the more difficult skills well. It would be productive for educationalists to take a look at the item difficulties and apply them to their educational programs.

To summarize: the MAAS-GP-scales of medical interviewing skills, with the exception of communicative skills, fit well in the Rasch-model and are considered to have adequate measurement properties. Furthermore, only one dimension is distinguished in our measurements of medical interview behavior. All scales measure medical interviewing skills.

7.4 Reliability of 68 MAAS-GP items.

In situations where the process of measurement is heavily dependent upon human observers, major problems with regard to the standardization of measurement are likely to occur. The quality of measurement in terms of inter-observer reliability has therefore to be studied in greater depth. In the following paragraphs, we report on the inter-observer reliability of the individual MAAS-GP-items determined by means of generalizability studies (Cohen, 1960; Cronbach et al, 1972; Shrout et al, 1979; Guilford et al, 1981).

Observers are usually regarded as potential sources of error in the measurement of observational data. Psychometricians have therefore developed a variety of indices which all purport to reflect inter-observer agreement and reliability (Tinsley et al, 1975; Mitchell, 1979). Cohen's kappa determines the proportion of agreement between observers corrected for chance (Cohen, 1960). Intra-class correlations reflect the ratio of the variance of interest over the sum of the variance of interest plus error (Shrout et al, 1979). Finally, Cronbach et al (1972) developed the theory of generalizability which makes use of estimates of variance components to determine several reliability coefficients for a variety of situations about which a researcher wishes to generalize. We applied all three procedures to our

data and compared the results. The generalizability coefficients for one observer were similar to Cohen's kappa and intra-classs correlations for one observer, whereas the generalizability coefficients for six observers resembled strongly the average intra-class correlations. Furthermore, generalizability coefficients provided information about the reliability of items when pairs of observers were used. We therefore decided to report only generalizability coefficients and to leave out Cohen's kappa and intra-class correlations.

7.4.1 Method and results.

Generalizability coefficients can be computed in several ways. As it is our intention to provide information about the reliability of MAAS-GP-items for random samples of observers, estimates of variance components were computed by means of the General Mixed Model Analysis of Variance with Equal Cell Sizes (BMDP-program P8V - Dixon et al, 1979) in which the physicians and observer facets were considered to be random, and in which each item was considered to be fixed. Data for computations were provided by twenty videotaped medical interviews all independently observed by six well-trained observers who filled in all the MAAS-GP-items.

Three types of generalizability coefficients were calculated using the formulae described by Thorndike (1982). The first type, presented in the left-hand column of table 7.2, provides information about the typical reliability of a single observer's scores for each item, a very common observation situation. The generalizability coefficient for one observer has to be interpreted as the correlation between scores of two randomly chosen observers after observing the same sample of videotaped interviews. For example, the correlation between scores of two observers on item 1 will be .34.

The second type of generalizability coefficient, presented in the second column of table 7.2, provides information on the reliability of an item for pairs of observers. The final score of a subject on an item is the summed score of both observers. This generalizability coefficient is interpreted as the correlation between summed scores for pairs of observers with summed scores of a different pair of randomly

chosen observers who observed the same medical interviews. Generalizability coefficients for pairs of observers were assessed because computations for the validity studies elaborated in chapters 8 and 9 and future studies about the nomological net are based on sumscores of observations by continually changing pairs of observers. Adding a second observer appears to be a powerful remedy against unreliability attributed to observer influences because the relative importance of errors of measurement made by observers is reduced.

The third generalizability coefficient, presented in the third column of table 7.2, provides information on the reliability for a group of six observers who all observed the same interviews. Since an increase in the number of observers enhances reliability, and since it is almost impossible to base analyses on larger groups of observers due to limited resources, the generalizability coefficients for six observers for each item can be considered to display the upper limits of reliability achieved by MAAS-GP-items.

7.4.2 Discussion.

Items in the scale "exploring reasons for encounter" display moderate generalizability coefficients. Appending a second observer significantly diminishes unreliability due to observer influences. The coefficients for six observers are very high, with the exception of item 2 which measures questioning behavior about the emotional impact of the complaint. The moderate reliability of this item is probably due to the fact that this issue is often elicited by a reflection of the patient's emotions rather than by question behavior, and that it is often difficult to discern what are the main and what are the related complaints.

Items in the scale "history-taking" show moderate to high levels of inter-observer reliability. Appending a second observer has a considerable impact on the reliability of items with moderate generalizability coefficients. In general, history-taking skills refer to single acts of interview behavior. They are therefore more easily defined and subsequently more easily recognized by observers when compared to other medical interviewing skills. Items in this scale display the upper limits of reliability to be achieved by observational

measurements of medical interviewing skills. Moderate reliability is reported for items referring to factors that appear to influence the complaints. This moderate reliability is attributed to an observer's confusion when a physician asks questions with a specific content which are recognized in the literature as influencing complaints such as, for example physical exercise, smoking, etc., instead of a general question about factors which trigger or increase the problem. Reliability is expected to be enhanced by sharper definitions. Moreover, low reliability is reported for the exploration of both somatic and psychological determinants, mainly a threshold problem. Observers generally agree on the topics elicited, but disagree on the depth of exploration.

Items in the scale "presenting solutions" show low to moderate levels of inter-observer reliability. Appending a second observer considerably enhances reliability. Low reliability is reported for the explanation of diagnosis, probably due to the fact that explanations are often mixed up with an explanation about the cause of the problem when they are presented in lay terminology. Low reliabilities for items pertaining to the patient's intention to comply and the check on the patient's understanding are probably due to the often implicit performance of this interview behavior. Sharpening of definitions by requiring explicit interview behavior is likely to enhance reliability.

Items in the scale "structuring the interview" show low to moderate levels of inter-oberver reliability. Adding a second observer enhances reliability, although still only moderate levels of reliability are achieved. A feature of this scale is that it attempts to measure phases of the interview and the quality of transitions. Observers are asked to recognize the phases and to qualitfy the transitions, a fairly difficult task because demarcations in a medical interview are, apart from at the beginning and the end, never that clear. We expect sharpening of definitions to enhance reliability at the cost of a decrease in validity. When, for example, a topic stemming from the patient's frame of reference is brought forward and discussed in a few sentences during history-taking, despite the fact that the physician has adequately explored the reasons for encounter, the question is raised of whether item 48 "Explores the reason for encounter before

history-taking" has to be scored "yes" or "no". The burden of taking this decision rests on the observer. Strict observation of interview behavior is not possible with regard to this kind of interviewing skill because a considerable degree of interpretation is always involved. The sizes of the generalizability coefficients adequately reflect the difficulties observers encounter during observation and scoring of medical interviewing skills which refer to the structuring of the interview.

Items in the scale "interpersonal skills" show low generalizability coefficients. Adding a second observer almost doubles the size of the coefficients, although still only low to moderate levels are achieved. Generalizability coefficients for six observers, representing finite reliability, even reach only a low to moderate level. Achievement of high reliability is hampered by our inability to define interpersonal skills in single acts of easily observed interview behavior. Some items require observers to position themselves empathically in the communication and to indicate whether they feel comfortable in it. Other items require observers to indicate whether what we call "non-behavioral" terms would be more appropriate, such as the item on meta-communication which tries to assess whether, in the case of inhibited communication, physicians fail to make meta-communicative comments. Inter-observer reliability is unlikely to be enhanced by sharpening of definitions since we have already tried to construct and define the items as behaviorally as possible. The question is how should the low to moderate reliability of this scale be treated: should we abandon the scale or should we accept this level of reliability because the scale measures - or at least attempts to measure - an important quality of the physician's medical interview behavior? In this thesis, items that fitted in the Rasch-model are applied during the validity studies. For future studies, we recommend the reconstruction of the scale "interpersonal skills". Many items referring to a variety of aspects of interpersonal skills should be included in the process of scale construction and the items should be described in behavioral terms in order to increase reliability and should be well defined in an accompanying manual. Reconstruction of this scale will not be easy: researchers should be wary of impairing the validity of the scale by wording items too behaviorally.

The scale "communicative skills" displays low levels of inter-observer reliability with the exception of the item on the quality of summaries. Increasing the number of observers has no impact on reliability. The low reliability is, in our opinion, due to the fact that the observers are required to make a summarizing judgment about frequently occurring and, in itself, already difficult to assess interview behavior, such as closed-ended questions and the provision of information in small units. We do not expect sharpening of definitions and better training of observers to increase reliability. A better approach for future studies might be to apply a different method of measurement in which the nature and quality of each utterance of the physician is determined and coded separately. With regard to this scale, the question is raised of whether it can be accepted as part of the MAAS-GP. Because the scale attempts to measure a theoretically important quality of medical interviewing skills, and because the data are available, we included the scale in the validity studies presented in this thesis.

We have analyzed and discussed here the inter-observer reliability for each of the MAAS-GP items separately. High levels of reliability are seen when items are worded in behavioral terms. Moderate reliability is observed when larger units of interviewing behavior are measured. Low reliability is reported in items which require interpretation by the observers. High inter-observer reliability is seen for the scale "history-taking", moderate reliability for "exploring reasons for encounter", "presenting solutions", and "structuring", whereas low reliability is observed for the scales "interpersonal skills" and "communicative skills". Inter-observer reliability is enhanced significantly when a second observer is added.

7.5 Inter-observer reliability on scale-level: a generalizability study.

Since the summed scores of items in each scale are treated in our study as indices of the quality of six distinct types of medical interviewing skills in the MAAS-GP, we decided to analyze the reliability of these measurements on scale level. Generalizability theory appears to be the most appropriate approach for the analysis of the size of a number of sources of variation in our measurements in one

and the same analysis. Analysis of variance and the subsequent estimation of variance components provides information about the contribution of each of the sources of variation. Based on these figures, generalizability coefficients can be computed which objectify the reliability of the measurements under a variety of situations (Cronbach et al., 1972); Mitchell, 1979; Thorndike, 1982; Nunnally, 1982).

7.5.1 Method: analyses and results.

Analysis of variance and generalizability theory ideally requires data to be gathered in a completely crossed, multidimensional design. In our research situation, 20 videotaped interviews between physicians and a total of 5 simulated patients who presented complaints of either a myocardial infarction or the inception of diabetes mellitus, were all observed by six obervers on all the items of the MAAS-GP. The research situation is described more extensively in chapters 5 and 8. Analyses of variance were carried out for each MAAS-GP scale by means of the General Mixed Model Analysis of Variance with Equal Cell Sizes (BMDP-program P8V - Dixon and Brown, 1979) in which the physician and observer facets were considered to be random, and in which the item facet was considered to be fixed.

The size of the variance components was estimated for the facets physicians, observers and items, and for the interaction between physicians and observers, the interaction between physicians and items, the interaction between observers and items, and the interaction between physicians, observers and items including error. These estimates of variance components were used to calculate the percentage of the total variance induced by each component and the estimates provided the necessary information to compute generalizability-coefficients.

A summary of the estimates of variance components for the respective scales is presented in table 7.4.

Table 7.4: Percentage of variance of each MAAS-GP scale atributed to different sources of variation.

Source of variation	EE	HT	PS	STR	IPS	COM
1. physician	8.9	20.1	7.0	7.1	4.1	2.2
2. observer	2.1	0.2	3.5	2.9	0.9	6.6
3. item	21.8	6.9	19.0	14.4	26.4	16.3
4. p.o.	5.2	2.7	6.5	10.4	10.5	14.6
5. p.i.	24.5	31.5	16.4	17.0	5.2	2.7
6. o.i.	2.0	2.2	6.6	4.6	6.8	12.7
7. o.i.p + error	35.4	36.4	41.0	43.7	46.2	44.7

The results of the analyses of variance for each scale are presented in the appendix 9. Generalizability coefficients, in which the item component is considered to be fixed, were calculated by means of the formulae obtained from Thorndike (1982).

The items facet became a fixed facet rather than a facet that is sampled randomly from a larger universe of items, becasue the sets of items were selected carefully to constitute together the domain of interviewing skills in which we are interested. The universe of interviewing skills pertaining to each dimension is limited and it is the researchers' impression after reviewing the literature (see chapters 2 and 4) and after observing many interviews, that most interview behavior is covered by one of the MAAS-GP items. Moreover, the sets of items constituting the scales reflect the universe about which we wish to generalize because the groups of items used in our scalability, reliability and validty studies are always identical. Generalizability coefficients for increasing numbers of observers are presented in table 7.5.

Table 7.5: Generalizability coefficients for increasing numbers of observers of the MAAS-GP (over Rasch homogeneous scales).

Number of observers	EE	HT	PS	STR	IPS	COM
1	.49	.77	.37	.32	.21	.08
2	.65	.88	.54	.48	.35	.17
3	.74	.92	.64	.59	.45	.21
4	.78	.93	.70	.65	.51	.26
5	.80	.95	.74	.70	.57	.30
6	.84	.96	.78	.74	.62	.35
10	.91	.98	.90	.82	.72	.46

Unfortunately, we did not extend our generalizability study to a completely crossed design which also included the simulated patients and the cases. Out of curiosity, and because we wished to increase our understanding of the influence of simulated patients and cases on MAAS-GP measurements of medical interviewing skills, analyses of variance and subsequent generalizability coefficients were computed for each of the two cases and each of the five simulated patients. The generalizability coefficients for the two cases are presented in table 7.6, whereas the estimates of variance components are presented between brackets in the discussion section where applicable.

Table 7.6: Generalizability coefficients over MAAS-GP scales for myocardial infarction case and inception of diabetes mellitus case for one and two observers.

	EE	HT	PS	STR	IPS	COM
1 observer						
myocardial infarction	.43	.61	.26	.32	.21	.13
diabetes mellitus	.15	.58	.28	.27	.21	-.01
2 observers						
myocardial infarction	.60	.76	.34	.49	.34	.23
diabetes mellitus	.25	.74	.43	.43	.34	.02

To compare physicians' responses on the scales for each case, mean scores and t-tests over the scale-totals were computed (see table 7.7).

Table 7.7: **Mean scores and standard deviation for both cases on MAAS-GP scales (Rasch homogeneous) and t-tests between the cases (df=78).**

Scale	Myocardial infarction		Diabetes mellitus		t-value	probability (2-tail)
	m	sd	m	sd		
exploring reasons for encounter	2.7	1.5	2.1	1.1	-2.18	0.03
history-taking	3.6	1.8	0.7	0.9	-9.22	0.00
presenting solutions	5.4	1.9	6.7	1.4	3.70	0.00
structuring	2.8	1.6	2.8	1.5	0.00	1.00
interpersonal skills	4.4	1.4	4.9	1.5	1.54	0.13
communicative skills	3.1	1.6	2.6	1.3	-1.39	0.17

7.5.2 Discussion.

The first facet, which reflects differences between physicians, reveals that most scales measure a considerable amount of variance which can be attributed to the physicians (see table 7.4). The first facet represents, together with the interaction between physicians and items, the true-variance in which we are primarily interested. Since the quality of measurement of the traits had already been secured by Rasch analyses, the amount of variance that can be attributed to the first facet is not really of significance to us. We can still see that the scale "communication skills" has poor measurement properties because the variance component attributed to physicians appears to be very small.

The second facet, reflecting differences between observers, reveals varying degrees of an observer's influence on the variance in the

MAAS-GP-scales. This implies that differences between observers with regard to the overall severity of their grading standards will produce variation in the estimates of a physician's competence. Saal argues that a significant observer main effect, especially one that explains a sizable proportion of the rating variance, has to be interpreted as the traditional leniency effect, generally defined as the tendency of observers to assign a higher or lower rating than is warranted by a subject's behavior (Saal et al, 1980). Leniency refers to the phenomenon that some observers will indicate that behavior occurred, whereas others will state that it occurred insufficiently or only partly, although all of them observed the same behavior. This influence is lowest in the scales "history-taking" and "interpersonal skills", moderate in "exploring reasons for encounter", "presenting solutions" and "structuring" and greatest in the scale "communicative skills". Scales referring to larger units of interview behavior which are difficult to define and for which the criteria for scoring are less well described are more likely to be affected by a leniency effect. Remarkable is that the scale "interpersonal skills" is scarcely influenced by the leniency effect. Stricter definitions and training sessions for observers to reach agreement are ways to reduce the leniency effect in affected scales.

The third facet, reflecting differences between items, shows that a considerable amount of variance is induced by the items which form the scales. The interpretation of the item facet posed problems for the researchers because of the difficulty resulting from systematic variance being induced by items which are identical for each observation situation. We attributed the differences between items to a different source of variation in our study: namely, the combination of simulated patient and medical problem. Considerable differences on the third facet between the two cases are observed for the scales "exploring reasons for encounter", "history-taking" and "presenting solutions" (respectively 11.9%, 17.8% and 17.1% for myocardial infarction, and 48.2%, 4.0% and 36.0% for diabetes mellitus). Dissimilarity in case presentation by the simulated patients and differences between the medical problems are considered to be responsible for these effects.

The fourth source of variation, reflecting the interaction between physicians and observers, influences the scores on the scales in varying degrees. Once again, the scale "communication skills" is affected most significantly but this influence also acts upon structuring the interview and interpersonal skills. The interaction between physicians and observers is known in the literature as "halo-effect" and it is defined as an observer's failure to discriminate between conceptually distinct and potentially independent aspects of a subject's behavior (Saal et al, 1980).

Occurrence of halo-effects suggests that one characteristic of a subject will influence an observer's opinion on a variety of items. Halo-effects are likely to emerge when the behavior under study is not well defined or when a substantial degree of judgment is involved in answering an item. Halo-effects were expected to occur to a certain degree in the scales "structuring the interview", "interpersonal skills" and in "communication skills" because these scales measure either larger units of difficult-to-define interview behavior or require observers to indicate their personal opinion of a physician's skills. However, halo-effects were not expected to contribute that significantly to the measurement of "communication skills" because, beforehand, we had considered the interviewing skills in this scale to be defined more in behavioral terms and that they could be measured more easily when compared to interpersonal skills. Apparently, an observer's impression of the skill to communicate is reduced to a more global judgment and this probably occurs because the scale measures interview behavior which takes place more than once and, sometimes, even many times in the course of a consultation. No differences between the two cases were observed with regard to the impact of halo-effects on the MAAS-GP-scales.

To diminish the influence of halo-effects, items should be well-defined, a difficult requirement with regard to these scales. Items in the scales "structuring" and "interpersonal skills" are already described as behaviorally as possible. We consider that the rewording of these items to achieve more behavioral descriptions of interviewing skills would seriously impair the validity of the scales. With regard to "communication skills", we have earlier proposed the classification of every utterance of the physician. A quite different

approach might be the selection for measurement purposes of those observers who are known to evoke only low degrees of halo-effects. Nevertheless, it is our opinion that, even with this approach, halo-effect has to be accepted as a measurement feature of these scales.

The fifth source of variation, reflecting the interaction between physicians and items, is considered in the literature as indicating true variance in addition to the physician's facet (Thorndike, 1982). This source of variation refers to differences between physicians with regard to their interviewing styles. The figures show that interviewing styles are most pronounced in exploring reasons for encounter and history-taking, less in presenting solutions and structuring the interview and scarcely present at all in interpersonal skills and communication skills. Considerable differences in impact on the interaction between physicians and items were observed between the two cases, especially in "exploring reasons for encounter", "history-taking" and "presenting solutions" (respectively 30.7%, 25.6% and 13.3% for myocardial infarction case and 10.2%, 43.4% and 6.2% for inception of diabetes mellitus case).

The sixth source of variation, reflecting the interaction between observers and items, is strongest for communication skills, less for interpersonal skills, presenting solutions and structuring, and almost absent for exploring reasons for encounter and history-taking. The interaction between observers and items refers to differences between observers in interpreting the meaning of items and the criteria for scoring. The figures reveal that single acts of operationally defined interview behavior are interpreted very similarly, whereas situations in which the observer has to match the occurrence of interview behavior with MAAS-GP-items and their definitions are inclined to induce differences in interpretation. The item "explains diagnosis or problem definition understandably", for example, is difficult to score because the observer has to decide what of everything said by the physician to the patient pertains to the diagnosis, whether it provides an explanation and whether it is presented understandably. This source of error variance is assumed to be minimized by the use of additional resources such as a manual, instructions and articles which increase observers' understanding of the behavior under study. Moreover, the

training of observers by means of group observations of videotaped interviews is likely to reduce this effect. Since four of the six observers are experts as far as the MAAS-GP is concerned as they participated actively in the construction, the figures presented here are considered as reflecting the upper limits of agreement that can be achieved among observers on the interpretation of MAAS-GP-items. Moreover, only minor differences between the two cases are observed.

The size of the seventh source of variation reflects that the interaction between physicians, observers and items, including error, forms a considerable source of variation in our measurements. This suggests that, in addition to the known and controlled sources of variation, other influences can act upon the variation of our measurements. Error is one of these influences. Physicians' interview behavior is considered to be affected by conditions such as fatigue, motivation, willingness to participate, etc., whereas observers' ratings are likely to be influenced by fatigue, mood, inaccuracy and pressure of time. Moreover, the simulated patients who play opposite to the physicians and the cases they are presenting, will contribute to the patient-physician communication and therefore will induce variation in our measurements. Differences between the two cases with regard to the size of the seventh source of variation are observed for the scales "exploring reasons for encounter" and "presenting solutions".

In conclusion, physicians, observers and items are all capable of inducing variance in MAAS-GP measurements of medical interviewing skills as is evidenced by the size of the different variance components presented in table 7.4. Furthermore, the cases and/or simulated patients appear to elicit different amounts of true variance and induce differences in the facet items, the interaction of physicians and items and the interaction of physicians, items and observers including error. This phenomenon undermined the design of the generalizability study to some extent and posed interpretation problems for the researchers. This issue therefore is elaborated upon below.

The generalizability study was conducted to determine, in one study, the impact of physicians, observers and items on MAAS-GP measurements of medical interviewing skills and to calculate generalizability coefficients which were expected to provide information about the effect of strategies to remedy error induced by observers. This is of

importance because the number of observers participating in data recording is the only source of variation that can be manipulated by the researchers. The generalizability coefficients for one observer (see table 7.5) reveal that an acceptable level of reliability is achieved for the scale "history-taking", whereas the other scales display low levels of reliability (Mitchell, 1979). Adding a second observer increases the reliability of "exploring reasons for encounter", "history-taking", and "presenting solutions" to a moderate or even high level. "Structuring" barely reaches an acceptable level, whereas "interpersonal skills" and "communication skills" gain the least advantage from an increase in the number of observers to enhance reliability; this is probably because these scales do not measure a large component of true variance. It is clear that an increase in the number of observers mitigates error induced by observers and diminishes differences between observers, halo-effects and different interpretations of items. Reliability is most effectively increased by adding one or two observers. The addition of even more observers to the process of measurement will not add substantially to reliability. In conclusion, we observe that it is possible to alleviate the influence of observers on our measurements of medical interviewing skills by incorporating one or two additional observers in the process of measurement.

The reliability of the MAAS-GP measurement was studied by observing medical interviews of 20 physicians who talked with one out of five simulated patients who presented one out of two cases. We decided to study reliability over two different cases in order to enhance external validity. Although this design was completely crossed for physicians, observers and items, we did not include simulated patients and cases in the completely crossed design. Since the question can be raised of whether this procedure is acceptable, we studied the impact of the case as a medical problem, of simulated patients as human beings interacting with their physician, and of simulated patients as performers of certain tasks on our measurements of medical interviewing skills. Swanson et al (1981) studied the stability of medical interviewing skills over different cases and concluded that patients and cases seem to differ greatly and that the physician's adaptation to these differences is complex and difficult to measure.

The mean scores, which are displayed in table 7.7, of 40 physicians on the scales in the MAAS-GP for both cases reveal significant differences for the scales "exploring reasons for encounter", "history-taking" and "presenting solutions". Beforehand, we had expected differences to show up only for the scale "history-taking" because the cases that were presented differed considerably with regard to the amount of information necessary to solve the medical problem. The mean scores reveal that, in accordance with expectation, almost no questions were asked in the diabetes mellitus case, whereas several questions were asked in order to solve the myocardial infarction case. The generalizability coefficients for the scale "history-taking" are, however, considerable and almost identical for both cases. This suggests that, given a certain medical problem, differences in history-taking skills are determined solely by differences between physicians and - of course - observers. The case will not interfere with the measurement of physicians' interviewing skills during history-taking. The situation is rather different for "exploring reasons for encounter" and "presenting solutions", because the medical problem is not conceived as significantly influencing these interviewing skills. The simulated patients, however, are expected to have a considerable impact, especially because these phases of the interview enable them to voice their concerns and to obtain information about and treatment for their complaints. The simulated patients in the diabetes mellitus case were trained to worry about the imminent disease and to insist on obtaining information, whereas the patients in the myocardial infarction case were much less demanding. The generalizability coefficients presented in table 7.6 show considerable differences between the cases in "exploring reasons for encounter", "presenting solutions" (after diminishing interfering observer influences) and "communication skills". The results suggest that the goals which patients try to achieve in the interview interfere - a better term might be "interact" - with the measurement of differences between physicians with regard to the quality of their interviewing skills. An alternative explanation might be that differences between the simulated patients as private persons will influence the interaction between physician and patient. To study this alternative explanation, generalizability coefficients over the scale "exploring

reasons for encounter" were computed for each of the five simulated patients. Since chance capitalization was likely to occur because the number of physicians was very small, these results have to be treated cautiously. For the patients in the myocardial infarction case, the coefficients were respectively .51, .35, and .56, and for the diabetes mellitus case, .06 and .18. The results support the hypothesis that patients, as private persons, elicit different amounts of true variance because different coefficients were reported within each case but, more importantly, the coefficients were low for the diabetes case and moderate for the myocardial infarction case. Since the coefficients were rather stable within each case, they also support the hypothesis that the goals which patients try to achieve have a considerable impact on physicians' interviewing skills.

During the construction of the Patient Satisfaction with Communication Checklist, identical problems were encountered because patients did not distinguish between the dimensions "insight" and "providing information". It is clear that, during an interaction, both participants influence each other and constitute together the reality of a medical interview. Unfortunately, our methods of measurement are only able to measure either the physician's interviewing skills or the patient's opinion about the consultation and they are unable to deal with the interaction.

Given that we now know that a medical interview forms a dialogue between physician and patient and that both contribute to the communication, the following question is raised: what is the reliability of the MAAS-GP? Physician and patient influence each other and the interviewing skills which are displayed by the physician during a medical consultation are therefore partly dependent upon his interviewing style and partly dependent upon the patient's contribution. Reliability, on the other hand, is defined as the ratio of true variance and observed variance. True variance in a physician's medical interviewing skills is assumed to reflect the differences in style and quality that are attributed to the physician, but it is clear that a physician's skills are also dependent upon the patient's contribution to the communication and upon the interaction between physician and patient. An example will clarify this issue.

Generalizability coefficient for two observers for "exploring reasons for encounter" over both cases was .65 (moderate), over the myocardial infarction case, .60 (moderate) and over the diabetes mellitus case, .25 (low). Is reliability moderate or low? This question cannot be answered because both are true. We therefore conclude that the reliability of interactional data should be studied cautiously and that results should be interpreted more in a comparative way than absolutely as we have in the discussion of the estimates of variance components. Our study reveals, furthermore, that demanding patients who insist on discussing specific topics yield less information about physician's/student's ability to perform a medical interview, an important finding for teaching and evaluation situations. The patient's influence is most pronounced during the "exploration of reasons for encounter" and the "presentation of solutions", whereas "history-taking" is influenced considerably by case-differences. Future studies are necessary to study this issue in greater depth.

7.6 Conclusion.

In this chapter, the scalability and reliability of the Maastricht History-taking and Advice Checklist in General Practice have been studied. The scalability was assessed by means of Rasch analysis which determines whether all items of a scale are, to a satisfactory degree, measuring the trait of interest and whether the group of items collectively reflect different levels of possession of this trait. MAAS-GP-scales of medical interviewing skills fit well in the Rasch model with the exception of "communication skills". The scales are considered as having adequate measurement properties. Additional analyses reveal that our scales of medical interviewing skills are measuring only one dimension.

Since the process of measuring medical interviewing skills is heavily dependent upon human observers, inter-observer reliability was studied. High levels of inter-observer reliability are seen when items are worded in behavioral terms. Moderate reliability is observed when larger units of interview behavior are measured. Low reliability is reported for items which require considerable interpretation by observers. High inter-observer reliability is seen for the scale

"history-taking", moderate reliability for "exploring reasons for encounter", "presenting solutions", and "structuring", whereas low reliability is observed for the scales "inter-personal skills" and "communicative skills". Inter-observer reliability can be enhanced significantly by adding a second observer.

Additional generalizablity studies reveal that leniency, halo-effects, and differences in interpretation impair the quality of measurement to varying degrees. These studies reveal, furthermore, a complex interaction between physicians' interviewing skills and the goals patients try to achieve during a medical consultation. Simulated patients who insist on discussing certain topics yield less information about a physician's interviewing skills during the "exploration of reasons for encounter" and the "presentation of solutions". We conclude, therefore, that the reliability of interactional data on scale level should be studied cautiously and that results should be interpreted comparatively rather than absolutely.

REFERENCES

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960; 20: 37-46.

Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. The dependability of behavioral measurements: theory of generalizability for scores and profiles. John Wiley and Sons, New York, 1972.

Crijnen AAM, Thiel J van, Kraan HF. Evaluatie van consultvoering: een spreekuur nagebootst (Evaluation of a medical consultation: simulated consultation hours). Huisarts en Wetenschap, 1986; 29: 316-318.

Dixon WJ, Brown MB. Biomedical Computer Programs, P-series. University of California Press, Berkeley, 1979.

Guilford JP, Fruchter B. Fundamental statistics in psychology and education. McGraw-Hill, London, 1981.

Gustafsson JE. The Rasch-model for dichotomous items: theory, applications and a computer program. Reports from the Institute of Education, University of Göteborg, no. 63, Sweden, 1977.

Hambleton RK, Cook LL. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977; 14: 75-96.

Molenaar IW. Programma beschrijving van PML voor het Rasch-model (Description of the PML-program for the Rasch-model, version 3.1). Heymans Bulletin, Vakgroep Statistiek en Meettheorie, Universiteit van Groningen, Groningen, 1981.

Nunnally JC. Reliability of measurement. In: The encyclopedia of educational research. McMillan and Free Press, New York, 1982.

Saal FE, Downey FG, Lahey NA. Rating the ratings: assessing the psychometric quality of rating data. Psychological Bulletin, 1980; 88: 413-428.

Shrout PE, Fleiss JL. Intraclass correlation: uses in assessing rater reliability. Psychological Bulletin, 1979; 86: 420-428.

Swanson DB, Mayewski RJ, Norsen L, Baran G, Mushlin AI. A psychometric study of measures of medical interviewing skills. In: Proceedings of the 20th Annual Conference of Research in Medical Education, Washington, 1981.

Tinsley HEA, Weiss DJ. Interrater reliability and agreement of subjective judgments. Journal of Counseling Psychology, 1975; 22: 358-376.

Thorndike RL. Applied psychometrics. Houghton Mifflin Company, Boston, 1982.

CHAPTER 8 CONVERGENT AND DIVERGENT VALIDITY OF FOUR METHODS OF MEASUREMENT OF MEDICAL INTERVIEWING SKILLS: A MULTITRAIT-MULTIMETHOD APPROACH

A.A.M. Crijnen and H.F. Kraan

8.1 Introduction.

Assessing the validity of measurements of medical interviewing skills is rarely done. In 1955, Cronbach and Meehl recommended establishing the validity of such measurements in order to understand the constructs which account for variance in test performance. More recently, two editorials have emphasized the need for reliable and valid measurement of medical competence and medical interviewing skills and have suggested strategies for increasing the validity of the assessment of medical competence (Katz, 1982; Gonella, 1985). Katz proposed increasing validity by observing physicians in real or simulated situations, whereas Gonnella proposed the establishment of the convergent and divergent validity of measurements of medical competency. In this study, we respond to these recommendations by testing the convergent and divergent validity of the Maastricht History-taking and Advice Checklist in General Practice.

Establishing convergent and divergent validity is one of the procedures suggested by Cronbach and Meehl (1955) and further elaborated by Campbell and Fiske (1959) for validating the meaning of measurements. It presumes that when two tests measure the same construct, a substantial correlation between the tests will emanate. Campbell and Fiske argue that demonstration of convergent and divergent validity requires that several measurements of the same construct confirm the meaning of a construct whereas, at the same time, measurements of other constructs are expected to support their distinct character.

Convergent validity is then indicated by substantial correlations between independent measurements of similar constructs, whereas divergent validity is indicated by low correlations between measurements of unrelated constructs. The essential notion in Campbell and Fiske's approach is that each measurement forms a combination of a

trait and a method: measurements combine a particular content with a measurement procedure which is not specific to that content. They therefore suggest the determination of the relative contributions of trait and method variance to measurements by applying more than one trait as well as more than one method in the validation process. Support for convergent and divergent validity was obtained by framing a multitrait-multimethod matrix (MTMM-matrix), consisting of the correlations between a number of traits each measured by several methods. This matrix had to be examined subsequently according to four criteria proposed by Campbell and Fiske. These criteria refer to the convergence of the traits, to the divergence of the traits, to the divergence of the methods and to a pattern of relations between the traits. Several authors recommend the examination of the multi-trait-multimethod matrix as an ideal validation procedure in test development (Kerlinger, 1981; Thorndike, 1982). Although recommended, the MTMM-validation procedure has not often been applied to psychological research because of its demanding criteria.

In the present study, we investigate the convergent and divergent validity of the Maastricht History-taking and Advice Checklist in General Practice. In addition to the MAAS-GP, three distinct measurements of medical interviewing skills are applied in order to assess the quality of physicians' medical interviewing skills. The multitrait-multimethod matrix, consisting of the correlations between the four methods of measurement, is constructed and subsequently examined according to Campbell and Fiske's criteria.

8.2 Examination of the multitrait-multimethod-matrix according to the Campbell and Fiske criteria.

In this section, we elaborate the construction of the multitrait-multimethod matrix. Moreover, Campbell and Fiske's criteria and their relation to the matrix are presented. A MTMM-matrix consists of the correlations resulting when each of several traits is measured with each of several methods. Table 8.1 displays an example of a MTMM-matrix for two traits (1,2) and two methods (A,B).

Table 8.1: Theoretical multitrait-multimethod-matrix for two traits (1,2,) and two methods (A,B).

		Method A		Method B	
		1	2	1	2
Method A	1	rA1A1			
	2	rA1A2	rA2A2		
Method B	1	rA1B1	rA2B1	rB1B1	
	2	rA1B2	rA2B2	rB1B2	rB2B2

Each trait is measured by each method and, subsequently, correlated. The values on the diagonal (rA1A1, rA2A2, rB1B1, rB2B2) are reliability indices, mostly coefficients of internal consistency. The diagonal in the Method A/Method B-block, called "validity diagonal", contains the correlations between one trait measured by different methods (rA1B1, rA2B2). The value(s) in the Method A1/Method A2-triangle (rA1A2) or Method B1/Method B2 (rB1B2)-triangle are called "heterotrait-monomethod-triangles". The value(s) in the Method A1/Method B2-triangle (rA1B2) or Method A2/Method B1-triangle (rA2B1) are called "heterotrait-heteromethod-triangles".

Convergent and divergent validity are assesed by inspecting the MTMM-matrix according to the four Campbell and Fiske criteria (Campbell et al, 1959; Schmitt et al, 1986).

The first criterion refers to the convergence of independent methods with regard to the measurement of a similar trait. It states that values on the validity diagonal should be large enough to warrant further examination of validity. As a minimal requirement, correlations on the validity diagonals (rA1B1 and rA2B2) ought to be statistically significant.

The second criterion pertains to the verification of distinctions between traits. This criterion states that values on the validity diagonal should be higher than the heterotrait- heteromethod correlations of the column and row in which the individual validity value is located. This criterion can be studied by examining whether a correlation on the validity diagonal (rA1B1) exceeds the strength of the correlations on the adjoining column and row (rA1B2 and rA2B1).

The third criterion determines the extent to which method variance contributes to the scores. It states that values on the validity diagonal must be higher than the off-diagonal values in its monomethod triangle. Variables should correlate higher with measurements of the same trait obtained by different methods of measurement than with other traits measured with the same method. This criterion can be studied by examining whether correlations on the validity diagonals (rA1B1 or rA2B2) exceed the (median of the) correlation(s) of the monomethod-heterotrait block (rA1A2 or rB1B2).

The fourth criterion states that the patterns of trait inter-relationship should be the same in all heterotrait triangles in both monomethod and heteromethod blocks. Campbell and Fiske have not developed procedures to establish this criterion unequivocally.

In summary, convergent and divergent validity are ascertained when measurements behave according to these four criteria.

Over the course of time, several more advanced statistical procedures for analyzing MTMM-matrices have been developed (Marsh et al, 1983; Schmitt et al, 1986). Confirmatory factor analysis (LISREL-Joreskog, 1974) is usually seen as the most appropriate model for evaluating convergent and divergent validity of MTMM-matrices. We have tried to apply LISREL-analyses to our MTMM-matrix, but because of technical problems, we have been unable to analyse our data thoroughly. Moreover, analyses of variance are not recommended for the study of MTMM-matrices because, for among other reasons, there is no clear equivalence between the ANOVA-effects and the Campbell and Fiske criteria (Marsh et al, 1983). We therefore approach the analyses of our MTMM-matrix of measurements of medical interviewing skills by means of the original Campbell and Fiske criteria.

8.3 Method.

In the following section, we describe the four instruments which measure medical interviewing skills, the six traits that are discerned, the subjects, the experimental setting and the analyses.

8.3.1 Four instruments measuring medical interviewing skills.

The methods of measurement employed in the MTMM validation procedure are expected to meet several requirements. Firstly, the methods are required to be completely independent of each other. Secondly, the instruments are required to measure the traits as conceptualized appropriately. Fiske (1971) elaborated the first requirement by stating that the procedures to obtain scores consist of a chain of events between the original behavior of the subject and the assignment of an index. He distinguished three features in this process: the modes for measuring personality, the method of data-recording and the process of indices production.

We finally achieved the construction of four independent methods of measurement of medical interviewing skills by varying the modes of measurement and the procedures of indices construction. Two methods of measurement, namely, the MAAS-GP and the Global Expert-Rating Scale, utilize the observation of behavior mode, whereas two other measures, namely, the MAAS-self and the Global-Self Rating Scale, make use of the self-description mode. In the MAAS-GP and MAAS-self, researchers constructed the indices, whereas in the Global Expert-Rating Scale and the Global Self-Rating Scale, indices were produced by, respectively, experienced general practitioners and the interviewing physicians. Each measure therefore differed as much as possible from the others. In table 8.2, the modes, the indices producers and the resulting measurement instruments are depicted.

In the following paragraphs, each of the four methods is briefly characterized.

In the Maastricht History-taking and Advice Checklist in General Practice, trained observers indicated whether each of 68 discernable units of interview behavior occured in the course of a medical consultation. Items are organized around six scales, pertaining to six

Table 8.2: Modes, indices producer and resulting methods of four methods of measurement of medical interviewing skills.

Mode	: observation of behavior		self-description	
Indices producer	: researcher	expert	researcher	inter-viewing physician
Method	: 1.MAAS-GP	3.GERS	2.MAAS-self	4.GSRS

theoretical dimensions in medical interviewing skills (see chapters 2 and 4). Researchers combined item responses to these scales to construct indices of the physician's competency on the six dimensions. Scalability was secured by using items which appeared to fit to the Rasch-model, whereas reliability was enhanced by adding a second observer to the process of measurement. Summed scores of pairs of observers for each scale were used to constitute the MTMM-matrix. Issues of scalability and reliability are elaborated in chapter 7. The instrument is classified as an observation of behavior mode with indices production by the researchers.

The Maastricht History-taking and Advice Checklist-self (MAAS-self) is a self-assessment instrument. The MAAS-self was constructed by a slight transformation of the original MAAS-GP-items to clear self-descriptions of behavior, such as "I provided information about the cause of the presented problem". This transformation was possible because of the high face-validity of the MAAS-GP which was constructed to form a feedback tool in medical education (see chapter 4). The MAAS-self has to be filled in by the interviewing physicians at the end of a medical consultation. Physicians are requested to indicate whether they have performed each of 68 units of interview behavior in the course of the preceding interview. The units of interview behavior are present or absent. Responses on MAAS-self items identical to the items which formed the Rasch homogeneous scales in the MAAS-GP were combined

by researchers to constitute the indices of six dimensions in medical interviewing skills. This method is classified as a self-description mode with indices production by the researchers.

In the Global Expert-Rating Scale, experienced general practitioners rate the quality of six dimensions of a physician's medical interviewing skills on a global evaluative rating scale after observing a medical interview. To make their evaluative ratings, experts dispose of only rather implicit criteria which are based on the face validity of the items. Since no further definitions or criteria for scoring are given, experts are considered to be the indices producers. They have to respond with degrees of agreement or disagreement (5-point Likert-scale) to a set of six items each pertaining to one of the dimensions discerned in medical interviewing skills. This method utilizes the observation of behavior mode whereas indices are produced by general practitioners.

The Global Self-Rating Scale (GSRS) has to be completed by the interviewing physicians themselves after the interview with a (simulated) patient. Physicians are asked to report their subjective impressions of the quality of the interview on six dimensions. As in the previous method, only global definitions of the dimensions are given, based on the face validity of the items. The interviewing physicians have to evaluate the quality of their own interview behavior on rather implicit criteria and are therefore required to produce the indices themselves. Items are rated on 5-point Likert-type scales. This method is classified as a self-description mode whereas the indices are produced by the physicians.

Campbell and Fiske require the methods used in the MTMM- validation procedure to be independent in order to minimize the influence of method co-variance. Of the methods presented here, we can assume that MAAS-GP and GERS share method-variance because both are observation instruments, whereas MAAS-self and GSRS are likely to share method-variance because both are self-description instruments. Furthermore, MAAS-GP and MAAS-self have in common a similar set of behaviorally described items and a similar procedure of indices formation. GERS and GSRS partially converge in that both are global evaluative rating scales without a priori criteria or definitions. We

therefore have to acknowledge that even within discernible methods of measurement, convergent and divergent elements can be recognized simultaneously.

8.3.2 Six dimensions in medical interviewing skills.

In addition to the independency of methods, Campbell and Fiske require that the traits are also independent. This requirement was posed to achieve near-zero values in the heterotrait-heteromethod and the heterotrait-monomethod triangles, and to maximize differences between the validity coefficients, the trait intercorrelations and the method influences. On reviewing these requirements, Fiske (1971) stated that one can never establish that, in an empirical sense, a trait is uncorrelated with all other traits; this mitigated the requirement for trait independency. It is therefore sufficient that traits are theoretically distinct before they are employed in a MTMM validation procedure.

The theoretical considerations leading to the classification of medical interviewing skills into six dimensions are elaborated in chapters 2 and 4 of this thesis. We confine ourselves here to summarizing the characteristics of the dimensions.

The skill to "explore the reasons for encounter" refers to a physician's ability to clarify the patient's complaint and to explore the motives in the pre-patient phase which led to the patient visiting the physician. "History-taking" skills enable the physician to generate hypotheses about the nature of the complaint, to test these hypotheses and to describe the patient's complaint in medical explanatory terms. During the "presentation of solutions", physicians convey information on causes and prognosis of the medical problem; they negotiate with the patient about the medical problem and its solutions and they provide concrete information on the approach in the near future. Some interviewing skills enable the physician to "structure" the medical interview. Moreover, "interpersonal skills" enable the physician to establish an optimal rapport with the patient, whereas "communicative skills" are apt to promote an effective exchange of information between patient and physician and vice versa. These six dimensions constitute the elements of an appropriate initial interview in primary health care. A summary of methods and dimensions is shown in table 8.3.

Table 8.3: **Methods and dimensions used to study convergent and divergent validity of the Maastricht History-taking and Advice Checklist in General Practice.**

Methods	Dimensions
1. Maastricht History-taking and Advice Checklist	1. Exploring Reasons for Encounter
2. Maastricht History-taking and Advice Checklist-self	2. History-taking
3. Global Expert-Rating Scale	3. Presenting Solutions
4. Global Self-Rating Scale	4. Structuring
	5. Interpersonal Skills
	6. Communication Skills

8.3.3 Subjects.

All forty residents of the Department of General Practice who participated in the 1984-1985 residency program took part in this study. These 31 men and 9 women (mean age: 28.6 year) finished medical school at the age of 26.4. Before starting the residency program, they had worked for an average of 10 months in health care. Physicians were not selected on their interviewing skills before they were allowed to participate in the residency program. At the time of the study, 11 residents had almost completed the residency program and the others had just started their training (mean: 5 months in residency program; range: 1-11 months). During the program, the residents received 5-10 hours on average of courses in medical interviewing skills. More than half of these residents had followed the undergraduate curriculum at Maastricht Medical School (23); 11 came from Nijmegen, 1 from Utrecht, 2 from Amsterdam (UvA) and 3 from Groningen.

8.3.4 Research setting.

To secure optimal conditions for measurement, comparability and control, a simulated consultation hour was created in which 40 residents in General Practice interviewed four different simulated patients (Crijnen et al, 1986). Two weeks before the simulated consultation hour took place, residents were informed of the goal and procedures of the study.

During the simulated consultation hours, residents were asked to behave as if they had taken charge of a colleague's practice and to perform a complete medical consultation with each simulated patient. Since physical examinations formed no part of the research setting, information about the patient's physical condition was given to the resident on request. Six rooms with video-equipment were at our disposal in the Skills Laboratory at Maastricht Medical School.

Following a signal, observers switched on the video-equipment and a simulated patient entered the consultation room. Residents were allowed to speak for a maximum of 15 minutes with the patients. During the consultation, six well-trained observers filled in the MAAS-GP (instrument 1) and rated the physician's interview behavior. A second signal, at 15 minutes, indicated that the consultation had to be terminated. The residents then rated their interview on the Global Self-Rating Scale (instrument 4) and the MAAS-self (instrument 2), and they filled in the scale "problem-solving". In addition, simulated patients filled in the Patient Satisfaction with Communication Checklist. A third signal, at 30 minutes, indicated that simulated patients and observers had to change rooms. This procedure was repeated four times.

Several months later, eleven general practitioners recruited from the Department of Family Medicine and the Skills Laboratory, were asked to observe the videotaped medical interviews and to rate the quality of the physicians' medical interviewing skills by means of the Global Expert-Rating Scale (instrument 3). Each interview was observed by two randomly assigned general practitioners. These general practitioners were considered to be experts in general practice because of long experience in general practice and teaching positions in the undergraduate medical curriculum or the residency General Practice Program. Clinical experience was important in order to anchor MAAS-GP-scores to clinical reality and relevance, whereas educational experience was considered to enhance experts' understanding of what constitutes a good medical interview.

All videotaped medical interviews were observed a second time with use of the MAAS-GP by one of three observers. Summed scores on the scales of live and video observations were added to enhance the reliability of measurement.

The MTMM validation procedure is based on case presentations of three simulated patients presenting complaints accompanying a myocardial infarction. The myocardial infarction case, borrowed from a real case-history and documented by a general practitioner and psychologist, described a 50 year-old building contractor who was worried about his heart because he had experienced the night before a short attack of intensive chest pain. The patient had smoked for years and had recently gone through a period of severe problems as a result of the economic recession. The simulated patients were recruited from the Skills Laboratory and were instructed by a psychologist and general practitioner to present the case naturally.

8.3.5 Analyses.

Before constructing the multitrait-multimethod matrix, data were prepared for computations.

Missing data were estimated by regressing variables with missing data on the variable with which they were most highly correlated (BMDP-program PAM-single). In MAAS-GP and GERS, no data were missing. In MAAS-self and GSRS respectively, 0.2% and 0.1% data were missing. For MAAS-self, missing data were estimated and replaced. Since one GSRS was not filled out at all, missing data could not be estimated and this physician's interview was left out of analyses. Six additional observations on MAAS-GP and GERS could not be obtained because these consultations were not videotaped due to technical failure. Since only complete cases were used, data obtained on 33 medical consultations were left for computations. For MAAS-GP, indices of the six traits were computed by summing responses on items that fitted in the Rasch-model (see chapter 7) of live and video observations together. For MAAS-self, indices for the six traits were computed by summing responses on items that appeared to fit in the Rasch homogeneous MAAS-GP-scales. For MAAS-GP and MAAS-self, the scales "interpersonal skills" and "communication skills" were dichotomized according to predetermined criteria. For GERS, indices for the six traits were obtained by adding ratings on each item of pairs of experts who observed the same interview. For GSRS, indices for the six traits were not transformed.

Firstly, mean scores, standard deviations and ranges were computed for each trait measured by each scale.
Results are presented in table 8.4.

Table 8.4: **Scale midpoint, mean, standard deviation and range for scales of MAAS-GP, MAAS-self, GERS and GSRS.**

Instrument	Scale	Scale-midpoint	Mean	S.D.	range
MAAS-GP	EE	3.5	2.7	1.5	0- 5
	HT	5.5	3.6	1.8	1- 7
	PS	5.5	5.4	1.9	2- 9
	STR	3.0	2.8	1.6	0- 6
	IPS	4.0	4.4	1.4	1- 7
	COM	3.0	3.1	1.6	0- 6
MAAS-SELF	EE	3.5	5.1	1.3	2- 7
	HT	5.5	6.9	2.2	2-11
	PS	5.5	7.0	2.3	2-11
	STR	3.0	4.4	1.5	1- 6
	IPS	4.0	4.9	1.8	1- 8
	COM	3.0	3.6	1.5	1- 6
GERS	EE	2.5	3.3	1.1	1- 5
	HT	2.5	3.3	1.1	1- 5
	PS	2.5	3.4	1.1	1- 5
	STR	2.5	3.2	1.0	1- 5
	IPS	2.5	3.6	1.0	2- 5
	COM	2.5	3.3	1.0	2- 5
GSRS	EE	2.5	3.8	0.8	1- 5
	HT	2.5	3.7	0.9	2- 5
	PS	2.5	3.8	0.9	2- 5
	STR	2.5	3.4	1.0	1- 5
	IPS	2.5	3.7	0.8	2- 5
	COM	2.5	3.4	0.9	1- 5

Secondly, the multitrait-multimethod-matrix was constructed by computing Pearson product-moment correlations between the six traits measured by each of the four methods. The median for each validity-diagonal, for each heterotrait-heteromethod triangle and for each heterotrait-monomethod triangle were calculated. The value between the two middle values of each validity-diagonal was considered to be the median. Results are displayed in table 8.5, at the end of this chapter.

To establish the first criterion, the number of significant correlations ($p=\leq.05$) on the validity-diagonals was counted for both traits and methods. Results are displayed in table 8.6.

Table 8.6: **Number of significant correlations ($p=\leq.05$) on the validity-diagonal for each trait and method (maximum is 6 respectively 18).**

EE	HT	PS	STR	IPS	COM	MAAS-GP	MAAS-S	GERS	GSRS
0	3	3	5	5	3	11	11	9	7

To establish the second criterion, the number of times that correlations on the validity-diagonal exceeded the strength of the correlations in the corresponding column and row of the two adjoining heterotrait-heteromethod triangles was counted. Each value on the validity-diagonal was compared to 10 other correlations. Results are displayed in table 8.7.

Table 8.7: **Number of values on the validity diagonal higher than heterotrait-heteromethod values in corresponding column and row (maximum=10).**

Trait	MAAS/ MAAS-s	MAAS/ GERS	MAAS/ GSRS	MAAS-s/ GERS	MAAS-s/ GSRS	GERS/ GSRS	Total
EE	9	4	5	7	4	2	31/60
HT	10	10	4	7	3	7	41/60
PS	10	10	7	3	9	9	48/60
STR	8	10	8	8	9	8	51/60
IPS	7	10	9	5	2	10	43/60
COM	5	6	4	5	10	0	30/60
Total	49/60	50/60	37/60	35/60	37/60	36/60	

To establish the third criterion, the number of times that the three validity-values for each trait exceeded the median of each heterotrait-monomethod triangle was counted. Results are displayed in table 8.8. A second operationalization of the third criterion by comparing the validity-values with each of the correlations in the monomethod triangles separately revealed essentially the same information and is therefore not presented.

Table 8.8: Number of times that the three validity values of each trait are higher than median of a heterotrait-monomethod triangle (maximum=3).

Trait	MAAS-GP (.22)	MAAS-self (.40)	GERS (.63)	GSRS (.44)	Total
EE	0	0	0	0	0/12
HT	2	2	0	0	4/12
PS	2	2	0	1	5/12
STR	3	2	0	0	5/12
IPS	3	1	0	2	6/12
COM	1	1	0	1	3/12
Total	11/18	8/18	0/18	4/18	

To establish the fourth criterion, we considered factor-analyzing parts of our correlation matrix. Due to an insufficient number of participating physicians, factor-analyses would yield unstable results. It was therefore decided not to study the fourth criterion.

8.4 Results.

Inspection of table 8.4 reveals that, according to MAAS-GP-scores, most physicians display only a limited number of the interviewing skills that can be displayed during a medical consultation. Furthermore, averaged MAAS-GP-scores are lower than MAAS-self-scores, and GERS-scores are lower than GSRS-scores. MAAS-GP-scores are under or just above the scale midpoints, whereas scores of MAAS-self, GERS and GSRS are above the scale midpoints. Standard deviations for MAAS-GP and MAAS-self and for GERS and GSRS are almost identical. The range of

scores shows that MAAS-GP-scales never reach the upper limits of scoring, whereas all other scales do reach the upper limits.

With regard to the first criterion, for MAAS-GP and MAAS-self, about 60% of the correlations on the validity diagonal are significant, and for GERS and GSRS respectively, 50% and 40% (table 8.6). The validity of the dimensions "structuring" and "interpersonal skills" is supported almost always; of the dimensions "history-taking", "presenting solutions" and "communication skills", only half of the time; the validity of the dimension "exploring reasons for encounter" is never supported.

Taking into account that support by GERS and MAAS-self is most important, we conclude that the validity of the MAAS-GP-scales "history-taking", "presenting solutions", "structuring" and "interpersonal skills" is confirmed; that the validity of the scale "exploring reasons for encounter" is discredited and that the validity of the scale "communication skills" is neither supported nor discredited.

With regard to the second criterion, inspection of table 8.7 depicts that "presenting solutions" and "structuring the interview", and, to a lesser extent, "history-taking" and "interpersonal skills" can be clearly differentiated from each other. The results support the distinct character of these dimensions. "Exploring reasons for encounter" and "communication skills" are differentiated less clearly from other dimensions.

The number of times that validity-diagonals between pairs of methods exceed the correlations of the adjoining rows and columns, shown in table 8.7, discloses that MAAS-GP/GERS and MAAS-GP/MAAS-self distinguish the dimensions in medical interviewing skills adequately. The other combinations appear to differentiate the dimensions less well. MAAS-GP seems best able to discern different dimensions of medical interviewing skills. GERS and MAAS-self are second and third, whereas GSRS is almost unable to distinguish dimensions of interviewing skills.

With regard to the third criterion, the median of the monomethod triangles, shown in table 8.5, provides information on the considerable impact of the method on the measurement of medical interviewing skills.

The reported differences are amazingly great and vary from a median of .22 for MAAS-GP to a median of .63 for GERS.

The results summarized in table 8.8 reveal that MAAS-GP, followed by MAAS-self, are plagued least by a disturbing influence of the method of measurement. The measurement qualities of both rating scales, GERS and GSRS, are considered to be impaired by strong method influences. Moreover, inter-observer reliability for GERS, expressed in Pearson product-moment correlations between pairs of experts, is low to moderate. Correlations between experts on each item vary from .15 to .42.

8.5. Discussion.

The following section pertains to a discussion of the convergent and divergent validity of the Maastricht History-taking and Advice Checklist in General Practice examined according to Campbell and Fiske's criteria (1955). In addition, the validity of three other methods of measurement of medical interviewing skills are scrutinized.

8.5.1 Criterion 1.

Criterion 1 states that values on the validity-diagonal should be large enough to support convergent validity.

For MAAS-GP, the confirmation of convergent validity by GERS is especially encouraging because it indicates that general practitioners with experience in medical education and primary health care agree with the operationalizations by MAAS-GP-scales of the dimensions "history-taking", "presenting solutions", "structuring" and "interpersonal skills". Moreover, convergent validity of these scales is underscored by physicians' recordings of their own interviewing skills (MAAS-self).

The insufficient evidence of convergent validity for the scales "exploring reasons for encounter" and "communication skills" is disappointing and needs further elaboration. The lack of validity can be attributed to either vagueness of the underlying theoretical concepts, to inadequate operationalization of the theoretical dimension in the items of the scale or to insufficient measurement properties. With regard to the operationalization of the "exploration of reasons

for encounter", one aspect is considered to be missing. In addition to eliciting information about factors in the pre-patient phase leading to the visit, patients should be asked to formulate their request for help explicitly. Item 6 in the MAAS-GP pertains clearly to the patient's request for help, but it is our opinion that more attention should be given to this issue because of its steering influence on content and process of an initial interview. Eisenthal and Lazare (1976, 1983) found that interview behavior which helped the patient to put his request into words was related to feelings of being helped, of satisfaction and plan wanted. Patients find it difficult to verbalize their request for help whereas, at the same time, they consider this to be very important. A structuring activity by the physician and his collaborative involvement stimulates the patient to formulate his request for help. In the scale "exploring reasons for encounter", more items must focus on this issue. A second reason for insufficient support of convergent validity is found in the characteristics of global rating scales which are considered to impair the quality of measurement. This issue is discussed with the third criterion.

"Communication skills", on the other hand, are measured unreliably by means of the MAAS-GP which hinders determination of any form of validity (see chapter 7).

For MAAS-self, evidence of convergent validity is available for the scales "history-taking" and "structuring" and, to a lesser extent, for "presenting solutions" and "interpersonal skills". No evidence of convergent validity is obtained for "exploring reasons for encounter" and "communication skills". The same is essentially true for MAAS-self as for MAAS-GP, but because a classical test-retest design cannot be carried out, unreliability of MAAS-self has to be taken into account as a confounding influence on the validation process.

For GERS, the validity for "history-taking", "presenting solutions", "structuring" and "interpersonal skills" is confirmed; the validity of "exploring reasons for encounter" is discredited, and the validity for "communication skills" is neither supported nor discredited. With regard to the measurement characteristics of global rating scales, it is known that raters are unable to assess more than two dimensions of performance accurately. In medical education, physicians discern mostly

a problem-solving and interpersonal-skill dimension, which largely agrees with the results presented here (Dielman et al, 1980; Streiner, 1985). "History-taking", "presenting solutions" and "structuring" are considered as reflecting the problem-solving dimension, whereas "interpersonal skills" reflects the interpersonal dimension. "Exploring reasons for encounter" and "communication skills" are not clearly discerned by general practitioners.

For GSRS, convergent validity of "interpersonal skills" is unequivocally supported by the validity coefficients. Apparently, a physician's experience of his interpersonal skills displayed during the interview agrees with the impression of MAAS-GP-observers and experts. This is of importance because it confirms the validity of an important but difficult-to-measure quality of a medical interview. Since convergent validity of the other dimensions is only supported by strong correlations with MAAS-self and not by MAAS-GP or GERS, we conclude that the validity of global self-rating scales of medical interviewing skills has to be questioned with the exception of measures of interpersonal skills .

In conclusion, MAAS-GP, MAAS-self and GERS display evidence of convergent validity for "history-taking", "presenting solutions", "structuring" and "interpersonal skills". Insufficient evidence was obtained to support convergent validity of the "exploration of reasons for encounter" and "communication skills". For GSRS, convergent validity is obtained for the measure of "interpersonal skills", whereas the validity of the other measures is discredited.

8.5.2 Criterion 2.

Criterion 2 states that values on the validity-diagonal should be higher than the values of the corresponding column and row in the heterotrait-heteromethod triangle to support divergent validity with regard to the dimensions of interest. Campbell and Fiske's goal was to verify a method of measurement's capability of distinguishing the dimension of interest from several other dimensions. They required the median of each heterotrait-heteromethod triangle to approach zero in order to enhance determination of divergent validity. The median values shown in table 8.5, reveal that none of them approaches zero, which

suggests that the methods and/or the dimensions are related. We expected this to occur because we were not able to construct totally independent methods of measurement and because the theoretical dimensions which were discerned in medical interviewing skills will be related to some extent.

The results, shown in the right-hand column of table 8.7, depict that the dimensions "presenting solutions" and "structuring" and, to a lesser extent, "history-taking" and "interpersonal skills", are clearly differentiated from each other. These results support the distinct character of the dimensions and underscore the theoretical considerations that led to the differentiation of medical interviewing skills into six distinct dimensions. Once again, "exploring reasons for encounter" and "communication skills" are less well discerned due to low correlations on the validity diagonals.

The combination of MAAS-GP/GERS and MAAS-GP/MAAS-self distinguishes the different types of medical interviewing skills most adequately as is shown in table 8.7. The other combinations of methods differentiate the dimensions less well. MAAS-GP thus appears to be the best able to discern different types of medical interviewing skills. GERS and MAAS-self are second and third, whereas GSRS is almost unable to discern dimensions with the exception of "interpersonal skills".

In conclusion, four dimensions of medical interviewing skills that were discerned theoretically and used to construct the MAAS-GP-scales can be distinguished empirically. "History-taking", "presenting solutions", "structuring the interview" and "interpersonal skills" are distinct types of medical interviewing skills. Difficulties arise in distinguishing the dimensions referring to the "exploration of reasons for encounter" and "communication skills".

8.5.3 Criterion 3.

Criterion 3 states that values on the validity diagonal must be higher than the off-diagonal values in the monomethod triangle. The third criterion was formulated to secure optimal measurement of the dimensions because every psychological measurement device is characterized by features that are specific to the dimension of interest and other features which are characteristic for the method

being employed. Since the process of measuring always elicits irrelevant method variance, measurements are considered to be inval$ated to the extent that method variance contributes to the scores obtained.

A look at table 8.8 reveals that features of the measurement process impinge strongly upon the scores obtained with GERS and GSRS, whereas a smaller influence of the method is observed for MAAS-self and MAAS-GP. Since the interview behavior on which the data are based was similar for all methods, the differences can be attributed to the methods that were employed.

The third criterion undeniably discloses the difficulties that arise in psychological measurement. Of all methods, MAAS-GP demonstrates the best measurement properties because it evokes a low degree of method variance and shows considerable correlations on the validity-diagonals. Once again, "exploring reasons for encounter" and "communication skills" are measured improperly and are therefore primarily responsible for the failure of MAAS-GP on the third criterion. As we performed a generalizability study, discussed in chapter 7, we know that "exploring reasons for encounter" in particular is measured fairly reliably with low levels of method variance. Furthermore, one of Campbell and Fiske's requirements is that each of the methods employed should measure the dimensions as conceptualized appropriately. It is therefore our opinion that the lack of success of the "exploration of reasons for encounter" on the third criterion can be partly attributed to the failure of the other methods to measure this dimension properly.

MAAS-self displays more method-variance when compared to MAAS-GP, and less when compared to GERS and GSRS. A look at table 8.4 shows that the mean of each MAAS-self scale is considerably higher than the mean of identical MAAS-GP scales. This interesting finding demonstrates that interviewing physicians believe that they perform more facets of interview behavior than they actually do. We have often had the experience in examination situations of noting that medical students mixed up information given by the patient with their own questioning behavior: when physicians received information, they often thought they had asked for it. This induces unreliability in MAAS-self measures. In conclusion, we remark that influences of the self-description mode

applied in MAAS-self are likely to interfere negatively with the measurement of the dimensions.

It is evident that GERS is affected strongly by the method of measurement, which consists of observation of behavior and, subsequently, a rating by experts. Although a considerable influence of the method was expected to occur, we were surprised by the strength of the halo-effect. Halo is conceptualized by an observer's failure to discriminate among conceptually distinct and potentially independent aspects of a subject's behavior, and it is operationalized by high intercorrelations between different dimensions (Saal et al, 1980). We asked general practitioners with experience in both general practice and medical education to participate in this study, especially because they were supposed to be able to distinguish the occurrence and quality of different medical interviewing skills. However, even experts experience difficulties in discerning dimensions in medical interviewing skills when no clearly-worded and well-defined items are available. Ratings of distinct types of interviewing skills displayed during a medical consultation are reduced to a judgment about a problem-solving and an interpersonal-skills dimension (Dielman et al, 1980; Streiner, 1985). A second type of method influence, so-called "leniency", a rater's tendency to assign a higher or lower rating to a subject's behavior, also appears to occur because all averaged ratings of GERS are above the midpoint of the scales. Most experts use the positive part of the scale continuum. Restriction of range, finally, seems to take place because most raters do not use the extreme ends of the scales. The negative side in particular is almost never used. We therefore conclude that strong halo-effects, leniency and central tendency, are all likely to impede the measurement properties of the Global Expert-Rating Scale.

The influence of the method, especially of halo-effect, on GSRS is considered to come close to MAAS-self, because the median of the monomethod triangle is near the median for MAAS-self. We expected the influence of halo in GSRS to approach halo in GERS, because both global rating scales have the feature in common that the behavior of interest is not well defined. It seems that physicians who are in the actual interview situation experience more differentiation than experts.

Moreover, leniency is suspected of influencing GSRS strongly, because averaged ratings on GSRS are significantly above the midpoint of the scales, leading to a decrease in the amount of variance. Interviewing physicians rate the quality of their own interview behavior more positively in comparison with observers' ratings. Restriction of range, finally, seems to take place because categories on the negative side of the scales are almost never used. In conclusion, we observe that halo, leniency and restriction of range in particular, are inclined to diminish the measurement properties of the GSRS. Since GSRS utilizes one item which is not well-defined to represent each dimension, this method of measuring medical interiewing skills is considered to be highly unreliable.

How should the third criterion be regarded? Campbell and Fiske constructed this criterion in order to determine the major sources of variation in measurements and in order to conclude that enough trait variance was measured to sustain optimal measurement. These precautions were taken to secure the process of measurement. With regard to our study, the third criterion undeniably discloses the difficulties that arise in psychological measurement. Halo, leniency and restriction of range appear to occur in varying degrees in our measurements.

8.5.4 Criterion 4.

The fourth criterion states that the patterns of trait interrelationship should be the same in all heterotrait triangles in both monomethod and heteromethod blocks in order to provide evidence for divergent validity. Satisfaction of this criterion would suggest that the underlying traits are really correlated, whereas failure of this criterion would imply that the observed correlation between traits assessed by a given method is due to a method or halo bias (Marsh et al, 1983). The interpretation of the fourth criterion has posed problems for us and for several other researchers because Campbell and Fiske did not operationalize it. Some authors have merely mentioned the fourth criterion in a publication but have not applied it to their data (Marsh et al, 1983). Other authors have considered this criterion to be too strict and therefore unrealistic (Magnusson, 1966). In our study it is unrealistic to interpret a correlation matrix consisting of 288

correlations. Each interpretation can be refuted by other correlations which will then suggest a different explanation. Furthermore, (parts of) the correlation matrix cannot be factor analyzed because of the small number of subjects (see also chapter 12). We therefore decided not to apply the fourth criterion to our correlation matrix. As this decision was taken, it can be concluded that no clear pattern of interrelations between the dimensions was observed in our data and that method influences are likely to interfere in the strength of the correlations but this had already been revealed during the interpretation of the third criterion.

8.6 Concluding remarks.

In this chapter, the convergent and divergent validity of the Maastricht History-taking and Advice Checklist in General Practice, in addition to the validity of three other methods of measurement of medical interviewing skills, is studied by means of the multitrait-multimethod matrix. In the multitrait-multimethod matrix, several dimensions in medical interviewing skills are measured with several methods. The resulting correlation matrix is scrutinized by means of four criteria which were developed by Campbell and Fiske.

For the Maastricht History-taking and Advice Checklist in General Practice, the convergent validity of "history-taking", "presenting solutions", "structuring the interview" and "interpersonal skills" is clearly warranted by the strength of the correlations, whereas the "exploration of reasons for encounter" and "communication skills" fail to provide evidence of convergent validity. Essentially, the same conclusions can be drawn for a self-evaluation variant of the MAAS-GP and the Global Expert-Rating Scale, whereas for the Global Self-Rating Scale, insufficient evidence for convergent validity is obtained with the exception of a measurement of "interpersonal skills".

Divergent validity of dimensions in medical interviewing skills is established for "history-taking"," presenting solutions", "structuring the interview" and "interpersonal skills". Difficulties arise in distinguishing dimensions referring to the "exploration of reasons for encounter" and "communication skills". Furthermore, MAAS-GP appears to be the most effective in discerning dimensions, followed by GERS and MAAS-self, whereas GSRS is unable to distinguish dimensions.

Moreover, MAAS-GP displays the best measurement properties because it evokes only low degrees of method-variance when compared to other methods. Halo, leniency and restriction of range are inclined to diminish the measurement properties of GERS and GSRS, and partly of MAAS-self.

All in all, MAAS-GP appears to be the best method of measurement of medical interviewing skills because it displays evidence of convergent and divergent validity, and is minimally influenced by the method of measurement.

Table 8.5: Multitrait-multimethod matrix of the Maastricht History-taking and Advice Checklist in General Practice and other methods of measuring medical interviewing skills.

		MAAS-GP						MAAS-SELF					
		EE	HT	PS	STR	IPS	COM	EE	HT	PS	STR	IPS	COM
MAAS-GP	EE	-											
	HT	.16	-			.22						.20	
	PS	.04	.23	-						.40			
	STR	.22	.54	.10	-								
	IPS	.09	.39	.17	.48	-			.12				
	COM	.19	.26	-.13	.41	.34							
MAAS SELF	EE	.21	.19	.18	.04	.10	-.04	-					
	HT	.01	.63	.16	.22	.58	.08	.25	-			.40	
	PS	-.07	.36	.50	.04	.24	-.01	.48	.60	-			
	STR	.28	.35	.42	.37	.43	.06	.29	.29	.40	-		
	IPS	.13	.46	.08	.24	.43	.12	.21	.66	.41	.38	-	
	COM	.03	.20	.13	.18	.21	.10	.22	.47	.52	.35	.45	-
GERS	EE	.02	.22	.42	.41	.40	.04	.19	.12	.26	.15	.16	.28
	HT	-.01	.47	.46	.47	.36	.17	-.04	.41	.42	.43	.34	.43
	PS	.06	.24	.59	.38	.28	.05	.04	-.02	.25	.45	.03	.36
	STR	-.01	.35	.43	.62	.47	.20	.16	.23	.30	.45	.19	.37
	IPS	.05	.26	.30	.43	.53	.28	.21	.30	.34	.38	.27	.34
	COM	-.01	.36	.45	.49	.42	.29	.06	.16	.27	.53	.14	.29
GSRS	EE	.15	.26	.20	.10	.31	.20	.17	.24	.33	.33	.60	.14
	HT	.05	.01	.02	-.01	.21	.09	.01	.20	.23	.11	.39	.51
	PS	.07	.06	.17	-.12	.24	-.05	.12	.38	.48	.25	.50	.42
	STR	.18	.37	.05	.32	.48	.26	.18	.38	.18	.41	.65	.14
	IPS	-.17	-.05	.22	.07	.45	.07	.32	.37	.47	.14	.34	.27
	COM	.09	-.11	-.07	.03	.17	.05	-.07	.18	.27	.22	.50	.58

Below diagonal: Pearson correlation coefficients between 4 methods and 6 traits of medical interviewing skills.

Above diagonal: Median for validity-diagonals, heterotrait-heteromethod triangles, heterotrait-monomethod triangles.

In Brackets: Correlation between two experts in global expert rating scale.

N = 33 physicians, case: myocardial infarction
$r \geq .29$ than $p \leq .05$
$r \geq .40$ than $p \leq .01$

GLOBAL EXPERT RATING SCALE						GLOBAL SELF RATING SCALE					
EE	HT	PS	STR	IPS	COM	EE	HT	PS	STR	IPS	COM
			.50	.30					.16	.06	
		.36						.20			
			.28	.21					.38	.22	
		.34						.33			
(.22)											
.43	(.26)			.63					.24	.17	
.62	.62	(.30)									
.71	.70	.72	(.42)								
.56	.50	.41	.67	(.39)				.12			
.62	.63	.77	.80	.66	(.15)						
.06	.18	.12	.14	.11	.12	–				.44	
.18	.25	.16	.15	.12	-.01	.43	–				
.04	.35	.21	.15	.13	-.04	.44	.57	–			
-.01	.35	.05	.26	.08	.01	.73	.46	.53	–		
.21	.28	.14	.35	.47	.11	.24	.25	.53	.36	–	
.07	.17	.20	.03	.05	-.08	.41	.65	.71	.40	.26	–

REFERENCES

Campbell DT, Fiske DW. Convergent and discriminant validation by the multi-trait multi-method matrix. Psychological Bulletin, 1959; 56: 81-105.

Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin, 1955; 52: 281-302.

Crijnen AAM, Thiel J van, Kraan HK. Evaluatie van consultvoering: een spreekuur nagebootst (Evaluation of a medical consultation: simulating consultation hours). Huisarts en Wetenschap; 1986: 29: 316-318.

Dielman TW, Hull AL, Davis WK. Psychometric properties of clinical performance ratings. Evaluation and the Health Professions, 1980; 3: 103-117.

Eisenthal S, Lazare A. Expression of patient's request in the initial interview. Psychological Reports, 1977; 40: 131-138.

Eisenthal S, Koopman C, Lazare A. Process analysis of two dimensions of the negotiated approach in relation to satisfaction in the initial interview. Journal of Nervous and Mental Disease, 1983; 171: 49-54.

Fiske DW. Measuring the concepts of personality. Aldine Publishing Company, Chicago, 1971.

Gonella JS. Evaluation of clinical competence (editorial). Journal of Medical Education, 1985; 60: 70-71.

Jöreskog KG, Sörbom D. Lisrel IV: A general computer program for estimation of linear structural equation systems by maximum likelihood methods. University of Uppsala, Uppsala, 1978.

Katz FM. Trends in assessment (Editorial). Medical Education, 1982; 16: 61-62.

Kerlinger FN. Foundations of behavioral research. Holt, Rinehart and Winston, Inc., New York, 1981.

Magnussen D. Test theory. Addison-Wesley, Reading, Massachussetts, 1967.

Marsh HW, Hocevar D. Confirmation factor analysis of multitrait-multimethod matrices. Journal of Educational Measurement, 1983; 20: 231-248.

Saal FE, Downey FG, Lahey NA. Rating the ratings: assessing the psychometric quality of rating data. Psychological Bulletin, 1980; 88: 413-428.

Schmitt N, Stults DM. Methodology review: analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 1986; 10: 1-22.

Streiner DL. Global rating scales. In: Neufeld VR, Norman GR (Eds.). Assessing clinical competence. Springer Publishing Company, New York, 1985.

Thorndike RL. Applied psychometrics. Houghton Mifflin Company, Boston, 1982.

CHAPTER 9 INTERVIEWING SKILLS AND MEDICAL COMPETENCE

A.A.M. Crijnen, G.J. Post, H.F. Kraan, C. van der Vleuten, T. Imbos, and J. Zuidweg.

9.1 Introduction.

In this chapter, the validity of the Maastricht History-taking and Advice Checklist in General Practice Primary Care is examined by correlating its results with other measurements of medical competence. In addition to the measurement of medical interviewing skills, physicians' medical knowledge, interpersonal skills, care and concern for the patient and problem-solving skills are assessed. In general, the validity coefficients support the validity of the Maastricht History-taking and Advice Checklist. Medical interviewing skills are confirmed by measurements of interpersonal skills, by measurements of care and concern and by measurements of problem-solving skills as far as they concern information exchange. The study reveals furthermore, that medical interviewing skills are clearly discerned from medical knowledge and the formulation of a treatment plan.

9.2 Validity studies of measurements of medical interviewing skills.

A plea has been made recently for the study of the validity of measurements of medical competence (Katz, 1982; Gonella, 1985). Some have said that greater attention should be given to the valid assessment of physicians' actual job performance, whereas others have observed that medical competency cannot be conceptualized as a single variable, since knowledge, problem-solving skills, interviewing skills and attitudes should be distinguished from each other.

Validity studies of methods of measurements of medical competency should be designed according to generally accepted, scientific criteria. On reviewing the literature, the following criteria were detected (Cronbach et al, 1955; Thorndike, 1982).

The first criterion requires a clear definition of and theory for the competence under study in order to ascertain the match between the conceptualization of the competence and its empirical measures, and to

distinguish the competence under study from other distinct competences: eg "history-taking" skills should be defined clearly in order to achieve agreement among researchers about the content of a test designed to measure history-taking skills. Moreover, the definition of "history-taking skills" should augment the distinction from other competences, such as "interpersonal skills" or "medical knowledge".
The second criterion requires that validity is empirically underscored by different methods of measurement of the same competence. Evidence from different sources gathered in different ways should all point to the same competence domain: eg. different methods of measurement of a physician's interpersonal skills are expected to correlate higher with each other than with measurements of his other competencies.

The third criterion requires that methods of measurement of the competence under study are differentiated empirically from measurements of other competences (Cronbach et al, 1955; Kerlinger, 1973). Measurements ought to show the difference between theoretically discernable medical competences: eg. the skills to explore the reasons for encounter are expected to correlate low with medical knowledge or the quality of physician's treatment plan.

These three criteria constitute a scientific model of how the validity of measurements of medical competency ought to be examined. Three validity studies of measurements of medical interviewing skills were scrutinized by applying the criteria (Stillman et al, 1977; Brockway, 1978; Swanson et al, 1981).

In the first study by Stillman et al (1977), who examined the validity of the Arizona Clinical Interview Rating Scale (ACIR), the 16 interviewing skills measured by the ACIR-scale are grouped into six major subsections which are treated ultimately as if they all contribute to a similar competence domain. Since the items purport to different skills, the correctness of this procedure can be questioned. No methods intending to measure the same competence domain were taken into account. However, the scores on the ACIR-scale were correlated with Medical College Admissions Tests which are supposed to measure a different competence. This study therefore fulfils only one of the three aforementioned criteria.

The second study under consideration is by Brockway (1978), whose analysis of the validity of her interview rating scale shows some shortcomings because the content of the two subscales, respectively "relationship skills" and "problem-solving skills", is rather heterogeneous and can easily be interchanged. The differences between the subscales are not well defined. No measurements indicating a similar competence domain were taken into account. The study, however, fulfilled one of the three criteria, because the rating scale was compared to two different measurements of medical competence; namely, data collection and problem identification.

The third study under consideration is by Swanson et al (1981), who compared the validity of three measurements of medical interviewing skills, the ACIR-scale, an Interaction Analysis-rating and a History and Physical Exam checklist. The heterogeneity of the measures hindered the researchers in adequately interpreting the matrix of correlations between the methods. They concluded that, in essence, no evidence for construct validity through comparison between measurements had been obtained. By reinterpreting the correlation matrix, it is possible to arrive at a different conclusion. When all measurements are assigned as measurements of interpersonal skills, communicative skills or physical examination skills, each of the three measurements of interpersonal skills appears to correlate significantly with the others: this conclusion supports validity. Moreover, measurements of interpersonal skills can be distinguished from communicative skills and physical examination skills. The same is not true for measurements of communicative skills. The study fulfilled two of three criteria, because measurements with the same and a distinct meaning were applied simultaneously.

All studies reveal difficulties in defining the measurements under examination and only one provides evidence of validity by applying measurements indicating a similar competence domain. Two studies differentiated medical interviewing skills from measurements of medical knowledge, data collection, diagnosis and treatment.

In this chapter, the validity of the Maastricht History-taking and Advice Checklist in General Practice is established by taking into account the three criteria mentioned previously. Five distinct types of

medical interviewing skills distinguished in the MAAS-GP, were defined and treated separately. Furthermore, measurements with a similar theoretical intention to the MAAS-GP-scales and measurements indicating distinct competences were applied simultaneously. This preliminary study was conducted before the reliability and scalability of the MAAS-GP were examined thoroughly by means of the procedures described in chapters 5 and 7. All items are thus taken into account as is the scale "basic interviewing skills" which had not yet been divided into "interpersonal skills" and "communicative skills".

The validity of the MAAS-GP was examined by correlating MAAS-GP-measurements of five types of medical interviewing skills with seven distinct measures of medical competence. These five types of interviewing skills pertain to the exploration of reasons for encounter, history-taking skills, presenting solutions, structuring the interview, and basic interviewing skills. The seven measurements of medical competence are: medical knowledge tests, expert judgments on interpersonal skills, expert judgments of care and concern for the patient, and four measurements of medical problem-solving. The delineation of constituents of medical competence was derived largely from the work of Fabb and Marshall (1983).

Simulated patients, lay-people staging a medical complaint, were used because they provide the opportunity for assessing medical interviewing skills as well as physical examination skills when real patients cannot be used (Stillman, 1983). In this study, the simulated patients were lay people, chosen from a large group of simulated patients who were trained by the Skills Laboratory at Maastricht Medical School to simulate complaints in the undergraduate-medical curriculum. Two simulated patients presented respectively complaints of fatigue and dyspnoe in exertion, and low back pain without irradiation.

The correlations between all measurements are expected to support either the similar meaning or the distinct character of MAAS-GP-measurements of medical interviewing skills and, as a secondary goal, to highlight the interrelations between interviewing skills and other domains of medical competence.

9.3 Methods.

9.3.1 Subjects.

All 45 physicians who graduated in the Summer of 1982 from Maastricht Medical School were asked to participate in this study: 28 decided to take part. The participating physicians did not differ significantly from the total group of graduated physicians in terms of age, sex distribution or scores on a medical knowledge test. The mean age of these 10 women and 18 men was 26. All subjects had gone through the 6 year problem-based medical curriculum, part of which is a continuous teaching program of medical interviewing skills with use of simulated patients under expert supervision. The study was carried out 2-3 months after graduation with the original goal of following up on physicians' competence after medical school (Post et al, 1985). We used this opportunity to validate the MAAS-GP.

9.3.2 Instrument to be validated (the MAAS-GP).

The Maastricht History-taking and Advice Checklist in General Practice is a 68-item observation instrument for the assessment of medical interviewing skills. Expert observers view videotapes of (simulated) medical interviews and rate these interviews on the items. Items are grouped into five scales measuring distinct types of interviewing skills.

The first scale, "exploring reasons for encounter", measures the physician's ability to clarify the patient's complaint, to explore the motives and expectations in the pre-patient phase leading to the visit and to obtain information about the patient's causal attributions. It measures the patient-centered part of the medical interview.

The second scale, "history-taking", measures skills which enable the physician to generate hypotheses about the nature of the patient's complaint, to test these hypotheses and to describe the complaint in medical explanatory terms. It measures the collection of present and past medical data.

The third scale, "presenting solutions", measures the quality of information exchange on diagnosis, aetiology, prognosis, treatment and the negotiation between physician and patient about the treatment plan.

The fourth scale, "structuring the interview", measures the physician's skill in opening, closing and phasing the interview.

The fifth scale, "basic interviewing skills", measures the ability to enhance effective information exchange and to establish rapport with the patient.

The items in the MAAS-GP refer either to content or process of medical interviewing skills during initial consultations. They are described behaviorally and have to be scored by skilled observers. Items and criteria for scoring are defined in an accompanying manual which is available in Dutch and English (Kraan et al, 1986; see this thesis). Items are described in behavioral terms to enhance both the reliability and practical application of the MAAS-GP in educational situations. The items in the first four scales are scored on a two-point scale (behavior is present or absent), whereas items in the fifth scale are scored on a three point rating scale (positive, indifferent, negative). In this study, items were rated by two skilled observers.

9.3.3 Instruments used to validate the MAAS-GP.

Medical knowledge was assessed by means of the Medical Knowledge Progress Test (Verwijnen, et al, 1982). The knowledge test was part of the examination system at Maastricht Medical School and consisted of approximately 250 true-false statements pertaining to medical knowledge. This score was obtained by counting the number of correct-scores. Each knowledge test was administered to all students at Maastricht Medical School four times a year. Since the present study was conducted only 2-3 months after graduation, the total correct-scores of the four tests administered during the physicians' final year in medical school were included in this study as a measure of medical knowledge.

The quality of the physician's interpersonal skills was rated by expert observers by means of a 10-item instrument, pertaining to the physician's attention for the patient and warmth in the communication. The items were scored by general practitioners on Likert-type, 5-point scales after observing the videotapes of physicians interviewing simulated patients. In this study, each interview was observed by three

general practitioners who were randomly chosen for each subject out of a pool of 10 general practitioners who served as expert raters in this study. The score was obtained by averaging the scores of these three general practitioners for each interview.

Care and concern during physical examination was measured by means of a four-item instrument, focusing upon the physician's care in reducing patient's anxiety and his efficiency during the physical examination. The items were expert ratings on Likert-type, 5-point scales by the same experts who rated the interpersonal skills. The care and concern score was obtained by averaging the scores of these three general practitioners for each physical examination.

Medical problem-solving skills were assessed by a paper and pencil test, called Summative Evaluation of Initial Medical Problem-solving (SIMP) (de Graaff et al, in press). SIMP requires physicians to read a short case-vignette and to write down narrative answers to four open-ended questions which reflect the process of medical problem-solving. These questions are: 1. What additional Subjective Information do you wish to obtain by means of interviewing a patient? 2. What additional Objective Information do you wish to obtain by performing a physical examination or additional laboratory tests? 3. What is your Assessment of Diagnosis or problem-definition? 4. What is your Plan for Treatment or further diagnostic examinations? Physicians' responses to each question were compared to criterion-answers obtained from a group of experienced general practitioners who answered the questions for each case-vignette themselves and attained agreement about the correct answers. For example, the Subjective Information-score was the number of matches between the physician's subjective information narrative and the experts' preset criteria.

9.3.4 Procedure.

The 28 physicians all filled in 6 SIMP's within one hour. They were then randomly assigned to one out of two simulated patients, yielding two subgroups of 14 physicians. Physicians were asked to behave as if they had taken charge of a colleague's practice. They were expected to interview the patient, to conduct a physical examination and to present a treatment plan to the patient. The available time was not to exceed

45 minutes. Interview and physical examination were videotaped.

These videotapes were observed independently by a general practitioner and a fourth-year medical student who both rated each interview on all the items of the MAAS-GP. Three general practitioners, who were randomly chosen for each subject, independently viewed the videotapes and rated them on the interpersonal skill-variable and the care- and concern-variable.

Finally, the physicians' scores on four medical knowledge progress tests administered during their final year of medical school were included in this study.

9.4 Results.

In table 9.1, the reliability of the instruments used in this study are shown.

Since subjects were assigned to one of the two simulated patients, two subgroups were formed. Mean scores for none of the criterion measures differed significantly between the groups. However, scores on three scales of the MAAS-GP, namely, exploring reasons for encounter, presenting solutions and structuring the interview, showed significant differences between the two subgroups ($t = -2.66$, $p \leq .05$; $t = -3.67$, $p \leq .01$; $t = -1.75$, $p \leq .05$; $DF = 26$; two-tail) which reflect an influence of the cases on interviewing skills.

Because of the significant influence of the cases on the MAAS-GP-measurements, variance due to cases was partialed out from the correlations between the five MAAS-GP-measurements and the seven other measurements of medical competence. Since two physicians exceeded the available time to answer the 6 SIMP's, their scores are omitted from the analyses. The number of complete cases on which the correlation matrix is based is thus limited to 26. Moreover, correlation coefficients have been corrected for attenuation in the criterion variables by taking the internal consistency as indication for reliability. Thus, validity is expressed as if the coefficients were based on completely reliable criterion measures.

Table 9.1: Internal consistency and inter-rater reliability of measurements of medical competence.

measurements of medical competence	number of items	internal consistency (Cronbach's alpha)	inter-rater reliability (Pearson correlations between observers/raters*)
MAAS			
1. exploring reasons for encounter	8	.41	.56
2. history-taking	23	.48	.82
3. presenting solution	12	.48	.47
4. structuring the	8	.43	.32
5. basic interviewing skills	17	.55	.44
Measurements of medical competence.			
1. medical knowledge progress test	194 278 252 262	.90 .91 .92 .92	
2. interpersonal skills	10	.91	.33 - .63
3. care and concern	4	.89	.80 - .88
Summative evaluation of initial medical problem-solving.			
4. subjective information	6	.68	.79 - .90
5. objective information	6	.38	.80 - .90
6. assessment of diagnosis	6	.52	.33 - .89
7. treatment plan	6	.58	.68 - .95

*) In case of more than 2 observers or raters, ranges are given.

Table 9.2: Validity coefficients between the Maastricht History-taking and Advice Checklist in General Practice and measurements of medical competence after partialing out case-influences and correcting for attenuation.

Maastricht History-taking and Advice Checklist in General Practice	medical knowledge	inter-personal skills	care and concern	medical problem-solving (SIMP) subjective information	objective information	assessment diagnosis	treatment plan
1. exploring reasons for encounter	-.13	.48**	.42*	.66**	.42*	.47*	.09
2. history-taking	-.15	.34*	.28	.60**	.66**	.43*	.00
3. presenting solutions	-.03	.39*	.36*	.37*	.06	.31	.25
4. structuring the interview	.06	.42*	.53**	.38*	.31	.06	-.01
5. basic interviewing skills	-.24	.58**	.39*	.49**	.44*	.43*	.14

(* $p \leq .05$; ** $p \leq .01$; DF= 23; one-tail)

Validity coefficients between the five types of medical interviewing skills and the seven other measurements of medical competency are shown in table 9.2.

The scale "exploring reasons for encounter" is strongly correlated with the measurement of interpersonal skills, subjective information and assessment of diagnosis. The scale is moderately correlated with care and concern and objective information.

The scale "history-taking" correlates strongly with subjective information and objective information, and is moderately correlated with interpersonal skills and assessment of diagnosis.

The scale "presenting solutions" correlates moderately with interpersonal skills, care and concern, and subjective information. The scale "structuring" correlates strongly with care and concern and moderately with interpersonal skills and subjective information.

The scale "basic interviewing skills" is strongly correlated with measurements of interpersonal skills and subjective information. It correlates moderately with care and concern, objective information and assessment of diagnosis.

9.5 Discussion.

In general, the validity coefficients between MAAS-GP-measurements of medical interviewing skills and seven other measurements of medical competence support the validity of the MAAS-GP, although some unexpected dissonants are found.

The scale "exploring reasons for encounter" converges with ratings of interpersonal skills and subjective information as expected. The low combination with medical knowledge and treatment plan, which have also been found in other studies (Stillman et al, 1977; Brockway, 1978), underscore the distinct character of these competence domains. The unexpectedly high validity coefficient with assessment of diagnosis may be explained by the necessity to ask patient-centered questions in order to include in the diagnosis issues defined by the patient as a problem. This is especially true for cases with a combination of somatic and psychological problems in which patient's concerns and real-life circumstances have to be included in the assessment of the diagnosis (Mishler, 1982). The moderate correlation with the care- and

concern-variable supports the validity of the scale "exploring reasons for encounter" because an attitude of caring and reassurance during the physical examination reflects a human dimension in the physician's approach to the patient, also reflected in the skills to explore the reasons for encounter.

The scale "history-taking", which measures the collection of present and past medical data, is highly correlated with measurements of subjective and objective information during medical problem-solving. Medical problem-solving is seen as a collaboration of several competences of which the search for additional data by means of history-taking is considered to be the communicative aspect (Neufeld et al, 1981). Since history-taking skills are determined by the process of medical problem-solving, the strong correlations underscore the validity of the MAAS-GP-scales. The low correlations of history-taking with medical knowledge and treatment plan support the distinct character of these measures of medical competence. The moderate correlation between history-taking and assessment of diagnosis suggests that physicians who collect more data have a better chance of establishing an accurate diagnosis. Although this finding is supported by Brockway (1978), other studies reveal that the number of questions asked during history-taking is not unequivocally related to the quality of diagnosis (Kassirer et al, 1978). Specialists collect less information and mention the correct diagnosis earlier in comparison with non-specialists who often revert to a general review of organ systems. The scale "presenting solutions" does not correlate strongly with any of the other competences. The validity coefficients with measurements of interpersonal skills, care and concern as well as with subjective information are moderate. The strength of these correlations indicates the continuous nature of information-exchange and the human factor of patient-centeredness and reassurance, which are also constituents of the scale "presenting solutions". The low correlation between this scale and treatment plan was expected because the MAAS-GP-scale deals only with the process of information-exchange and negotiation and not with the content of the exchanged information, the treatment plan itself. This argument also holds for the low validity coefficients with medical knowledge, objective information and

assessment of diagnosis, which underscores once more the distinct character of the scale "presenting solutions".

The scale "structuring the interview" is strongly correlated with the care- and concern-variable, which measures the physician's care in reducing the patient's anxiety during the physical examination. Items in this MAAS-GP-scale measure the quality by which the physician structures the interview into natural segments enabling the patient to voice his concerns and to understand the goal of certain interview behavior. Although not yet studied, it is conceivable that interviewing skills which structure the interview will reduce the patient's anxiety. In analogy with the reassuring effects of introduction and explanation of the procedures during the physical examination, similar behavior by the physician during the interview may entail similar effects. The low correlation with medical knowledge, objective information and treatment plan indicate a divergency with knowledge and problem-solving skills as expected.

The scale "basic interviewing skills" accompanies measurements of interpersonal skills and of subjective information during medical problem-solving. This agrees with the underlying aim of the scale, which is to measure a physician's ability to establish an optimal rapport with the patient and to induce an effective exchange of information. This pertains also to the moderate correlation with care and concern. The moderate correlations with objective information and assessment of diagnosis might be explained by the statement of DiMatteo and DiNicola (1982), that a physician's competence is likely to involve a scientific and technical ability translated into practice through both interpersonal skills and the art of medicine. The low validity coefficients with medical knowledge and treatment plan underscore the distinct character of the pertinent competences.

Methodologically, the present study posed several problems. Firstly, the reliability indices of the MAAS-GP were, unexpectedly, low to moderate: in prior studies, higher inter-observer reliabity has been obtained. Further exploration by means of generalizability studies revealed that the coefficients differed markedly for both patients (respectively, .62 and .37). This difference in reliability was attributed at least partly to the controlling communication style of

one of the simulated patients. Since MAAS-GP-items are mainly directed at the physician's interviewing skills, the observers easily make mistakes when patients take the initiative so intrusively. Item definitions and criteria for scoring cannot deal appropriately with this situation.

Secondly, the strong impact of the cases on physicians' interviewing skills influenced the validity study to some extent. The influences of the different cases may be attributed to two factors: the presentation of the case by the simulated patients and the characteristics of the medical problem. Prior studies with the MAAS-GP provided evidence that the case as medical problem did not influence the physician's interview behavior significantly (Kraan et al, 1986), whereas others have pointed to a considerable impact of the characteristics of the medical problem on the physician's interview behavior (Norman et al, 1981). Interpretation of the study presented here suggests that case influences seem to result from the case presentation by the simulated patients. This conclusion is corroborated by the finding that history-taking skills were influenced least by difference in case whereas, at first sight, the greatest influence of the case as medical problem was expected on this type of interviewing skills.

9.6 Conclusion.

It can thus be concluded that the MAAS-GP validly measures 5 distinct types of medical interviewing skills. Moreover, the validity coefficients between MAAS-GP and related medical competences, such as interpersonal skills, care and concern, and information-exchange, confirm the validity of the MAAS-GP. The validity coefficients between MAAS-GP and unrelated medical competences, like knowledge and the quality of a treatment plan, display the distinct character of MAAS-GP-measurements of medical interviewing skills and therefore support the validity of the MAAS-GP.

The study supports the model of distinct medical competences as delineated by Fabb and Marshall (1983) and underscores that the evaluation of students' or physicians' medical competency can no longer be based solely on the assessment of their medical knowledge as is the case in most examinations. Medical interviewing skills should be taken into account.

REFERENCES

Brockway BS. Evaluating physician competency: what difference does it make? Evaluation and Program Planning, 1978; 1: 211-220.

Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin, 1955; 52: 281-302.

DiMatteo MR, DiNicola DD. Achieving patient compliance: the psychology of the medical practitioner's role. Pergamon Press, New York, 1982.

Fabb WE, Marshall JR. The assessment of clinical competence. Lancaster, England, MTP Press Limited, 1983.

Gonella JS. Evaluation of clinical competence (editorial). Journal of Medical Education, 1985; 60: 70-71.

Graaff E de, Post GJ, Drop MJ. Validation of a new measurment of clinical problem-solving. Medical Education, (accepted for publication).

Kassirer JP, Gorry GA. Clinical problem solving: a behavioral analysis. Annals of Internal Medicine, 1978; 89: 245-255.

Katz FM. Trends in assessment (editorial). Medical Education, 1982; 16: 61-62.

Kerlinger FN. Foundations of behavioral research. Holt, Rinehart and Winston Inc., New York, 1973.

Kraan HF, Crijnen AAM, DeVries MW, Zuidweg J, Imbos T, Vleuten C van der. Are medical interviewing skills teachable? Perspectief, 1986; 4: 29-51.

Kraan HF, Crijnen AAM, Zuidweg J. The Maastricht History-taking and Advising Checklist: an observation instrument for the measurement of physicians' interviewing skills in initial medical consultations in primary care, manual for scoring. Department of Social Psychiatry, University of Limburg, Maastricht, 1986.

Mishler EG. The discourse of medicine: dialectics of medical interviews. Ablex Publishing Corporation, Norwood, New Jersey, 1982.

Neufeld VR, Norman GR, Feightner JW, Barrows HS. Clinical problem-solving by medical students: a cross-sectional and longitudinal analysis. Medical Education, 1981; 15: 26-32.

Norman GR, Feightner JW. A comparison of behavior on simulated patients and patient management problems. Medical Education, 1981; 15: 26-32.

Post GJ, Hellemons-Boode BSP, Heyden PFA van der, Graaff E de, Drop MJ. Medische competentie: een vergelijking tussen verschillende meetinstrumenten (Medical competence: comparing different measurement instruments). Rijksuniversiteit Limburg, Maastricht, 1985.

Stillman PL, Brown DR, Redfield DL, Sabers DL. Construct validation of the Arizona Clinical Interview Rating Scale. Educational and Psychological Measurement, 1977; 37: 1031-1038.

Stillman PL, Burpeau-Di Gregorio MY, Nicholson GI, Sabers DL, Stillman AE. Six years of experience using patient instructors to teach interviewing skills. Journal of Medical Education, 1983; 58: 941-946.

Swanson DB, Mayewski RJ, Norsen L, Baran G, Mushlin AI. A psychometric study of measurement of medical interviewing skills. In: Proceedings of the 20th Annual Conference of Research in Medical Education, Washington DC, 1981.

Thorndike RL. Applied psychometrics. Houghton Mifflin Company, Boston, 1982.

Verwijnen GM, Imbos T, Snellen A, Stalenhoef B, Pollemans M, Luyk S van, Sprooten SM, Leeuwen Y van, Vleuten C van der. The evaluation system at the Medical School of Maastricht. Assesssment and Evaluation in Higher Education, 1982; 3: 225-244.

CHAPTER 10 SCALABILITY AND RELIABILITY OF THE MAAS-PRIMARY MENTAL HEALTH CARE

H.F. Kraan and A.A.M. Crijnen

10.1 Introduction.

In the present chapter, results of an investigation into the scalability and reliability of the MAAS-PMHC are reported.

Scalability is attained when items aiming to measure one underlying theoretical dimension fit well the assumptions of the Rasch-model. To ascertain its scalability, the fit of 8 MAAS-PMHC-scales to the Rasch-model is determined. In addition, it is checked to see whether each scale measures a different dimension. This study is useful because, in forthcoming studies, such as validity research, we shall use the 8 scales as indices for important theoretical concepts such as "exploration of the reasons for encounter", "psychiatric examination", etc. The scalability study is presented in 10.2. Furthermore, the reliability of the MAAS-PMHC is investigated.

The question is addressed of whether measurement with the MAAS-PMHC is stable and consistent when conditions of subjects, observers and mental health problems are varying. Reliability is studied on the level of the 8 scales of the MAAS-PMHC for the summed scores of the 8 scales of the MAAS-PMHC (10.3). Reliability is also studied for each MAAS-PMHC item. This subject is treated in the next chapter where results are presented from studying item reliability in a content validity perspective. Generalizability analyses are used as a method to gain insight in the amount of true variance i.e. of the physician's essential ability in interviewing, about the agreement among observers, about measurement biases caused by the instrument itself and about errors of measurement.

A special question of reliability is that of inter-case reliability: this refers to the stability of interviewing skills over different mental health problems. The impact of the "case influence" on the MAAS-scores is also studied by means of generalizability analysis (10.4).

We refer the reader to chapter 5 for clarifying remarks on the objectives and the theoretical background of the methodology of these

studies.

The chapter ends with summaries and conclusions (10.5).

10.2 Scalability of the MAAS-PMHC.

A first step in scale construction of a measurement instrument is the assembly of a set of items which all measure the latent trait we wish to measure. The latent trait in our study is the physician's ability to perform initial interviews in primary mental health care. We are interested in whether the 8 theoretically assembled scales have one underlying trait.

We also investigate whether the items of these scales are able to differentiate between competency levels on this latent trait.

We investigate whether these scales measure different underlying traits or whether there is also uni-dimensionality between the different scales.

In this thesis, we have chosen to use the probabalistic scale models, especially the one-parameter logistic (Rasch) model, to support the process of scale construction because of its attractive, though demanding, features. We have extensively described the characteristics of this model in chapter 5.

In the Rasch-model, the items should fit in a logistic function, relating the probability of a succesful score on an item to the subject's position on the ability scale. The shape of this curve is almost indistinguishable from the normal ogive (Hambleton et al, 1977).

The fitting of items in this item-characteristic curve has two important consequences. In the first place, the probability of an individual subject providing a "correct score" on an item (when a certain interviewing skill is present) is independent of the distribution of the subject's ability in the population of subjects of interest. The probability of a correct score for a physician does not depend on how many other physicians are located at the same point on the ability continuum (or at a different point). In the second place, all items in the Rasch-model are assumed to have equal discriminating power, but only varying in terms of "difficulty". In our study, "difficulty" means the physician's level of ability to apply a certain interviewing skill.

10.2.1 Method.

Subject and research setting.

Interviews of 102 (future) physicians who talked with a patient simulating a major depression have been scored with the MAAS-PMHC. This sample consists of two subsamples of subjects: 40 residents in general practice and 62 psychiatric clerks during the 6th and final year of their medical curriculum. The characteristics of the residents group and the experimental conditions of this study have been described in chapter 8.

The simulated patients presented a case of a middle-aged woman with a long standing undertreated depression (Diagnostic and Statistical Manual of Mental Disorders - III: major depression), after migration from her native village to a neighbouring town.

The forty resident interviews were scored "live" with the MAAS, whereas the 62 clerk interviews were scored after being videotaped.

10.2.2 Analysis

Over the 102 interviews, Rasch analysis was carried out of each of the 8 MAAS scales using the PML-program (Gustaffsson, 1977; see also 7.3.1). The following five steps were taken.

Firstly, the item scores of the scales "interpersonal and communicative skills" were dichotomised. On theoretical grounds, three constructors of the MAAS dichotomized each three-point scale by determining whether the second scale point was to be considered as good or bad interviewing. Their agreement was high (80-90%). With respect to controversial items, a consensus was attained after discussion.

Secondly, descriptive statistic analyses were carried out, revealing no missing data.

Thirdly, the binominal and Allerups graphical test selected the items not fitting in the Rasch-model (Molenaar, 1981).

Fourthly, the fit of the scales in the Rasch-model was ascertained by means of the Martin Löf chi-square test (Martin Löf, 1973; see also 7.3.1).

Finally, the uni-dimensionality of each MAAS-scale was tested by determining whether pairs of Rasch homogeneous scales which were considered to measure distinct traits could be positioned in one, similar scale.

Table 10.1: Rasch analysis of the eight MAAS-PMHC scales.

SCALES	NUMBER OF ITEMS		MARTIN LÖF TEST		
	ORIGINAL* SCALE	RASCH* SCALE	CHI-SQUARE	PROBABILITY	INTERNAL CONSISTENCY (KR-20)**
I. Exploration of the reason for encounter	13	13	111.30	.70	.57 (.67)
II. History-taking	13	13	80.02	.88	.46 (.57)
III. Psychiatric examination	18	9	37.34	.24	.46 (.65)
IV. Socio-emotional exploration	20	18	166.13	.22	.54 (.57)
V. Presenting solutions	15	13	75.31	.94	.61 (.68)
VI. Structuring the interview	8	5	11.91	.45	.66 (.83)
VII. Interpersonal skills	10	10	80.11	.07	.34 (.51)
VIII. Communicative skills	7	7	20.32	.91	.21 (.43)

*) See appendix B for the items belonging to the original and Rasch homogeneous scales.

**) (in parentheses) The internal consistency figures corrected for a test length of 20 items are given.

10.2.3 Results

The Rasch analysis of the 8 MAAS-PMHC scales is given in table 10.1.

Firstly, it is shown that the 8 scales fit the assumptions of the Rasch-model after elimination of only a few items by the binominal and Allerups graphical tests (first and second column of table 10.1). From the scales "psychiatric examination", "socio-emotional exploration" and "presenting solutions" and "structuring the interview", several items are excluded in order to fit the scales in the Rasch-model.

In the scale "psychiatric examination", some items are excluded because in the cases the simulated patients presented, certain interviewing skills of this scale were not applicable.

Secondly, results of the Martin Löf tests (column three to five of table 10.1) show the empirical proof that the 8 scales fit in the Rasch-model. The higher the probability of the chi-square test, the better the fit of the scale to the assumptions of the Rasch-model.

Thirdly, the internal consistency (KR-20) of the 8 scales are presented as a comparison of the probabalistic scale model with the classical test model. The alphas of the 8 scales are given in column 5 of table 10.1. To compare them, their values have been corrected for a test length of 20 items by means of the Spearman-Brown formula (Guilford et al., 1982). The alphas are moderate except for the scale "interpersonal and communicative skills", where they are low.

Fourthly, the item difficulties that reflect the point of the ability scale, where subject have a 50% chance to score an item positively, have been calculated. They are presented in column 5 of table 10.2.

In table 10.3 the subject's scores on the scales are pair wise compared by means of the Martin Löf chi-square test to examine whether the scales measure the same underlying dimension. The higher the probability (p), the higher is the probability of the uni-dimensionality between both scales. Other statistics provided by the Martin Löf tests are also given such as chi-squares and Pearson correlations (Molenaar, 1981).

Table 10.2 Generalizability coefficients of the MAAS-PMHC-items for 1, 2 and 6 observers; variance components of item scores, to be attributed to observers; item difficulties on the Rasch scales; p-values of items.

I. EXPLORATION OF THE REASON FOR ENCOUNTER

ITEMS	GENERALIZABILITY COEFFICIENTS			% OF TOTAL VARIANCE ATTRIBUTED TO OBSERVERS	ITEM DIFFICULTY	P-VALUE
	1 OBSERVER	2 OBSERVERS	6 OBSERVERS			
1. reason for visit	0.09	0.17	0.37	26	-0.84	0.83
2. description complaint	0.13	0.23	0.47	23	0.25	0.35
3. emotional impact	0.07	0.14	0.32	25	0.61	0.31
4. problem presentation now	0.49	0.66	0.85	1	0.16	0.38
5. opinion about the cause	0.36	0.53	0.77	1	-1.52	0.78
6. discussion in family	0.57	0.73	0.90	5	-1.92	0.61
7. patient's own solutions	0.22	0.36	0.62	15	-0.26	0.51
8. consequences for daily life	0.20	0.34	0.61	16	0.55	0.28
9. life circumstances	0.60	0.75	0.90	0	0.30	0.30
10. habitual solutions	0.14	0.24	0.48	2	0.83	0.15
11. burden to others	0.15	0.26	0.52	0	2.25	0.08
12. recent life-events	0.40	0.57	0.80	17	0.89	0.36
13. desired help	0.64	0.78	0.91	0	-1.31	0.56

Table 10.2: (continued)

II. HISTORY-TAKING

ITEMS	GENERALIZABILITY COEFFICIENTS			% OF TOTAL VARIANCE ATTRIBUTED TO OBSERVERS	ITEM DIFFICULTY	P-VALUE
	1 OBSERVER	2 OBSERVERS	6 OBSERVERS			
14. intensity of complaint	0.13	0.22	0.46	2	0.29	0.23
15. course during the day	0.29	0.45	0.71	4	-1.03	0.36
16. history complaint	0.30	0.46	0.72	13	-2.55	0.99
17. provoking factors	0.24	0.39	0.65	10	-2.32	0.59
18. increasing factors	0.13	0.23	0.48	4	0.04	0.28
19. maintaining factors	0.00	0.00	0.00	6	0.84	0.06
20. decreasing factors	0.20	0.33	0.60	4	-0.39	0.16
21. functionality/gains	0.06	0.11	0.27	18	1.04	0.09
22. mental problems in past	0.24	0.38	0.65	0	1.40	0.09
23. prof. treatment in past	0.20	0.33	0.59	0	-0.67	0.24
24. present consultations	0.00	0.00	0.00	-	2.46	0.02
25. (ab)use medication	0.74	0.85	0.95	3	-1.03	0.32
26. (pseudo) hereditarity	0.70	0.82	0.93	0	1.92	0.04

Table 10.2: (continued)

III. PSYCHIATRIC EXAMINATION

ITEMS	GENERALIZABILITY COEFFICIENTS			% OF TOTAL VARIANCE ATTRIBUTED TO OBSERVERS	ITEM DIFFICULTY	P-VALUE
	1 OBSERVER	2 OBSERVERS	6 OBSERVERS			
27. disturbances in mood/affect	0.33	0.50	0.75	6	-2.41	0.21
28. depression	0.27	0.42	0.69	4	-1.69	0.16
29. depresssive cognitions	0.05	0.10	0.25	1	-1.57	0.14
30. suicidal behavior	0.69	0.82	0.93	1	-2.2	0.18
31. anxiety symptoms	0.24	0.39	0.65	2	1.50	0.16
32. phobic symptoms	0.29	0.45	0.71	22	-	0.13
33. anxiety in-/decreasing factors	0.22	0.36	0.63	8	-	0.05
34. consequences of anxiety	0.00	0.00	0.00	-	-	0.03
35. drowsiness, impaired concentration	0.00	0.00	0.00	-	0.46	0.04
36. disturbed orientation	0.00	0.00	0.00	-	-	0.00
37. disturb. immed. memory	0.00	0.00	0.00	-	2.23	0.01
38. disturb. memory	0.00	0.00	0.00	-	1.50	0.01
39. disturb. remote memory	0.00	0.00	0.00	-	-	0.00
40. distinguishes hallucinations	0.00	0.00	0.00	-	-	0.01
41. character hallucinations	0.00	0.00	0.00	-	-	0.00
42. disturb. stream thought	0.00	0.00	0.00	-	2.23	0.01
43. disturb. content thought	0.00	0.00	0.00	-	-	0.00
44. disturb. own process thought	0.00	0.00	0.00	-	-	0.00

Table 10.2: (continued)

IV. SOCIO-EMOTIONAL EXPLORATION

ITEMS	GENERALIZABILITY COEFFICIENTS			% OF TOTAL VARIANCE ATTRIBUTED TO OBSERVERS	ITEM DIFFICULTY	P-VALUE
	1 OBSERVER	2 OBSERVERS	6 OBSERVERS			
45. feelings love/affect.	0.27	0.44	0.69	7	1.99	0.04
46. aggressive feelings	0.01	0.02	0.07	0	1.45	0.03
47. perspective/aspirations	0.27	0.42	0.69	17	-	0.26
48. care-giving	0.01	0.01	0.04	19	0.47	0.04
49. responsibility	0.00	0.00	0.00	12	-	0.01
50. religious feelings	0.00	0.00	0.00	-	3.12	0.01
51. character/self-image	0.32	0.49	0.74	6	0.29	0.04
52. relations family	0.22	0.36	0.62	8	-2.56	0.55
53. social support	0.25	0.40	0.67	12	-2.08	0.31
54. cultural differences	0.38	0.55	0.79	5	-1.41	0.15
55. prof. functioning	0.53	0.69	0.87	6	-1.82	0.62
56. leisure time	0.45	0.62	0.83	7	-1.64	0.41
57. sexual functioning	0.00	0.00	0.00	-	1.45	0.01
58. sleeping habits	0.49	0.66	0.85	7	-1.64	0.51
59. eating habits	0.88	0.93	0.98	0	-1.14	0.29
60. substance (ab)use	0.00	0.00	0.00	-	2.41	0.03
61. housing condition	0.39	0.56	0.79	13	-2.22	0.29
62. financial situation	0.65	0.79	0.92	1	1.69	0.05
63. education/profession	0.00	0.00	0.00	-	0.20	0.06
64. development	0.00	0.00	0.00	-	1.45	0.01

Table 10.2: (continued)

V. PRESENTING SOLUTIONS

ITEMS	GENERALIZABILITY COEFFICIENTS			% OF TOTAL VARIANCE ATTRIBUTED TO OBSERVERS	ITEM DIFFICULTY	P-VALUE
	1 OBSERVER	2 OBSERVERS	6 OBSERVERS			
65. conveys diagnosis	0.16	0.28	0.53	21	0.90	0.31
66. info causal factors	0.24	0.39	0.66	16	-0.04	0.39
67. info prognosis	0.30	0.46	0.72	12	1.81	0.05
68. patient's expectations	0.30	0.47	0.72	6	-1.37	0.41
69. responsib. treatment	0.10	0.19	0.41	46	-0.35	0.16
70. proposal help	0.01	0.02	0.05	60	-3.95	0.81
71. explains proposal	0.10	0.17	0.39	29	-0.89	0.51
72. pros/cons proposal	0.10	0.19	0.41	4	2.13	0.13
73. opinion proposal	0.12	0.22	0.45	14	-1.81	0.63
74. influence by others	0.25	0.40	0.67	12	0.83	0.13
75. discus. any diff. opinion	0.11	0.20	0.43	18	0.76	0.14
76. choice proposal	0.02	0.03	0.09	5	0.55	0.13
77. concrete info. advice	0.04	0.09	0.22	17	-	0.41
78. advice understood?	0.02	0.03	0.10	62	1.44	0.16
79. appointments follow-up	0.36	0.53	0.77	3	-	0.79

Table 10.2: (continued)

VI. STRUCTURING THE INTERVIEW

ITEMS	GENERALIZABILITY COEFFICIENTS			% OF TOTAL VARIANCE ATTRIBUTED TO OBSERVERS	ITEM DIFFICULTY	P-VALUE
	1 OBSERVER	2 OBSERVERS	6 OBSERVERS			
80. introduces himself	0.58	0.73	0.89	0	-	0.38
81. plan consultation	0.32	0.49	0.74	5	-	0.07
82. summarizes reason visit	0.52	0.68	0.87	3	0.90	0.36
83. ordering HIT, PE, SE	0.16	0.28	0.54	11	1.42	0.19
84. EE precedes HIT, PE, SE	0.11	0.20	0.43	1	-0.91	0.49
85. PS after EE, HIT, PE, SE	0.06	0.13	0.54	29	-1.79	0.52
86. starts PS with diagn.	0.14	0.25	0.49	22	0.38	0.33
87. main problems discussed?	0.01	0.03	0.08	27	-	0.13

Table 10.2: (continued)

VII. INTERPERSONAL SKILLS

ITEMS	GENERALIZABILITY COEFFICIENTS			% OF TOTAL VARIANCE ATTRIBUTED TO OBSERVERS	ITEM DIFFICULTY	P-VALUE
	1 OBSERVER	2 OBSERVERS	6 OBSERVERS			
88. facilitation	0.15	0.26	0.52	18	-2.42	0.86
91. reflects emotions	0.23	0.37	0.64	8	-0.70	0.59
92. react express. emotions	0.11	0.19	0.41	4	3.99	0.06
93. feelings of the moment	0.15	0.26	0.51	20	0.90	0.12
97. meta-communication	0.13	0.22	0.46	13	2.01	0.11
99. proper history-taking	0.03	0.06	0.15	38	1.43	0.26
100. puts at ease	0.24	0.38	0.65	19	0.50	0.38
101. proper pace	0.05	0.10	0.26	6	-1.48	0.79
103. congruent non-verbals	0.23	0.38	0.65	21	-2.68	0.92
104. proper eye-contact	0.01	0.02	0.07	92	-1.55	0.76

Table 10.2: (continued)

VIII. COMMUNICATIVE SKILLS

ITEMS	GENERALIZABILITY COEFFICIENTS			% OF TOTAL VARIANCE ATTRIBUTED TO OBSERVERS	ITEM DIFFICULTY	P-VALUE
	1 OBSERVER	2 OBSERVERS	6 OBSERVERS			
89. proper closed questions	0.11	0.20	0.42	33	-0.39	0.56
90. concretizes	0.10	0.18	0.40	16	-0.45	0.58
94. summarizes	0.38	0.55	0.79	13	-1.70	0.73
95. info. small units	0.02	0.03	0.10	26	0.93	0.58
96. checks understanding	0.03	0.05	0.14	54	0.42	0.34
98. proper confrontations	0.06	0.12	0.28	6	0.27	0.24
102. understandable language	0.03	0.05	0.18	2	1.60	0.91

Table 10.3: Results of the Martin Löf chi-square tests for unidimensionality of the MAAS-PMHC scales.

	exploration reason encounter	history-taking	psychiatric examination	socio-emotional exploration	presenting solutions	structuring the interview	interpersonal skills
cy-	X^2=77.4 p=1.00 r=0.19						
iatric ıation	X^2=49.3 p=1.00 r=0.31	X^2=30.9 p=1.00 r=0.38					
- ɔnal :ation	X^2=123.7 p=0.99 r=0.44	X^2=80.3 p=1.00 r=0.26	X^2=40.8 p=1.00 r=0.43				
ıting ions	X^2=126.2 p=0.99 r=0.22	X^2=96.8 p=0.99 r=0.27	X^2=64.8 p=0.99 r=0.18	X^2=100.5 p=0.99 r=0.27			
ɔuring nterview	X^2=50.9 p=0.99 r=0.32	X^2=58.0 p=0.99 r=0.33	X^2=39.6 p=0.99 r=0.07	X^2=76.6 p=0.99 r=0.10	X^2=47.8 p=0.99 r=0.39		
ɔersonal s	X^2=96.4 p=0.99 r=0.01	X^2=75.6 p=0.99 r=0.03	X^2=54.0 p=0.99 r=0.20	X^2=78.8 p=1.00 r=0.10	X^2=81.0 p=0.99 r=0.07	X^2=46.6 p=0.98 r=0.04	
nicative s	X^2=74.7 p=0.88 r=0.05	X^2=58.4 p=0.99 r=0.13	X^2=57.2 p=0.65 r=0.06	X^2=84.2 p=0.99 r=0.09	X^2=65.9 p=0.97 r=0.20	X^2=42.3 p=0.70 r=0.17	X^2=32.8 p=0.99 r=0.34

10.2.4 Discussion.

It is notable that most of the items of the 8 original scales fit in a Rasch homogeneous scale despite the strictness of the Rasch-model. This fact might be explained by the homogeneity of our sample. It is striking that the scales fit so well in the Rasch-model, given that internal consistencies in general are only moderate. The reason for this discrepancy might be that the variance in the MAAS-scores is rather restricted because of the homogeneity of our sample. This

relatively low variance suppresses the reliability figures. This fact indicates a fallibility of the classical test theory in which reliability figures, calculated from rather homogeneous samples, turn out to be relatively low because of their relatively low variances in the subject's scores on the scales. In 5.3.1 we have already discussed weaknesses of the classical test theory resulting from sample dependency.

We now take a closer look at the loss of items from the scales "psychiatric examination", "socio-emotional exploration" and "presenting solutions", caused by a non-fit in the Rasch-model (see table 10.2; column 5).

The loss of items from the scale "psychiatric examination" is considerable. This loss concerns those items in particular which pertain to the exploration of disturbances in consciousness and orientation, memory, perceptual and thought disturbances. It is explained by the zero or near-zero variances of the item scores in our experiments. Possibly these items fit in the Rasch-model when tested in interviews that yield higher variances in these item scores. The loss of three items pertaining to the psychiatric examination of anxiety can only be partly attributed to low variances in item scores, but is mainly due to a poor operationalization of the underlying theoretical concepts of anxiety in the items.

The item loss in the scale "socio-emotional exploration" is not severe. Only two items are lost, although the item on perspectives and ambitions in life is important, especially for depressive disorders.

Finally, the unexpected exclusion of two important items from the scale "presenting solutions", "conveying concrete information about the execution of a given advice" and "making appointments for the further follow-up", is considerable in a theoretical sense.

In sum, all MAAS-PMHC scales fit in the Rasch-model. In addition, they all share one similar underlying dimension according the Martin Löf test for uni-dimensionality. They differ in a content validity aspect (see next chapter). However, some caution in interpretation of these results is necessary pertaining to the application of these Rasch homogeneous scales to different populations and to the influence of "case specificity".

Although our population of 6th year psychiatric clerks and residents in general practice is rather broadly defined, our Rasch homogeneous scales might not be applicable in a population of - for example- experienced general practitioners or inexperienced undergraduate students. This restriction is made for three reasons. In the first place, the patterns of interviewing ability in other populations might be so different that the present scale would no longer fit. For such a population, similar research should be carried out to ascertain anew the Rasch homogeneity of the scales.

Secondly, the distribution of item parameters within scales is often unequal and narrowly-spaced, so some item-characteristic curves fall within the confidence bands of others. According to Birnbaum (1974), the informativity of such "narrowly"-scaled items on the subject's ability level is relatively low in these instances. In more simple terms, this means that some of the adjacent items are interchangable. Thirdly, and - partly - the most likely cause of the latter observation, the numbers of subjects used in these analyses are relatively low. This is partly due to the relatively low number of observations for too many cells in the bivariate distributions of score groups and of numbers of subjects within these score groups. Therefore, the probability of capitalization on chance might be rather high.

Another critical remark pertains to the restrictive influence of case specificity. As shown previously (e.g. in chapter 7), case specificity, which can be attributed to differences in mental health problems and to differences in presentation by patients, might cause restraints on the generalizability of the physician's interviewing skills from one "case" to another. Although these findings based on the classical test theory do not hold automatically for probabalistic scaling, they entail the exercise of some caution. The use of only one case of a depressive patient means that not all items of the MAAS-PMHC are used to a sufficient degree as is witnessed by the scoring pattern on the scale "psychiatric examination". For instance, symptoms concerning anxiety, disturbance in perception and thinking etc. are not used.

10.2.5 Conclusions.

All scales of the MAAS-PMHC fit in the Rasch-model without losing items of great theoretical importance except for some items in the scales "psychiatric examination" and "presenting solutions". These scales all have one similar underlying dimension.

10.3 Reliabilities on the scale level: a generalizability study.

Reliability is analyzed on the scale level by means of a generalizability study. The scale level has been chosen because each MAAS-PMHC scale is constructed around a theoretical dimension of medical interviewing skills and because each scale fits in the Rasch-model. An additional reason for the use of the scales as the unit of analysis is the fact that the sum scores of items within each scale are taken as indices for the underlying theoretical dimensions of medical interviewing in the studies of validity.

We selected generalizability analysis as a measurement of reliability because it is the most appropriate approach to the analysis of different sources of variation in a set of scores such as the one studied here. The sources of variation and their interactions may be rooted in differences in agreement between observers, in the physicians' interviewing styles and ability and in measurement properties of the method.

For each scale, different coefficients of generalizability are calculated as reliability measurements. Generalizability coefficients are calculated for measurement situations where one, two and six observers are used. These coefficients are scrutinized in order to investigate how the observers, the physician's interviewing ability and the measurement properties of the method itself all influence the reliability of the MAAS-scores.

10.3.1 Method.

This study in reliability was carried out in a similar sample to that described in chapter 5 and 7 and which is briefly presented here.

In order to secure optimal conditions for measurement, comparability and control, we created a consultation hour in which 40 residents in general practice interviewed four different simulated patients. The

characteristics of this group of 40 residents in general practice have already been described, as has the experimental situation (see chapter 7). Two of the four simulated patients presented a mental health problem to the physician. One patient represented a middle-aged woman with a long-standing, under-treated depression (Diagnostic and Statistical Manual of Mental Disorders-III: major depression) after migration from her native village to a neighbouring town. The second patient represented a middle-aged man who, after losing his job, developed a panic disorder (Diagnostic and Statistical Manual of Mental Disorders-III) with insomnia, mild depression and feelings of shame resulting in family tensions.

In the present studies, we use the interview which the 40 residents held with both patients simulating the mental disorders. A pool of six trained observers rated the interviews with the MAAS-PMHC ("live" scores). The interviews were videotaped simultaneously.

From this collection of 80 videotaped interviews, we randomly selected 20 interviews, 10 of each case. Each of these 20 interviews was scored by the whole pool of 6 trained observers using the MAAS-PMHC. The data base was composed in this way for two reasons: to secure a good impression of the observers' influence, we raised their number to six and to alleviate the burden of scoring, we decreased the number of residents to 20.

The scores of these 120 (6x20) rated interviews were submitted to a three-way analysis of variance where physicians, observers and items were the sources of variation.

In this way, we may generalize from our 20 subjects, 6 observers and 104 items to the "universe" of subjects, observers and items. We fixed the variance component of items, however, because the items of the scale have been carefully selected in order to attempt to encompass the "entire domain" of the underlying theoretical dimension. Such a universe of interviewing skills, pertaining to one dimension, is, however, not endless. We therefore assume that most interviewing skills are covered by the MAAS-items.

The analyses of variance were carried out for each MAAS-scale by means of the General Mixed Model Analysis of Variance with Equal Cell Sizes (BMDP-program P8V; Dixon and Brown, 1979). The size of the

variance components was estimated for the facets physicians, observers and items, and their interactions. These estimates were used to calculate the percentage of the total variance induced by each component.

Coefficients of generalizability were then calculated for scoring situations with respectively one, two and six observers. These coefficients are the ratio between the expected "true" variance (or universe variance) and the observed variance (see also 5.4.3).

To study better the observers' influence on the MAAS-scores, we considered the effects on the generalizability coefficients when we generalized from research situations with one, two or six observers. These effects can be determined, in the formula of the generalizability coefficient, by dividing the number of observers by 6 in the case of one observer and by 3 in the case of 2 observers (Thorndike, 1981). This procedure was accomplished in each of the 8 scales.

10.3.2 Results.

The analyses of variance over the eight Rasch homogeneous scales can be found in appendix G. The estimates of the variance components pertaining to the sources of physicians, observers, items and their interaction effects are given in table 10.4 for each scale of the MAAS-PMHC. The coefficients of generalizability of the Rasch homogeneous scales in research situations, using one, two and six observers are given in table 10.5.

10.3.3 Discussion.

The generalizability coefficients for physicians with the sources of variation "observers" random and "items" fixed are, in general low for 1 observer, moderate for 2 observers and reasonable to good for 6 observers. Mitchell (1979) considers the range of 0.50 to 0.80 as moderate to good when three main sources of variance and their interaction components are taken into account.

The scales show that the coefficients are reasonable for the scales "exploration of the reason for encounter", "history-taking", "psychiatric examination" and "socio-emotional exploration". The scales "presenting solutions", "structuring the interview" and "interpersonal skills" have moderate coefficients, whereas the coefficient for "communicative skills" is low.

Table 10.4: Variance components in the sumscores of MAAS-PMHC scales after three way analysis of variance (in percentages of the total variance).

Source of variation	exploration of the reasons for encounter	history-taking	psychiatric examination	socio-emotional exploration	presenting solutions	structuring the interview	interpersonal skills	communicative skills
physician (p)	2.0	1.7	3.6	2.2	3.7	8.7	2.6	1.7
observer (o)	2.6	1.7	0.5	2.2	8.3	6.7	3.3	4.9
items (i)	18.1	19.9	10.4	27.5	21.1	5.2	43.4	15.2
p.o	4.5	5.0	5.0	2.4	6.9	19.7	4.9	8.3
p.i	24.7	22.4	34.2	28.3	7.3	7.4	3.7	6.4
o.i	6.1	2.5	1.8	2.5	8.7	5.0	10.8	18.6
o.i.p + error	42.0	46.9	44.5	35.0	44.0	47.5	31.4	44.9

Table 10.5: Generalizability coefficients of the MAAS-PMHC scales, based on 1, 2 and 6 observers.

Scales of MAAS-PMHC	Generalizability coefficients (items fixed; physicians and observers random)		
	1 observer	2 observers	6 observers
I. Exploration of the reasons for encounter	0.26	0.42	0.68
II. History-taking	0.26	0.42	0.68
III. Psychiatric examination	0.24	0.39	0.66
IV. Socio-emotional exploration	0.40	0.57	0.81
V. Presenting solutions	0.18	0.31	0.59
VI. Structuring the interview	0.22	0.36	0.71
VII. Interpersonal skills	0.20	0.33	0.59
VIII. Communicative skills	0.11	0.19	0.41

When we scrutinize the variance components of the various sources (table 10.4), it becomes clear why some generalizability coefficients are low or moderate.

The first source of variation, physician (p), together with the fifth source of variation, the interaction between physicians and items (p x i), represents the "true" variance of the physician's interviewing ability in the eight theoretical dimensions represented by the MAAS-PMHC scales (see also Thorndike, 1982). These variance components are

generally low, in particular in the scales "presenting solutions", "structuring the interview", "interpersonal and communicative skills". Since the quality of measurement of the traits has already been secured by the Rasch analyses, the magnitude of this first source of variation is not that important in itself.

The second source of variation, observers (o), concerns the variance caused by systematic differences in scoring by observers. Strict or lenient observers, for example, cause systematic variations in the assessment of the physicians' interviewing skills. According to Saal et al. (1980), a significant proportion of this observers' variance should be interpreted as the traditional leniency effect, defined as the systematic tendency by observers to assign a higher or lower rating than is warranted by the subject's interviewing skills.
The scales "presenting solutions", "structuring the interview" and "communicative skills", seem to be most vulnerable to the leniency effect, whereas the scales "history-taking" and "psychiatric examination" are less vulnerable. However, all these effects are not very strong.

The background of this leniency effect may be "threshold problems" in the scoring of more complex interviewing behavior. For instance, observers may know and agree upon the criteria for scoring an item, such as "discussion of the pros and cons of the proposed help". In observing an interview, observers may disagree on whether such a discussion took place completely or only partly. Observers holding the latter opinion will decide that the threshold for positive scoring has not been reached.

The third source of variation, items (i), reflects the variation in the items "used" during the scoring of the MAAS-PMHC. The variation in which items are scored positively depends on two factors. First, the physician's interviewing style reacts more or less flexibly to the mode of self-presentation of the patient. This flexibility requires a variation in the skills needed for the process of interviewing. Examples of item variation due to differences in style are found in relatively high scores of "presenting solutions" and - in particular- "interpersonal skills". Second, the variation in mental health problems requires different content elements in the interview. Examples of the

influence of this factor are found in the relatively high components in the scales "exploration of the reasons for encounter", "history-taking" and "socio-emotional exploration".

This frequently considerable source of variation reflects in general the difficulties of translating medical interviewing skills into items.

The fourth source of variation is the interaction between physicians and observers (p x o), known in the literature as "halo-effect". It is defined as the observer's failure to discriminate among conceptually distinct and potentially independent aspects of a subject's behavior (Saal, et al., 1980). In general, halo-effects are not very high in the MAAS-PMHC (variance components of about 5%), except for the scales "structuring the interview" and, to a lesser extent, "communicative skills". We expected halo-effects to be high in the scales "presenting solutions" and "interpersonal skills" as well, because these scales measure larger units of difficult - to - define interview behavior. Moreover, they require observers to indicate their personal opinion of the physicians' skills. It is commonly acknowledged that halo-effects are high under such conditions. It is not clear whether the high halo-effect in the scale "structuring the interview" should be explained by the complex interviewing behavior causing unreliable scoring or by the low validity of the underlying concept.

The latter argument is supported by the fact that, in primary mental health care, the distinction between the phases "exploration of the reasons for encounter" and "history-taking" is not clear because, in both phases, much patient-centered information is collected. The structuring of the interview in phases is thus difficult to measure.

The fifth source of variation, the interaction between physicians and items (p x i), is considered in the literature as true variance and is therefore added to the physician's facet (p) (Thorndike, 1982). This source of variance is the highest in the scales measuring skills characteristic for the three phases of the medical interview. Content elements play a major role here. Depending on the specific content aspects of a case, the physician will ask certain questions which might well not be posed in other cases. Parallel to the discussion of items as source of variation, one should conclude that different cases yield different patterns of scored items. In the scales which measure process

aspects of the interview that are less dependent on the case, the physician and items interaction component is less prominent. These findings suggest a degree of case-specificity to which we address ourselves further in 10.4.

The sixth source of variance, the interaction between observers and items (o x i), refers to the differences between observers in interpreting the meaning of items and the criteria for scoring. This variance component is the greatest for the scale "communicative skills". It is clear that this source of variance may be responsible for the low generalizability coefficient of this scale. Apparently, items like confrontation, concretization, conveying information in small units etc. are liable to differences in interpretation because the criteria for scoring may be confusing. The combination of qualitative (how adequately accomplished) and quantitative criteria (how many times accomplished) might be too intricate for appropriate scoring. Improvement will be made by the splitting of items and by introducing more unequivocal criteria.

The seventh source of variation is the interaction between physicians, observers and items also including an error component (p x o x i + error). This considerable variance component suggests that influences other than the above-mentioned sources of variation also impinge on the variation of the scores, such as error. The observers' ratings may be influenced by fluctuating attention, mood, fatique and pressure of time. All scales of the MAAS-PMHC are considerably affected by this variance component (ranging from 31.4 to 47.0). Although in this source of variation, the error is not distinguishable from the p x o x i component, we have to conclude that much of the physician's interviewing style is not covered by MAAS-PMHC items. This is the price paid for our operationalization of the physician's interviewing behavior into clearly defined, teachable skills. In particular, the exclusion of many non-verbal behavior from the MAAS may be a reason for this considerable, unexplained component of variance.

The observer's influence on the MAAS-scores becomes clear from the comparison of the generalizability coefficients based on measurement by one, two and six observers. Rising from one to six observers, the generalizability coefficients increase from low to good. This finding suggest a marked observers' influence on the scores.

Despite this conclusion, for the forthcoming validity studies with the MAAS-PMHC we use scores summated over two observers to control the observers' effect.

10.4 What is the inter-case reliability of the MAAS-PMHC?

The question treated in this section concerns the sensitivity of the MAAS-PMHC to differences in cases. Instruments measuring medical competence, including medical interviewing skills, are always susceptible to "case influence" (Swanson, et al., 1981). What does "case influence" mean?

The physician's interviewing style is influenced by the "case", in the sense of the nature of the medical problem (its complexity, severity, aetiology etc.), but also by the way the patient presents it. Self-presentation by the patient is determined by personality traits (intro-extroversion, etc.), affective states (anxiety, mood), attitudes, social norms, locus of control etc.

In chapter 7, we attempted to disentangle different aspects of this case influence. The results suggest that the influence caused by the patient's self-presentation explains most of the variance, especially in the scale "exploration of the reasons for encounter" and "presenting solutions". The case influence in the "history-taking" is caused by the case as a medical problem.

However, our design did not allow us to replicate this study because simulated patients and cases are not completely crossed (simulated patients are irregularly nested in the cases). We therefore relied once more upon the generalizability analysis to assess the "not disentangled case influences" in the variance of MAAS-PMHC scores.

10.4.1 Method.

The subjects for this study were once more the 40 residents in general practice, each interviewing two patients simulating respectively a major depression and a panic disorder. The interviews were scored "live" on the MAAS-PMHC (in its "Rasch version") by trained observers. This data base has been described in 10.4.2. From this data base, we took the "live" scored 80 interviews (40 residents x 2 cases). We again applied a three-way analysis of variance to the item scores of

these interviews. Our purpose was to generalize to physicians with the sources of variation "cases" random and "items" fixed. To avoid a more difficult - to - interpret four-way analysis of variance, we excluded the observers' variation from our generalizability analysis. This observers' effect was thus lost in the error term of the variance components. A coefficient of generalizability was then calculated for physicians with "cases" random and "items" fixed for each scale of the MAAS-PMHC.

To gain an insight into the impact of case influences on the MAAS-scores, we estimated the generalizability coefficients when the number of cases was raised to 20 and 40 cases. These estimations were calculated by substitution of these numbers of cases in the equation of the generalizability that was previously calculated for 2 cases.

10.4.2 Results/discussion.

The case influence on the MAAS-scores was studied in a generalizability analysis. Its results, the generalizability coefficients for the 8 scales, are presented in table 10.6.

In general, these generalizability coefficients are low, with the exception of "socio-emotional exploration". These results indicate a lack of inter-case reliability of the MAAS-PMHC.

Raising the number of cases to twenty makes the coefficients moderate to very reasonable, except for the scales "history-taking" and "interpersonal skills" which appear to be very susceptible to "case influences". The influence in the first scale is due to differences in mental health problems, whereas in the second scale, the impact of the self-presentation by the patient is sensible.

Low inter-case reliability due to susceptibility to differences in mental health problems is only avoidable when items pertaining to content elements are more abstractly formulated to the level of "grouped topics" (see 4.2). For example, items are constructed that assess the exploration of the patient's social functioning in general instead of its distinctive aspects. When this is not possible, all content elements would be removed from the MAAS-PMHC, which is, of course, absurd.

Table 10.6: Generalizability coefficients for cases and physicians' interviewing skills (physicians and cases random; item fixed) and estimated generalizability coefficients for 20 and 40 cases.

Scales of the MAAS-PMHC (Rasch homogeneous)	Generalizability coefficients calculated from 2 cases	Estimated generalizability coefficients for 20 cases	for 40 cases
I Exploration of the reasons for encounter	0.20	0.72	0.84
II History-taking	0.01	0.15	0.26
III Psychiatric examination	0.12	0.58	0.73
IV Socio-emotional exploration	0.32	0.82	0.90
V Presenting solutions	0.14	0.62	0.77
VI Structuring the interview	0.24	0.76	0.86
VII Interpersonal skills	0.08	0.48	0.64
VIII Communicative skills	0.22	0.73	0.85

However, the low inter-case reliability due to self-presentation by patients can be improved by using better-trained simulated patients whose self-presentation has been standardized. This is, in fact, only possible in test situations in medical education.

10.5 Conclusions from the generalizability analyses of the MAAS-PMHC scales.

In generalizability analyses, the influence of observers, of method of measurement (items) and of cases on the physician's interviewing skills is determined.

To gain an impression of the observers' influence, generalizability coefficients for physicians are calculated with the sources of variation "observers" kept random and the source of variation "items"

kept fixed. The coefficients are generally low when one observer, moderate when two observers and satisfactory when six observers are used.

A moderate to reasonable interrater reliability may be concluded from these findings.

Regarding the "performance" of the scales, we notice that "exploration of the reasons for encounter", "history-taking", "psychiatric examination" and "socio-emotional exploration" are satisfactory. However, "presenting solutions" and "interpersonal skills" have moderate, and "communicative skills" have even low coefficients.

Causes of unreliability are studied by inspecting the main and interaction components from the analysis variance.

The variation of the most important component, the physician's interviewing skills, is substantial but rather low compared to the other variance components.

Leniency effects i.e. systematic high or low rating by the observers, are not high in general, but notable in the scales "presenting solutions", "structuring the interview" and "communicative skills". It is apparent that the threshold to considering an interviewing skill as present or absent is an important reason for these differences between observers.

Halo-effects in scoring originate from characteristics of subjects (appearance, attitude, etc) which are not related to their interviewing skills, but which influence the observers to systematic high or low ratings. These effects are seen in the scales "structuring the interview" and "communicative skills", witnessing poor operationalization and/or low validity of the underlying concepts.

The high "item component" is characteristic for measurement of skills. This phenomenon in the measurement of interviewing skills is due to the method's susceptibility for different mental health problems and - in particular - for differences in self-presentation by patients.

Finally, it is notable that a considerable variance component is not explainable (the phy x obs x item + error component). Although this component is generally high in measurement because of the high number of degrees of freedom, this may also be due to our exclusion of non-verbal interviewing behavior that is not easily teachable.

"Case influence" turns out to be a high variance component in MAAS-scores. It is marked in particular in the scales "history-taking" and "interpersonal skills". In the analysis, where we generalized to the universes of physicians and cases, it could not be clearly discerned whether this inter-case unreliability should be attributed to real differences in medical problems or to the influence of the way patients present their problem to the physician. Evidence from the MAAS-General Practice points to the latter of the two causes. Reducing these "case influence" effects requires about twenty cases.

Is the MAAS-PMHC usable in the light of these reliability considerations?

In general, causes of unreliability, like halo-effects and non-systematic error of measurement, should be improved by simplification of items based on more clearly-defined and delineated interviewing behavior and by better-described criteria for scoring. However, these interventions imply a further selection of items, perhaps entailing a more simple and reliable but less valid item domain. This price may be too high for a measurement method used in research into medical interviewing skills.

Leniency effects, a typical observers' characteristic, may be diminished by better training and - sometimes - by selection of observers.

In test situations using 1 observer, there will be a considerable observer effect. This can be compensated for by observer training, or by the use of two or more observers.

"Case influence" may be reduced by standardizing simulated patients in their presentation of cases.

In research situations, where interviews are videotaped, the MAAS-PMHC is very useful. In these situations, at least two observers and more cases can be used to reduce respect observers' and case influences respectively.

REFERENCES

Birnbaum A. Some latent trait model and their use in inferring an examiner's ability. In: Lord FM, Novick MR (Eds.). Statistical theories of mental test scores. Addison-Wesley, Reading, 1974 (2nd ed.).

Dixon WJ, Brown MD, BMDP-79: biomedical computer programs, P-series. Univ. Calif. Press, London, 1979.

Diagnostic and Statistical Manual of Mental Disorders, third edition (DSM-III). American Psychiatric Association, 1980.

Guilford JP, Fruchter B. Fundamental statistics in psychology and education. MCGraw-Hill, Auckland, 1982 (6th ed.).

Gustafsson JE. The Rasch model for dichotomous items: theory applications and a computer program. Reports from the Institute of Education. Univ. of Göteborg, nr. 85.

Hambleton RK, Cook LL. Latent trait models and their use in analysis of educational test data. Journal of Educational Measurement, 1977; 14: 75-96.

Martin-Löf P. Statistical models. Notes from seminars 1969-1970 by Rolf Sunberg. Instute för försäkrings-metamatik och matematik vid Stockholms Universitet, Stockholm, 1973.

Mitchell SK. Inter-observer agreement, reliability and generalizability of data collected in observational studies. Psychological Bulletin, 1979; 86: 376-390.

Molenaar IW. Programma-beschrijving van PML (versie 3.1) voor het Rasch model. Heymans Bulletins, Vakgroep Statistiek en Meettheorie, Universiteit van Groningen, Groningen, 1981.

Saal FE, Downey FG, Lahey NA. Rating the ratings: assessing the psychometric quality of rating data. Psychological Bulletin, 1980; 88: 413-428.

Swanson DB, Mayewski RJ, Norsen L, Baran G, Mushlin AI. A psychometric study of measures of medical interviewing skills. Proceedings of the 20th Annual Conference on Research in Medical Education. 1981: 3-8.

Thorndike RL. Applied psychometrics. Houghton, Mifflin Cie., Boston, 1982.

CHAPTER 11 CONTENT VALIDITY OF THE MAAS-PMHC

H.F. Kraan and A.A.M. Crijnen

11.0 Introduction.

The core question in this study of content validity reads as follows: is the content concerning (initial) medical interviewing skills in primary mental health care adequately operationalized in the MAAS-PMHC items?

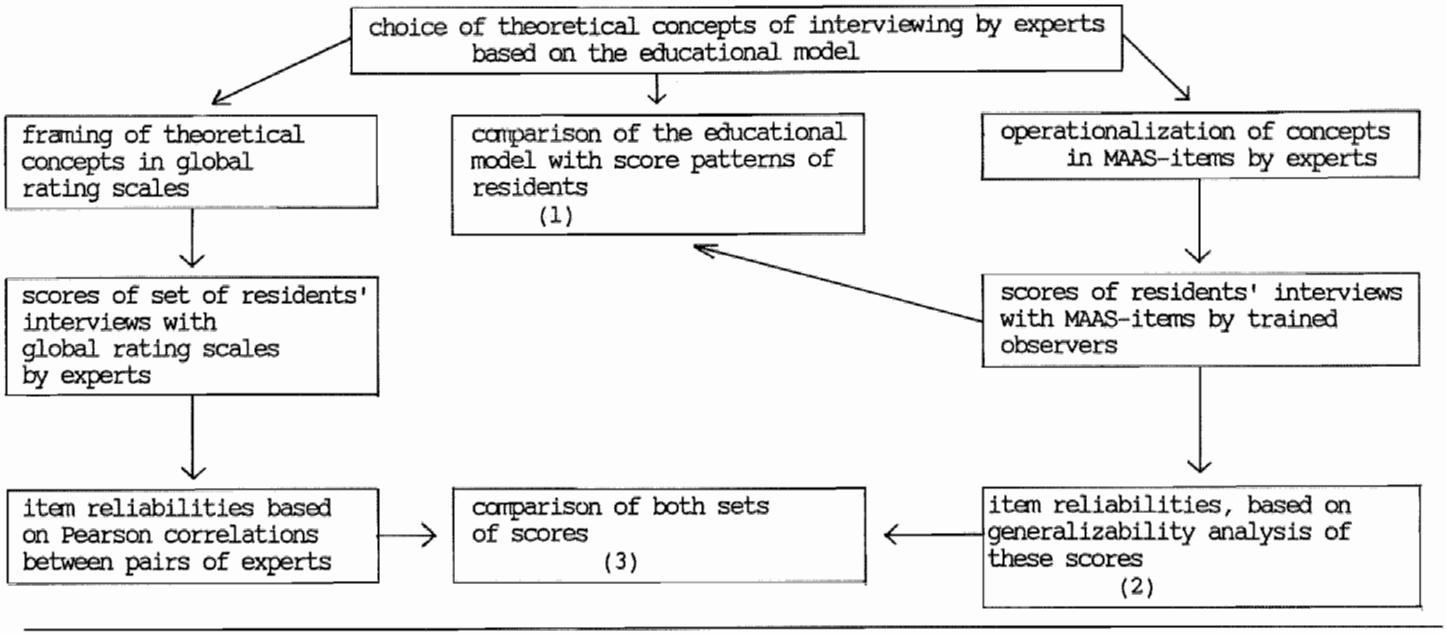
Unfortunately, no procedure for empirical quantitative assessment of content validity has been recognised in the literature (De Groot, 1961).

We therefore propose an indirect procedure for the study of the content validity: this is illustrated in table 11.1.
This procedure runs as follows:

- Starting-point is the construction of the MAAS-PMHC for which we have chosen the theoretical concepts of interviewing and where the operationalization of items has been performed (described in chapter 4).
- Scoring of a set of residents' interviews with the MAAS-PMHC by trained observers. Assessment of item reliabilities by generalizability coefficients.
- Scoring of the same set of interviews by experts who use the Global Expert-Rating Scales. This scale measures the same dimensions of interviewing which also underly the MAAS-PMHC. Item reliabilities are calculated as Pearson correlations between pairs of experts.
- Comparison of both sets of scores by correlating the MAAS-scores on the scale level with the corresponding items of the Global Expert-Rating Scale.
- Inspection of the residents' scores on the MAAS-PMHC for the representativeness of the original theoretical concepts in their scoring patterns.

Although all steps of this procedure should be checked for their correctness, we carry out this check on 3 points. The most important step in this procedure, however, the operationalization of the content of interviewing into items, cannot be investigated directly. When these

Table 11.1: Procedure to study the content validity of the MAAS-PMHC.



checks turn out to be positive, it supports the content validity of the MAAS-PMHC. In checking these steps, the three following questions are asked (indicated by their number in table 11.1):

1. What interviewing skills are used by residents in general practice? How representative is our educational model of interviewing skills as operationalized in the MAAS-PMHC for this group?
   These questions are answered by studying the interviewing skills residents in general practice display during the simulated consultation. By this approach, content validity on the item level is investigated (11.1).
2. Do each of the items pertaining to a single interviewing skill show a sufficient reliability?
   To answer this question from the perspective of content validity, the sources of unreliability should be explored. Are the theoretical concepts as described in chapter two adequately operationalized, resulting in agreement between observers or is the item construction insufficient in a technical sense, this acting as a source of error? These questions are answered by using a generalizability design which allows the study of the different components of variance in the MAAS-PMHC-scores on the item level (11.2).
3. According to the opinion of a panel of experts, do the scales of the MAAS-PMHC reflect the theoretical dimensions of initial interviewing in primary mental health care?
   In a correlative study, MAAS-PMHC-scores are compared with the ratings by experts over the same interviews. In this approach, content validity research takes places on the scale level (11.3).

11.1 Content validity and interview behavior of residents in general practice.

A contribution to content validity of the MAAS-PMHC emerges from the scoring patterns of residents in general practice. These patterns of scoring provide a picture of the interviewing skills of this group. The contribution of these scoring patterns to the content validity is based on two rather self-evident arguments:

- In their under-and postgraduate education, the residents have been made more or less familiar with the theoretical aspects of interviewing in Primary Mental Health Care.
- In their present postgraduate training, the residents have to cope at a primary care level with the mental health problems of their patients.

11.1.1 Method.

We studied the frequency of positive MAAS-PMHC item scores of the residents in both cases of the simulated consultation hour. A picture of their interviewing skills thus evolves. As a criterion for infrequently used interviewing skills, we have taken the mean minus 1 standard deviation as a cutting point. Cutting points of one standard deviation above or below the mean are common in group referenced educational measurement as used at the University of Limburg Medical School. This implicates that the cutting point for infrequently used interviewing skills is below the p-value of 0.16.

11.1.2 Results.

In table 11.2 can be seen which of the percentage scores (p-values) of the physicians' interviewing skills are below 0.16.

In general, all interviewing skills are used in the scale "exploration of the reasons for encounter" except for questions about coping with problems in the past and about the impact of the problems on others.

The "history-taking" scale is also generally supported, but the two important items about the factors, maintaining problems and functionality (gains), are under-used.

In the scale "psychiatric examination", scores on examination of disturbances in consciousness, thought, perception and memory are lacking, because these items have barely been used in the consultations of the residents with simulated patients (see also 10.2.4).

In the scale "socio-emotional exploration", many items (11 out of 18 Rasch homogeneous items) show low percentage scores. Conspicuous are the few questions asked about aggressive, affectionate and religious feelings. In addition, questions on caring, responsibility, substance

Table 11.2: Generalizability coefficients on the scale level, item generalizabilities, observers variances and score patterns by residents.

Rasch homogeneous items (key words)	generalizability coefficient for 2 observers	percentage of total variance attributed to observers	percentage of positively scored items by residents in general practice
I. Exploration of the reasons for encounter.			
1. reason for visit	0.17	26	83
2. description complaint(s)	0.23	23	35
3. emotional impact	0.14	25	31
4. problem presentation now	0.66	1	38
5. opinion about cause	0.53	1	78
6. discussion in family	0.73	5	61
7. own solutions	0.36	15	51
8. consequences daily life	0.34	16	28
9. life circumstances	0.75	0	30
10. habitual solutions	0.24	2	15
11. burden to others	0.26	0	8
12. recent life-events	0.57	17	36
13. desired help	0.78	0	56
	averaged generalizability on the item level 0.44		generalizability on scale level 0.42
II. History-taking.			
14. intensity complaint(s)	0.22	2	23
15. course during day	0.45	4	36
16. history complaint	0.46	13	99
17. provoking factors	0.39	10	59
18. increasing factors	0.23	4	28
19. maintaining factors	0.00	6	6
20. decreasing factors	0.33	4	16
21. functionality/gains	0.11	18	9
22. mental problems in past	0.38	0	9
23. prof. treatment in past	0.33	0	24
24. consult. in present	0.00	0	2
25. (ab)use medication	0.85	3	32
26. (pseudo) heriditarity	0.82	2	4
	averaged generalizability on item level 0.35		generalizability on scale level 0.42

Table 11.2: (continued)

Items	gen.coeff. 2 obs.	% obs.var.	% pos. items
III. Psychiatric examination.			
27. disturbances mood/affect	0.50	6	21
28. biol. depres. symptoms	0.42	4	16
29. depressive cognition	0.10	1	14
30. suicidal behavior	0.82	1	18
31. anxiety symptoms	0.39	2	16
35. disturb. consc./orient.	0.00	0	4
37. disturb. immed. memory	0.00	0	1
38. disturb. short memory	0.00	0	1
42. disturb. stream thought	0.00	0	1
	averaged generalizability on item level 0.25		generalizability on scale level 0.39
IV. Socio-emotional exploration.			
45. feelings love/affec.	0.44	7	4
46. aggressive feelings	0.02	0	3
48. care giving	0.01	19	4
50. religious feelings	0.00	12	1
51. character/self-image	0.49	6	4
52. relations family	0.36	8	55
53. social support	0.40	12	31
54. cultural differences	0.55	5	15
55. prof. functioning	0.69	6	62
56. leisure time	0.62	7	41
57. sexual functioning	0.00	0	1
58. sleeping habits	0.66	7	51
59. eating habits	0.93	0	29
60. substance (ab)use	0.00	0	3
61. housing condition	0.56	13	29
62. financial situation	0.79	1	5
63. education/profession	0.00	0	6
64. development	0.00	0	1
	averaged generalizability on item level 0.36		generalizability on scale level 0.57

Table 11.2: (continued)

Items	gen.coeff. 2 obs.	% obs.var.	% pos. item
V. Presenting solutions.			
65. conveys diagnosis	0.28	21	31
66. info. causal factors	0.39	16	39
67. info. prognosis	0.46	12	5
68. patients' expections	0.47	6	41
69. responsib. treatment	0.19	46	16
70. proposal help	0.02	60	81
71. explains proposal	0.17	29	51
72. pros/cons proposal	0.19	4	13
73. pat. opinion proposal	0.22	14	63
74. influence by others	0.40	12	13
75. discus. diff. opinion	0.20	18	14
76. choice proposal	0.03	5	13
78. advice understood?	0.03	62	16
	averaged generalizability on item level 0.23		generalizability on scale level 0.31
VI. Structuring the interview.			
82. summarizes reason visit	0.68	3	36
83. ordering HIT, PE, SE	0.28	11	19
84. EE precedes HIT, PE, SE	0.20	1	49
85. PS after EE, HIT, PE, SE	0.54	29	52
86. starts PS with diagn.	0.25	22	33
	averaged generalizability on item level 0.39		generalizability on scale level 0.31
VII. Interpersonal skills.			
88. facilitation	0.26	18	86
91. reflects emotions	0.37	8	59
92. react. express. emotion	0.19	4	6
93. feelings of the moment	0.26	20	12
97. meta-communication	0.22	13	11
99. proper history-taking	0.06	38	26
100. puts at ease	0.38	19	38
101. proper pace	0.10	6	79
103. congruent non-verbals	0.38	21	92
104. proper eye contact	0.02	92	76
	averaged generalizability on item level 0.22		generalizability on scale level 0.33

Table 11.2: (continued)

Items	gen.coeff. 2 obs.	% obs.var.	% pos. items
VIII. Communicative skills.			
89. proper closed quest.	0.20	33	56
90. concretises	0.18	16	58
94. summarizes	0.55	13	73
95. info. in small units	0.03	26	58
96. checks understanding	0.05	54	34
98. proper confrontations	0.12	6	24
102. understandable language	0.05	2	91
	averaged generalizability on item level 0.17		generalizability on scale level 0.19

(ab)use, sexual functioning, housing and financial situation and developmental issues are scarce. This finding shows that residents do not adhere strongly to the preferred theoretical style of initial interviewing in primary mental health care which recommends a balanced combination of a general, non-directive style, with directive or systematic questioning (see 2.4.2.2.3).

The scale "presenting solutions" shows low percentage scores in some interviewing skills which have to do with the negotiation process: conveying of information about pros and cons of the treatment plan, discussion of differences in problem-definition and inviting the patient to make a choice from different treatment alternatives. These findings might shed some light on the imperfect style of negotiation of residents in general practice.

In the scale "structuring the interview", the items on the consultation plan and on the check whether the most important problems have been discussed are under-used. This scale is otherwise generally supported.

The same holds in general for the scales "interpersonal and communicative skills". The interviewing skills operationalized in these items are generally used by the residents. Only items about the feelings of the patient during the consultation and about meta-communicative comments are under-used.

11.1.3 Discussion.

In their 15-20 minutes interviews with simulated patients, residents show a variability in their interviewing behavior which generally covers the item domain of the MAAS-PMHC. In this way, the content validity of the MAAS-PMHC is supported by the interviewing pattern of the residents in general practice.

There are, however, some exceptions such as items concerning hypotheses generation and functionality of the problems which are under-used during history-taking. In addition, a considerable number of items from the scale "socio-emotional exploration", particularly emotion-related items, are under-used. Finally, the items on the negotiation process in presenting solutions are infrequently scored. The last two findings may be due to the fact that the simulated situation does not invite the student to explore emotional topics or to negotiate with the patient.

These minor deviations from our educational model seem to point to deficiencies in the interviewing styles of residents, or to restrictions of the simulated situation, rather than to the lack of content validity of the MAAS-PMHC.

11.2 Content validity and item reliability.

When an item is unreliable, the theoretical content covered is not appropriately represented in the measurement method. The consequence is "theoretical loss" during measurement. Although these statements show the relationship between reliability and content validity, item reliabilities studies are not a comprehensive check on the process of operationalization of theory into items. It is one step in the study of content validity (see table 11.1).

A generalizability design enables the detection of the sources of unreliability in items. In the MAAS-scores, variance components may be discerned such as variance due to the physicians' skills, due to the observers' interpretation of the items and due to interactive effects between physicians and observers. The latter two variance components are to be considered as sources of unreliability. Moreover, it is important to note that we have used Rasch homogeneous scales.

In the conclusion of this section (11.2.3), we summarize the content validity on the item level including the "theoretical loss" already suffered by exclusion of items during Rasch analysis to obtain homogeneous scales (c.f. 10.2.4).

11.2.1 Method.

Firstly, we computed the generalizability coefficients over each of the 104 items of the MAAS-PMHC. In 10.4, we have already justified this type of reliability research on the item level. We again used the same design as described in 10.3, where 20 videotaped interviews (10 for each case), all rated by 6 observers, provide the data for generalizability analysis.

Secondly, an analysis of variance was carried out where physicians and observers were taken at random, but each item was, of course, fixed. In this way, 3 components of variance originate concerning, respectively, the physicians' ability, the observers' interpretation and their interaction component, including error.

Thirdly, we inspected the generalizability coefficients and the variance components caused by the observers. This last component has to do with systematically different interpretations of items by observers of each item. This observer component was compared with its p-value in table 11.2. This source of unreliability is more serious when the item is regularly used (cutting point p-value $\geq .16$ as in 11.1).

Finally, we considered the "theoretical loss" caused by the removal of unreliable items. As a minimum of reliability, we took a generalizability of 0.35, a criterion given by Mitchell (1979) calculated over 2 observers, as we base the validity studies of the MAAS-PMHC on MAAS-PMHC-scores resulting from summed ratings by two independent observers.

11.2.2 Results.

Scrutinizing table 11.2 for items showing a generalizability coefficient lower than 0.35, we discuss the "theoretical loss" that takes place when these items are removed from the Rasch homogeneous scales.

In the scale "exploration of the reason for encounter", 6 out of 13 items would be dropped, amongst which are "reason for visit", "thorough exploration of the complaints and symptoms" (from the patient's frame of reference) and "exploration of the emotional impact of the problem upon the patient himself and his important others". In these items, the variance component attributed to the observers is substantial (≥20%). The size of this component may explain its low generalizability coefficients. Furthermore, the items about the "consequences of the problems on daily life" and the item about "problem solving and coping in the past" are unreliable. Some of these items represent a considerable loss of information: the emotional impact of the problem and its significance for daily life. This loss is harmful to the completeness of patient-centered information, a major aim of this phase.

In the scale "history-taking", the theoretical loss mainly concerns the exploration of the conditions pertaining to aetiology (factors that have increased, decreased or that have maintained the problem/complaint) and factors that have to do with the functions e.g. gains of the complaint. With this loss, this scale becomes somewhat vulnerable in measuring the data collection by the physician in order to generate explanatory and treatment hypotheses.

In the scale "psychiatric examination", evaluation of the content validity aspects by means of reliability criteria is hampered by the restrictions of our experimental situation. Since our patients did not present clinical pictures with marked disturbances in perception, thought, memory and consciousness, some of the items pertaining to these symptomatology show very low to none variance. Consequently, analysis of variance could not be performed to provide data for generalizability analysis. Finally, 5 out of 9 items would be removed from this scale, among which are important theoretical items pertaining to afore-mentioned symptomatology. We do not know the consequences of this loss. Reliability of these items should be assessed in interviews pertaining to this symptomatology. However, these findings do not immediately imply that the aim of this scale i.e. measurement of the thoroughness of psychiatric examination, is not attained. The restricted number of unreliable items shows that this item format is

promising for evaluation of the thoroughness in psychiatric examination.

In the scale "socio-emotional exploration", the exclusion of unreliable items is substantial (7 out of 18). The theoretical loss mainly concerns emotional aspects (exploring feelings of aggression, responsibility, caring and future expectations) and, to some extent, biographical data collection. This unreliability correlates with the low frequency with which these topics have been raised by the physician during the interview and may be due to threshold problems in scoring (see 10.3).

The scale "presenting solutions", although almost entirely Rasch homogeneous (13 out of 15 items), suffers from a general unreliability on the item level. The subsequent theoretical loss pertains to both functions dealt with during this phase of the interview: firstly, the conveyance of information on causal conditions of the problem, the rationale and pros/cons of the treatment plan and, secondly, the negotiation between physician and patient about problem definition, treatment methods and goals. Five items of this scale show a considerable (≥20%) observers' variance component, indicating a definition disagreement between observers which can be corrected.

In the scale "structuring the interview", a considerable theoretical loss has already taken place during Rasch analysis (3 items out of 8). The low reliabilities on the item level in particular concern the items on the structuring in the sequence of the phases in the initial interview. In analogy with the previous scale, it may be concluded that experts agree on some structuring of the phases, e.g. the introduction and closure of topics, but disagree on the fixed sequence of the phases "exploration of the reason for encounter", "history-taking", "psychiatric examination", "socio-emotional exploration" and "presenting solutions".

The scales "interpersonal and communicative skills" both suffer from a deficient reliability on the item level, especially the latter. Although both scales are entirely Rasch homogeneous, in the scale "interpersonal skills", seven out of 10, and in the scale "communicative skills", six out of 7 (!) items would be removed because of unreliability.

Looking closer, strikingly low reliability is found in items pertaining to facilitative behavior, such as proper history-taking and proper pacing of the interview (interpersonal skills), concretising, conveying information in small units, checking of understanding, proper confrontations and comprehensible language (communicative skills). Also conspicuous is the high number of items with a substantial observers' variance component (in interpersonal skills: 4 and in communicative skills: 3) again witnessing definition and criterion problems.

We end this section with two more general observations. Firstly, we notice that when the averaged generalizability of the Rasch homogeneous items are compared with the generalizability coefficients on the scale level (10.3.2), then several discrepancies are noted. The coefficients on the scale level are higher in comparison with the averaged coefficients on the item level, except "exploration for the reasons for encounter", "structuring the interview" and "communicative skills".

Secondly, there are many regularly used items with substantial ($\geq$ 16%) observers' variance components in the scales "exploration of the reason for encounter" (5 out of 13), "presenting solutions" (7 out of 13), "structuring the interview" (2 out of 5), "interpersonal skills" (5 out of 10) and "communicative skills" (4 out of 7).

11.2.3 Discussion.

Reviewing the items in the scales after a removal of non-Rasch homogeneous and unreliable items (see table 11.3 for an overview of the remaining items), we reassess the content validity on the scale level:

In the "exploration of the reasons for encounter" scale, there is theoretical loss in interviewing behavior pertaining to the emotional aspects of the problems and their significance for daily life. The loss of emotional aspects from the interviewing behavior is even more marked in the scale "socio-emotional exploration". In the "history-taking" scale, a qualitatively important loss concerns the items measuring the physicians' interviewing behavior serving clinical problem-solving. The scale "psychiatric examination" has promising features as to the measurement of thoroughness in exploration of symptomatology, though some loss took place due to artefacts in the experimental situation.

The findings concerning the "presenting solution" scale are somehow ambiguous. The reliability on the item level is often low, but the scale has a considerable Rasch homogeneity evidencing an increase in consistency and stability when one moves from the item level to the more abstract scale level. "Structuring the interview" shows a low to moderate reliability on the item level. The reasons for this may be two-fold. First, the phases in the interview are more difficult to distinguish in primary mental health. In 10.4.3 we have already argued that the distinction line between doctor- and patient-centered information is difficult to draw, as is the line between "history-taking" and "socio-emotional exploration". In general, interviewers use a more variable, less structured interviewing style. Second, observers might not adhere to our operationalization of the items and induce a source of unreliability.

Finally, the scales "interpersonal" and a fortiori "communicative skills" suffer from a low to moderate reliability on the item level although both are Rasch homogeneous. In 10.2.4 we have given an explanation for this discrepancy.

We conclude this section with three more general remarks.

Firstly, most scales, except "exploration of the reasons for encounter", "structuring the interview" and "communicative skills", show an upward jump in reliability when going from the item to the scale level. Apparently, observers adhere more strongly to our concepts of interviewing on the broader and more globally defined scale-level than on the stricter item level.

Secondly, we have seen that a considerable number of items in the scales measuring process skills suffer from unreliability caused by a high observer variance component.

This finding, based on systematic observers' biases, is repairable by improving item definitions and criteria, as well as ameliorating the observers' training.

Thirdly, we repeat once more that with this reliability approach to content validity study, we are not able to infer conclusions about the quality of the operationalization of concepts into items. All statements made so far pertaining to a lack of content validity or theoretical loss, may also be explained by deficits in operationalization which we cannot detect directly, as we argued at the beginning of these paragraphs.

Table 11.3: Review of numbers of item per MAAS-scales after Rasch analysis and removal of unreliable items.

	number of items		
MAAS PMHC SCALES	original scale	Rasch scale	Rasch scale after removal of unreliable items
expl.reas.enc	13	13	7
hist.taking	13	13	6
psych.exam.	18	9	4
soc-emot.expl.	20	18	11
pres.sol.	15	13	4
struct.interv.	8	5	2
interp.skills	10	10	3
comm.skills.	7	7	1
TOTAL	104	88	38

11.3 Content validity and experts' opinion.

According to the opinion of experts, do the MAAS-PMHC scales reflect initial interviewing in primary mental health care?

The procedure of construction of the MAAS-PMHC by a core group of experts has been described in chapter 4. The content validity of the MAAS-PMHC-items is studied by comparing MAAS-PMHC-scores with global expert ratings of the same interviews.

11.3.1 Method.

An expert panel, different from the core group of experts involved in the construction of MAAS-PMHC, evaluated the set of 80 videotaped interviews which had been scored previously by trained observers with the MAAS (see 10.2.1).

The expert panel consisted of 4 psychiatrists, 1 third-year resident in psychiatry, 1 androgologist, 2 social workers and 1 psychiatric nurse. They were all experienced in primary mental health care and in the practical training of undergraduate students and residents.

To evaluate the videotaped interviews, the panel used the Global Expert-Rating Scale, a 9-item, evaluative 5-point Likert scale (see appendix F). The items are very globally defined, asking an opinion about subjects' exploration of the reason for encounter, history-taking, presenting solutions, structuring the interview, interpersonal and communicative skills. History-taking is measured by two items: "history-taking to generate and test explanatory hypotheses" and "history-taking to generate treatment hypotheses". In this way, experts are invited to evaluate history-taking from a diagnostic and a therapeutic point of view. This instrument ends by asking for an overall evaluation of the interview. With this Global Expert-Rating Scale, the experts rated 80 videotaped interviews (40 residents in general practice each interviewing 2 simulated patients, see 10.2.1). Each of these 80 interviews were rated by 2 experts randomly drawn from the panel, yielding 160 rated interviews.

The 80 videotaped interview were also scored twice. The first time, live during the interviews; the second time, a few months later from the videotaped interviews.

In order to reduce the observers' source of unreliability, the Global Expert-Rating of both experts, as well as the scores of both observers with the MAAS-PMHC, have been summated.

The correlations between the summated MAAS-PMHC scores on the scale level and the summated Global Expert-Ratings are studied to assess to which degree the experts support the MAAS-PMHC-scores. The magnitude of the correlations serves as a measure of content validity. Special attention is given to the question of whether experts actually make the theoretical distinction between interpersonal and communicative skills. In addition, we examine the issue of what experts have in mind when they consider an interview to be good.

However, before studying these validity coefficients, we shall make some remarks about the Global Expert-Rating Scale in terms of its inter-rater reliability and internal consistency. First, we studied the inter-rater reliability by calculating the Pearson's correlation between the item scores of the randomly combined pairs of experts who rated the 80 videotaped interviews. Second, a Cronbach alpha for the Global Rating Scale was calculated for the 160 rated interviews.

11.3.2 Results.

The reliability figures of the Global Expert-Ratings (table 11.4) are the Pearson's correlations between the item scores of randomly combined pairs of experts. To simplify the analyses, we combined in the Global Expert-Rating Scale both history-taking items (explanatory and treatment hypotheses) to form one item.

These reliability figures are moderate, but one has to take into account that the reliabilities are always calculated over one item; a hard condition. The reliability figures for the items "history-taking", "presenting solutions" and "structuring the interview" are reasonable. The Cronbach alpha of the whole scale is 0.79, which is good for a 7-item rating scale.

Table 11.4: **Inter-rater reliability of the Global Expert-Rating Scale by calculating Pearson's correlations between randomly combined pairs of experts.**

1. Exploration of the reason for encounter	-0.08
2. History-taking	0.32*
3. Presenting solutions	0.23*
4. Structuring the interview	0.33*
5. Interpersonal skills	0.03
6. Communicative skills	0.13
7. Global evaluation of the whole interview	0.15
Cronbach alpha of the 7-item scale	0.79

*) Correlations ≥ 0.22 are significant on the $p \leq .05$ level (N=80, 2-tailed test).

The good coefficient alpha and the moderate inter-rater reliability of this method contrasts with each other. It is proof of high method co-variance in the items, indicating that differences between the traits will not be accurately measured.

Next we turn to table 11.5 showing the correlation matrix between MAAS-PMHC scales and the Global Expert-Ratings of the same traits. In the MAAS-PMHC, we have combined the scales "history-taking" with those of "psychiatric examination" and "socio-emotional exploration" to correlate both methods better with each other.

Experts give support to the combined history-taking dimension of the MAAS-PMHC (r=.42). They also support the scales "interpersonal and communicative skills" (resp. r=.27 and r=.25), and "structuring the interview" (r=.22), but they fail to do this with the scales "exploration for the reason for encounter" and "presenting solutions". Moreover, the experts' notion of "exploration of the reason for encounter" correlates with the MAAS-PMHC trait of "interpersonal and communicative skills" (resp. .27 and .33). In contrast, the "exploration of the reason for encounter" of the MAAS-PMHC bears no correlation at all with the experts' trait of "interpersonal and communicative skills". The "exploration of the reason for encounter" of the MAAS-PMHC also shows a substantial correlation with the experts' trait of "history-taking" (.29).

Whether experts support the theoretical distinction between interpersonal and communicative skills (Hess, 1969), is studied in table 11.6. It shows that experts support the scales pertaining to the combination of interpersonal- and communicative skills as measured by the MAAS-PMHC (0.42). However, experts do not make the theoretical distinction between interpersonal and communicative skills which has been elaborated in 2.3 and which has been operationalized in the MAAS-PMHC. Their validity coefficients (resp. 0.27 and 0.25) are about the same as the correlations of communicative- and interpersonal skills measured by the different methods (0.18 and 0.29).

Table 11.5: Correlations between the scores on 6 MAAS-PMHC scales with the corresponding items of the Global Expert-Rating Scale.

		MAAS-PMHC					
		exploration of the reasons for encounter	history-taking	presenting solutions	structuring the interview	interpersonal skills	communicative skills
GLOBAL EXPERT RATING SCALE	exploration of the reasons for encounter	0.17	0.19	0.12	0.19	0.27*	0.33*
	history-taking	0.29*	0.42*	-0.03	0.04	0.16	0.04
	presenting solutions	0.05	0.22*	0.19	0.24*	0.02	0.04
	structuring the interview	0.03	0.23*	0.09	0.22*	0.26*	0.17
	interpersonal skills	-0.04	0.09	0.00	0.06	0.27*	0.18
	communicative skills	-0.03	0.20	0.09	0.07	0.29*	0.25*

*) Correlations ≥ .22 are significant ($p \leq .05$ level; N=80, 2-tailed test).

Table 11.6: Correlations between interpersonal and communicative skills and their combination measured by the MAAS-PMHC, with the same traits measured by the Global Expert-Rating Scale.

		MAAS-PMHC		
		interpersonal skills	communicative skills	interpersonal and communicative skills
GLOBAL EXPERT-RATING SCALE	interpersonal skills	0.27*	0.18	0.26*
	communicative skills	0.29*	0.25*	0.31*
	interpersonal and communicative skills	0.35*	0.39*	0.42*

*) Correlations ≥.22 are significant (p≤.05 level; N=80, 20 tailed test)

Table 11.7: Correlations of the experts' evaluation of the interview as a whole with the traits measured by the MAAS-PMHC.

		MAAS-PMHC							
		EE	HT	PE	SE	PS	ST	IPS	CS
GLOBAL EXPERT RATING SCALE	overall evaluation of the interview	.14	.31*	.17	.23*	.12	.20	.28*	.27*

*) p ≤ .05 for correlations ≥ .22
legend: see table 12.2

Table 11.8: The correlations of the experts' overall evaluation of the initial interview with their own measurements of the different traits.

		GLOBAL EXPERT-RATINGS					
		EE	HT	PS	ST	IPS	CS
GLOBAL EXPERT-RATING SCALE	overall evaluation of the interview	.63*	.36*	.48*	.53*	.52*	.67*

*) $p \leq .05$ from correlations $\geq .22$
legend: see table 12.2

The question of which characteristics of the interview (in terms of MAAS-PMHC scales) experts have in mind when they consider the initial interview as good, is answered in table 11.7.

It turns out that the experts' overall evaluation of the interview correlates significantly, though moderately, with the MAAS-PMHC scales "history-taking", "socio-emotional exploration", "interpersonal and communicative skills". No significant correlations are found with the scales "exploration of the reason for encounter", "psychiatric examination", "presenting solutions" and "structuring the interview".

The question of what which characteristics experts have in mind themselves when they consider an initial interview to be good, is addressed in table 11.8. These correlations, which are generally substantial (.36 to .67), suggest that, according to the experts' opinion, the quality of good initial interviewing in primary mental health care is mainly based on good "interpersonal and communicative skills" and on the characteristics "exploration of the reasons for encounter" and "structuring the interview".

11.3.3 Discussion.

In this section, we compare the experts' evaluations of interviewing skills with the MAAS-PMHC scales. Since this evaluation has been measured with the Global Expert-Rating Scale, we have to take its restrictions into account. Although there is a wide-spread belief in the validity of expert judgment in medical competence, we have found in our Global Expert-Rating Scale its necessary condition, i.e. its inter-rater reliability, to be moderate. In this instance, the validity of experts' judgment in interviewing should be taken with some reservedness. Even experts fail to agree on when items are ill-defined or have no clearly described criteria for rating. These findings agree with Streiner's (1985) comments on global-rating scales. Reliability and validity of global-rating scales is hampered by halo-effects, idiosyncratic use by raters and by a restriction of being able to measure no more than two dimensions (see also 4.8.2).

Reviewing the correlations between MAAS-PMHC scores and expert ratings in the light of these shortcomings, experts moderately support the MAAS-PMHC on the scale level. However, some remarks should be made.

Between experts and the constructors of the MAAS, a conceptual difference can be noted in the scale "exploration of the reasons for encounter". From our theoretical stance, this concept pertains to patient-centered information necessary to clarify the request for help i.e. the way the patient wishes to be helped to fulfil his needs in seeking professional help (Lazare et al., 1975). However, experts consider "exploration of the reason for encounter", as we operationalized it in the MAAS-PMHC, as an extension of "history-taking", aiming to collect patient-centered information. This finding is also supported by the substantial correlation between the "exploration of the reason for encounter"-dimension measured by the MAAS-PMHC and the "history-taking"-dimension of the Global Expert-Ratings.

On the other hand, the experts' own concept of the "exploration of the reason for encounter" is significantly correlated with the MAAS-PMHC operationalizations of the "interpersonal and communicative skills"; in other words, with factors enhancing an interview climate of

trust and acceptance in order to promote a mutual exchange of information. In our view, this discongruence is caused by the unclear sense of the concept "request for help" amongst experts who, however, frequently use the term, apparently to denote some of the aforementioned factors in the interview: "the request for help" means the most appropriate way the patient desires his problems to be solved or needs to be fulfilled.

In addition, it is remarkable that the theoretical, relevant distinction between "interpersonal and communicative skills" (Hess, 1969) is not made by experts, although in the Global Expert-Rating Scale this distinction has been clearly defined. It is a pity that our concepts of "presenting solutions" and "structuring the interview" are not strongly confirmed. This finding may be due to the low to moderate reliability of these MAAS-PMHC scales on the item level (see 11.2).

Finally, we find some empirical support for the generally propagated style of good initial interviewing in primary mental health care (see 2.4.2.2.3). This style which is, in principle, patient-centered, allows the patient to tell his story in his own words, while the interviewer follows in a non-directive way. This pattern should be interrupted with periods of a more structured, systematic method of questioning when hypotheses are to be tested. These periods are directive and physician-centered. In the correlations between the experts' overall evaluation of the interview with their own global ratings and with the MAAS-PMHC ratings of the different traits, we find empirical support for this statement.

When experts consider the initial interview as well-conducted, they have in mind the dimensions "history-taking" in a broad sense (physician-centered, directive) as well as "interpersonal and communicative skills" (non-directive, patient-centered).

11.4 Concluding remarks about the content validity of the MAAS-PMHC.

Content validity of the MAAS-PMHC has been investigated on item and scale levels.

Content validity of the MAAS is difficult to study. Direct investigation into whether the item domain of the MAAS-PMHC is representative for initial medical interviewing in primary mental

health care is not possible. Moreover, the quality of the operationalization of content and process of medical interviewing skills into items is almost impossible to assess.

We approximated content validity by means of a three-step procedure.

Firstly, we investigated the score profiles of residents in general practice with the MAAS-PMHC in mental health interviews. These score profiles generally support the content validity.

Secondly, we studied which "theoretical losses" the MAAS-PMHC suffered from unreliability on the item level. The reliability of items measuring interviewing skills which pertain to process aspects (interpersonal and communicative skills, the ability to structure the interview and to present solutions), is rather low. Due to this lack of reliability, the MAAS-PMHC suffers some theoretical and conceptual loss in the measurement of the process aspects of the interview. Generalizability analysis of the items pertaining to interviewing skills from each of the three phases of the interview is satisfactory. "Theoretical loss" is consequently restricted to the items of the scales "structuring the interview", "interpersonal- and communicative skills". Fortunately, reliability on the scale level is proportionally much higher, except for "exploration of the reasons for encounter", "structuring the interview" and "communicative skills". This finding suggests that observers endorse our theoretical concepts on the scale level rather than their operationalizations in items.

As a third step, we compared experts' judgements of important dimensions of medical interviewing with the scores on MAAS-scales intended to measure the same dimensions. Experts support the MAAS-PMHC on the scale level with the exception of the scales "exploration of the reasons for encounter" and "presenting solutions".

Our conceptualization of the scale "exploration of the reason for encounter" is taken for an extension of history-taking with the collection of patient-centered information, rather than for the exploration of the "request for help".

The experts' support in content validity also seems in favor of a generally propagated interviewing style, basically non-directive, interrupted by periods of more directive and systematic data-gathering in order to generate and to test hypotheses.

REFERENCES

Groot AD de. Methodologie. Mouton, 's-Gravenhage, 1961.

Hess JW. A comparison of methods for evaluating medical student skills in relating to patients. Journal of Medical Education, 1969; 44: 934-938.

Lazare A, Eisenthal S, Wasserman L. The customer approach to patienthood. Attending to patient requests in a walk-in clinic. Archives of General Psychiatry, 1975; 32: 553-558.

Mitchell SK. Interobserver agreement, reliability and generalizability of data collected in observational studies. Psychological Bulletin, 1979; 86: 376-390.

Streiner DL. Global rating scales. In: Neufeld VR, Norman GR (Eds.). Assessing clinical competence. Springer Publ. Cie., New York, 1985.

CHAPTER 12 THE CONVERGENT AND DIVERGENT VALIDITY OF THE MAASTRICHT HISTORY-TAKING AND ADVICE CHECKLIST-PRIMARY MENTAL HEALTH CARE

H.F. Kraan and A.A.M. Crijnen

12.1 Introduction.

The study of the convergent and divergent validity is governed by the question: are the theoretical concepts of medical interviewing in primary health care that underly the MAAS-PMHC really measured when we apply this method to interviewing skills? For instance: are we really measuring the physician's ability to present solutions when we apply the scale of the same name? We assume that with this scale we are measuring the physician's ability to propose a treatment plan to the patient, to make the patient responsible for his choice and to negotiate about it etc. We could have, however, measured, for instance, the quality of the treatment plan itself instead of its manner of presentation by the physician to the patient.

The ideal way to investigate this problem is to compare the measurement of this scale with a method which is known to measure the concept of "presenting solutions". Vice versa, a method not intended to measure "presenting solutions" should, of course, not measure this concept. These types of validity questions are known as convergent and divergent validity issues. For more detailed information, the reader is referred to chapter 5.

In this chapter, we study the convergent and divergent validity of the MAAS-PMHC, comparing this method with 3 different methods: two self-rating methods and one method of expert-rating.

We first briefly describe in 12.2 the methodology used to study the convergent and divergent validity: the multitrait-multimethod matrix (Campbell and Fiske, 1959). In 12.3, the multitrait-multimethod-matrix is studied by means of 4 criteria, stated by Campbell and Fiske, by which we judge the convergent and divergent validity of the MAAS-PMHC in comparison with the 3 other methods. In 12.4, the results are discussed and explained in terms of the measurement properties of the MAAS-PMHC and the three other methods. In section 12.5, we end this chapter with conclusions.

12.2 The multitrait-multimethod matrix (MTMM-matrix) in the study of the convergent and divergent validity of the MAAS-PMHC.

In chapter 8, the MTMM-matrix was defined as consisting of the correlations between multiple traits and multiple methods when each of the traits is measured with each of the methods. An instructive example of the most simple version of a MTMM-matrix is presented in table 8.1. This MTMM-matrix is used in the study of validity in order to distinguish the method from the trait variance in the scores. This distinction is a prerequisite for the interpretation of the correlation coefficients in the MTMM-matrix because these correlations always consist of a component of trait covariance and a component of method covariance. The magnitude of this latter component should be estimated in order to evaluate these correlations for their property of validity coefficients.

In the following sections, the methods (12.2.1) and the traits (12.2.2) which have been used to fill the MTMM-matrix, are described. An overview of methods and traits is given in table 12.1. How the MTMM-matrix is analyzed is discussed in 12.2.3.

12.2.1 Methods used to measure interviewing skills.

Four methods have been used: the MAAS-Primary Mental Health Care, the MAAS-SELF, the Global Self-Rating Scale (GSRS) and the Global Expert-Rating Scale (GERS). These methods, abundantly elaborated in chapter 4 and the appencices A to F, are rehearsed only briefly.

In the MAAS-PMHC, trained observers indicate which of 104 clearly-defined and discernable units of interviewing behavior occur in the course of a medical consultation. The 104 items of the MAAS-PMHC have been grouped according to the 8 theoretical dimensions ("traits") of medical interviewing skills. These 8 groups of items are the scales of the MAAS-PMHC. The scores of the items within these scales have been combined to form construct indices on these theoretical dimensions.

The MAAS-SELF was filled in by the interviewing physicians themselves at the end of the interview. On this checklist, they were asked which of the 92 items of interviewing behavior they performed in the preceding interview.

Table 12.1: Methods and traits in the measurement of physicians' interviewing skills used in the MTMM-matrix.

METHODS	TRAITS
1. MAAS-PMHC (104 items) observational measurement of the content and behavioral aspects of interviewing skills. 2. MAAS-SELF (92 items) self-rating of content and behavioral aspects of interviewing skills. 3. Global Self-Rating Scale (GSRS) (7 items) global, evaluative self-rating scale, of major dimensions of interviewing skills. 4. Global Expert-Rating Scale (GERS) (7 items) global, evaluative rating scales of interviewing skills to be rated by experts.	1. Skills to explore the reasons for encounter (EE). 2. History-taking skills (HT) which in the MAAS-PMHC and MAAS-SELF, are subdivided into: - history-taking - psychiatric examination - socio-emotional exploration In Global Self- and Global Expert Ratings Scales, history-taking skills are subdivided into: - history-taking skills in order to generate explanatory hypotheses - history-taking skills in order to generate treatment hypotheses. 3. Skills to present solutions (PS) 4. Skills to structure the interview (ST) 5. Interpersonal skills (IPS) 6. Communicative skills (CS).

The MAAS-PMHC and the MAAS-SELF have a similar item content and format, except for the scale "psychiatricexamination" where the items in the main classes of symptoms have been reformulated on a more abstract level. This change caused a reduction in the number of items. evertheless, the same 8 theoretical dimensions of medical interviewing are measured. The indices on these dimensions are constituted by combining the item scores within these 8 scales as in the MAAS-PMHC.

The Global Self-Rating Scale was rated by the interviewing physicians themselves after completion of the interview. They evaluated the quality of their interview on six theoretical dimensions and gave one global overall evaluation of the whole interview. Only broad definitions of these dimensions have been provided in the items. The items are to be rated on a 5-point Likert scale.

In the Global Expert-Rating Scale (see 11.3), a multidiscipline panel of experienced care-takers in primary mental health care rated the quality of the interviewing skills of 40 residents in general

practice on six theoretical dimensions (the same as in the previous method). In addition, they were expected to give one global overall evaluation of the whole interview.

Campbell and Fiske (1959) demanded that each of the methods used in the convergent and divergent validation procedure should be independent in order to minimize the influence of shared method variance. The MAAS-PMHC and the Global Expert-Rating Scale might share method variance, both being observation instruments. In addition, the Global Self-Rating Scale and the MAAS-Self have the aspect of self-evaluation in common. In addition, the MAAS-PMHC and MAAS-SELF have a similar format of "behaviorally" described items and the same indices formation (by summation of the item scores within scales). Finally, the Global Expert-Rating Scale and the Global Self-Rating Scale also have a negative characteristic in common: only global definitions of items and no sharp criteria for scoring have been given.

Moreover, we have not included existing methods in our study because they differ in objectives of measurement. Too many differences in underlying traits (see next paragraph) would hamper the comparison of the methods.

12.2.2 The traits measured by the four methods of measurement.

In the methodology of the multitrait-multimethod matrix, the term "trait" is not used in the sense of a psychological trait. It signals here the underlying theoretical dimension of a group of interviewing skills. These traits we extensively described in chapter 2 and 4. We here summarize the definitions and indicate how these traits are measured (see also table 12.1).

The exploration of reasons for encounter (EE) measures the physician's ability to clarify the patient's complaint, to explore the motives in the pre-patient phase leading to the visit to the physician and, finally, to gain insight into the way the patient expects his needs to be met. It is measured with MAAS-PMHC and MAAS-SELF scales of the same name and in both Global Rating Scales with the item of the same name.

History-taking (HT) enables the physician to generate explanatory hypotheses about the complaint/problem and to test these hypotheses. Furthermore, we may generate hypotheses about interventions to alter

the patient's condition. In the MAAS-PMHC and MAAS-SELF, this trait is measured by combining the scales "history-taking", "psychiatric examination" and "socio-emotional exploration".

In the Global Expert-Rating Scale and the Global Self-Rating Scale, the history-taking ability is measured by two items: one pertaining to the data that contribute to explanatory hypotheses and one pertaining to data necessary to generate hypotheses for treatment interventions. The scores on both items are taken as the trait measures in both methods.

During the presentation of solutions (PS), physicians convey information about causes and prognosis of the problem, negotiate with the patient about problem-definition and possible solutions and provide concrete information about possible treatment interventions. In the MAAS-PMHC and MAAS-SELF, they are measured by the scales of the same name and in the Global Rating Scale, by the items of the same name.

By "skills to structure the medical interview" (ST) is understood the ability to open and to terminate the interview and to pass from one phase to another in a way that is perspicuous to the patient.

Interpersonal skills (IPS) aim to establish an optimal rapport with the patient.

Communicative skills (CS) should serve to promote an effective exchange of information between patient and physician.

The last three traits are measured in the same manner as the "presenting solutions trait".

However, Campbell and Fiske (1959) demand that these "traits" ought to be independent and therefore have near zero to low correlations. In our situation, these ideal circumstances are not met. Different traits sometimes have process aspects in common such as questioning, conveying information etc. Sometimes there is also an overlap in the content aspects: for instance, between questioning about causal conditions of the mental health problems in the scales "exploration of the reasons for encounter", "history-taking" or "socio-emotional exploration". On theoretical grounds, some correlative relationship between the traits may thus be expected.

12.3 Constructing and analyzing the multitrait-multimethod matrix.

In principle, the MTMM-matrix presented in table 12.2 was

Table 12.2: Multitrait-multimethod matrix, consisting of the crossed (partial) correlations of 6 traits of interviewing skills, each measured by 4 methods.

		MAAS-PMHC						MAAS-SELF					
		EE	HT	PS	ST	IPS	CS	EE	HT	PS	ST	IPS	CS
MAAS PMHC	EE												
	HT	.34*				.25						.01	
	PS	.34*	.25*							.16			
	ST	.14	.38*	.37*									
	IPS	.01	.27*	.11	.34*				.06				
	CS	.05	.21*	.04	.27*	.55*							
MAAS SELF	EE	.23*	.30*	-.01	.12	.21*	-.13						
	HT	.03	.38*	-.10	.08	.26*	-.07	.66*				.38	
	PS	.15	.25*	.29*	.09	.12	-.07	.44*	.43*				
	ST	.01	.17	-.09	.14	.10	.00	.32*	.30*	.31*			
	IPS	.13	.08	-.07	.07	.06	-.07	.34*	.42*	.23*	.38*		
	CS	-.03	.14	-.20*	.10	.04	-.17	.42*	.42*	.28*	.43*	.26*	
GLOB SELF-RATING	EE	-.08	.07	-.14	.10	.05	-.02	.29*	.37*	.21*	.47*	.49*	.42*
	HT	.16	.17	-.02	.26*	.24*	.16	.30*	.42*	.36*	.37*	.36*	.36*
	PS	.10	.09	.16	.15	.09	-.10	.23*	.37*	.31*	.05	.18	.25*
	ST	.06	.05	-.08	.06	.12	-.09	.25*	.33*	.19*	.24*	.39*	.38*
	IPS	-.10	.09	-.13	.18	.10	-.08	.26*	.32*	.10	.24*	.19*	.42*
	CS	-.00	.12	-.01	.29*	.16	-.15	.23*	.32*	.13	.34*	.23*	.40*
GLOB EXPERT-RATING	EE	.15	.17	.11	.20*	.28*	.36*	-.01	.07	.02	.04	.03	.16
	HT	.24*	.37*	-.08	.06	.18	.09	.27*	.31*	.19*	.20*	.17	.28*
	PS	.07	.26*	.20*	.24*	.02	.04	.07	.13	.21*	.20*	.07	.19*
	ST	.08	.32*	.11	.21*	.25*	.14	.03	.04	.14	.20*	.14	.16
	IPS	-.03	.12	.01	.05	.26*	.18	.12	.24*	.08	.02	.25*	.13
	CS	-.01	.25*	.11	.07	.29*	.24*	.00	.10	.13	-.08	.09	.08

Legend:
EE = exploration of the reason for encounter
HT = history-taking
PS = presenting solutions
ST = structuring the interview
IPS = interpersonal skills
CS = communicative skills

		GLOB SELF-RATING						GLOB EXPERT-RATING					
		EE	HT	PS	ST	IPS	CS	EE	HT	PS	ST	IPS	CS
MAAS PMHC	EE												
	HT					.07						.13	
	PS				.04						.24		
	ST												
	IPS		.05						.15				
	CS												
MAAS SELF	EE												
	HT					.26						.09	
	PS				.31						.17		
	ST												
	IPS		.30						.14				
	CS												
GLOB SELF-RATING	EE												
	HT	.44				.32						.08	
	PS	.37	.32								.06		
	ST	.38	.32	.14					.07				
	IPS	.28	.34	.10	.29								
	CS	.41	.43	.27	.24	.40							
GLOB EXPERT-RATING	EE	-.06	.13	.07	.14	.00	.08						
	HT	-.01	.23	.05	.19	.15	.13	.49				.39	
	PS	-.01	.06	.06	.10	-.05	.11	.45	.48				
	ST	.01	.17	.12	.06	-.09	.10	.37	.43	.51			
	IPS	.10	.07	.35	-.01	-.01	-.06	.32	.20	.24	.25		
	CS	-.03	.15	.12	.18	-.00	.09	.56	.32	.37	.52	.39	

Correlations of .19 or higher are significant ($p \leq .05$; two-tailed test; N=80).

constructed as a completely crossed intercorrelation of the scores of the 6 traits of interviewing skills, each trait measured by the 4 methods. The scale scores were taken from the performance of 40 residents in general practice interviewing the two simulated mental patients.

However, in order to improve the reliability of these scale scores, the following arrangements have been made:

- The inter-rater reliability has been improved by adding "live" MAAS-scale scores obtained during the simulated consultation hour and MAAS-scale scores by observers rating the videotaped interviews several months later. As a result, we have at our disposal a summated set of MAAS-scores from two independent, well-trained observers of each case. We used the Rasch homogeneous scales of the MAAS-PMHC in these validity studies, also taking advantage of their slightly better reliability figures.
  A similar summation to improve reliability was carried out with the Global Expert-Rating for which there are also two sets of scores (each by an independent observer) per case available.
- The inter-case reliability has been enhanced by partializing the "case-influence" from the correlation matrix. Partialization has been carried out by including "case influence" in the original correlation matrix as a dummy variable, taking the value 1 for case 1 (depression) and value 2 for case 2 (anxiety). The correlations with this dummy variable are subtracted from the original matrix. The resulting first order correlations are used in the MTMM-matrix (Guilford et al., 1982, 6th ed.). After this partialization, it is permitted to add the scores of both cases.

Before calculating the matrix, the missing data (less than 0,5% on the item level) in the MAAS-SELF and the Global Self-Rating Scale have been estimated by regressing items with missing values (as dependent variables) on the items (as independent variables) with which they were most highly correlated (BMDP-program: PAM-single). In the scores of the Global Expert-Rating Scale, no missing data have been detected.

The multitrait-multimethod matrix has been built up by intercorrelations of the trait scores obtained by each of these four

methods. The resulting matrix counts 24 (6 [traits] x 4 [methods]) intercorrelated variables, yielding 288 (24x24/2) correlations coefficients.

To provide a better overview of the matrix, the means of the (validity) diagonals and of the triangles have been written down on the right-hand side of the grand diagonal, the axis of symmetry. These means have been calculated after transformation of the correlations into Fisher Z-scores.

On studying the matrix, correlations of .19 or higher are taken as significant ($p \leq .05$; two tailed test; N=80). This means that approximately 14 of the 288 correlations are significant by chance.

This study of a MTMM-matrix proceeds along the lines of the four criteria proposed by Campbell and Fiske (Campbell and Fiske 1959; Schmitt and Stults, 1986) which we have already been extensively described in chapter 8. We repeat them in the following section.

If the investigated method (MAAS-PMHC) meets these four criteria, then a perfect convergent and divergent validity of this method has been attained.

12.4 Results and discussion.

This section describes the study of the MTMM-matrix by means of the 4 criteria of Campbell and Fiske (12.4.1). In addition, a further part of this section (12.4.2) is devoted to the phenomenon of method variance, a source of systematic variance in measures of interviewing skills whatever method is used. This method variance, often caused by halo-effects, may act as a considerable confounder in measurement.

12.4.1 First criterion: convergent validity.

According to this criterion, to judge the MTMM-matrix the number of significant correlations on the validity diagonals are counted. These validity coefficients are the correlations between similar traits measured by different methods. When our four methods are compared with each other, each method has 18 validity coefficients in common with the three other methods. The significant coefficients are counted for each method (see table 12.3).

Table 12.3: Number of significant correlations ($p \leq 0.05$, 2 tail test) on the validity diagonal between each method and the three other methods (maximum = 18).

MAAS-PMHC:	8
MAAS-SELF:	13
Glob. Self-Rating:	7
Glob. Exp.-Rating:	10

Fifty three percent of the validity coefficients are statistically significant. Moreover, 3 or 4 correlations of the possible maximum of 72 may be significant by chance ($p \leq .05$; two tailed test). A closer view of these moderate findings reveals support for the MAAS-PMHC by the Global Expert Ratings, except for the "exploration of the reasons for encounter".

These findings have already been reported in the previous chapter in which we also concluded that experts do not support the theoretical content of the scales "exploration of the reason for encounter" and "presenting solutions" because of differences in theoretical orientation in the former and because of lack of reliability on the item level in the latter.

Furthermore, we note a lack of support from the Global Self-Ratings. This fact is probably due to the deficient reliability and validity of global-rating scales which has been universally noted in the measurement of other domains of medical competence (a.o. Streiner, 1985). The lack of validity is hardly due to the self-evaluation aspect in this method because the MAAS-SELF supports the validity of the MAAS-PMHC at least in the scales "exploration of the reason for encounter", "history-taking" and "presenting solutions". It seems evident that interviewers are better able to evaluate their own technical, problem-solving aspects in the three phases of initial interviews than they are the process aspects such as interpersonal and communicative skills.

Criterion 2: Divergent validity with regard to the traits.

This criterion requires the entries on the validity diagonal to be higher than the heterotrait-heteromethod values in the column and row in which a certain validity coefficient is located.

The entries on the validity diagonal mainly consist of covariance of the same trait measured by two different methods. In the heterotrait-heteromethod triangles, the correlations are mainly built up from covariance arising from pairs of different traits measured by these two different methods. It is thus implied by this second criterion that covariance of the same trait measured by different methods is compared with the covariance of one of these traits with a different trait. This comparison is a test of whether each trait can be discerned from another, the measurement methods being kept in control.

To investigate this criterion, the number of times an entry on the validity diagonal is higher than the entries in the related row and column in the two adjoining heterotrait-heteromethod triangle is counted. Each entry on the validity diagonal is therefore compared with two other values as is shown in table 12.4.

Table 12.4: **Number of times where the entries on the validity-diagonal are higher than the heterotrait-heteromethod-values in corresponding column and row (maximum per cell = 10; maximum per row or column = 60).**

pairs of methods / traits	MAAS/ MAAS-Self	MAAS/ Glob Self-Rating	MAAS/ Glob Expert-Rating	MAAS-Self/ Glob Self-Rating	MAAS-Self/ Glob Expert-Rating	Glob Expert-Rating/ Glob Self	Total
EE	9	2	5	5	0	0	21/60
HT	10	8	10	10	10	10	58/60
PS	10	10	8	8	10	4	50/60
ST	9	4	7	3	10	3	36/60
IPS	3	6	8	2	10	4	33/60
EFC	1	0	7	8	2	4	22/60
TOTAL	42/60	30/60	45/60	36/60	42/60	25/60	

Legend: see table 12.2.

The column totals show the "capacity" of a certain combination of methods to discriminate the six traits. The row totals show how each trait "allows itself" to be discriminated by the six possible method combinations.

These marginals of rows and columns are expressed as ratios between obtained and maximum possible scores (60) providing a measure of discriminating capability. In general, it can be stated that the marginals of the columns are a "measure of quality" of a certain combination of methods, whereas the marginals of the rows are a measure of the divergent validity of a certain trait.

In table 12.4, the sum scores of the rows reveal that the "history-taking" and "presenting solutions" traits show strong evidence of divergent validity, whereas the traits pertaining to "structuring the interview" and to "interpersonal skills" only have a moderate divergent validity. Their validity coefficients contrast substantially with the comparable correlations in the heterotrait-heteromethod triangles, irrespective of the methods used. The traits of "exploration of the reasons for encounter" and of "communicative skills" again show insufficient evidence of divergent validity.

On looking at the marginals of the column of table 12.4, it turns out that combinations of the MAAS-PMHC, MAAS-SELF and the Global Expert-Rating exhibit the best ability to discriminate the six traits. The relatively great ability of the MAAS-SELF and the low ability of the Global Self-Ratings as methods to discriminate the traits are noteworthy.

In this respect, the Global Self-Ratings suffer, apart from error variance, from two other sources of unreliability due to the characteristics of global-rating scales (see 4.8.1): high method variance and poor operationalization of underlying theoretical concepts (content validity). Under the heading of criterion 3 and 4, we discuss this subject in greater depth.

Criterion 3: Divergent validity with regard to the methods.

According to this criterion, the validity coefficients must be higher than the off-diagonal correlations in their monomethod triangle.

Measurement of similar traits obtained by different methods should intercorrelate higher than measures of different traits obtained by the same method. When the validity coefficients are lower than the off-diagonal correlations in their monomethod triangle, there is evidence that the traits are highly intercorrelated or/and there is a high

(confounding) method variance (Campbell and Fiske, 1959; Schmitt and Stults, 1986).

To study this criterion, we counted the number of times that the three validity coefficients for each trait are higher than the mean of the heterotrait-monomethod triangles (see table 12.5).

Table 12.5: The number of times that three validity values of each trait are higher than the mean of a heterotrait-monomethod triangle (maximum per cell = 3; maximum per row = 12; maximum per column = 18).

Method	MAAS-PMHC	MAAS-SELF	Glob Self	Glob Exp	Total
Trait					
EE	0	0	0	0	0/12
HT	2	2	1	1	6/12
PS	1	0	0	0	1/12
ST	0	0	0	0	0/12
IPS	1	0	0	0	1/12
CS	0	1	1	0	2/12
Total	4/18	3/18	2/18	0/18	

Legend: see table 12.2.

For each trait, three validity coefficients arise when the measurement of this trait by one method is correlated with the measurement of this trait obtained by the three other methods.

Looking at the column marginals, which can be considered as a measurement of the divergent validity of a certain method, we notice that the MAAS-PMHC meets this criterion 22.2% of the time, the MAAS-SELF about 16.7% of the time, the Global Expert-Rating Scale 11.1 of the times and the Global Self Rating-Scale, never. The row marginals can be taken as measurement of the divergent validity of each trait. These row marginals reveal that "history-taking skills" meets this criterion 50% of the time, whereas the other traits vary from zero to twice meeting this criterion.

These results, which seem rather modest for the MAAS-PMHC, have to be judged in the light of some critical methodological remarks as to this third criterion. In the application of this criterion, Campbell and Fiske have aimed to compare method variance with trait variance, considering the validity coefficient as common trait variance and the values in the heterotrait-monomethod triangle as results of method variance.
This statement does not entirely hold true in our situation.

First, in the discussion of the second criterion, we noticed the confounding of the validity coefficients with shared variance between the two methods compared. This method co-variance is highly probable in the combination of MAAS-SELF and Global Self-Ratings because of the comparability of both self-evaluation methods.

Second, the off-diagonal correlations in the monomethod triangles, which might be indicative for the method variance, may be confounded and inflated by existing trait intercorrelations. It is highly improbable that the traits are independent of each other because of their fitting in one unidimensional Rasch scale (see 10.2.4.).

To summarize: this criterion is difficult to meet because the method variance, which is already fairly high, has, in addition, been inflated with both afore-mentioned influences.

As stated in chapter 8, Fiske (1971) approached this criterion with greater subtlety, accepting method influences as an inseparable aspect of measurement. In our opinion, measurements are not invalidated when the correlations in the heterotrait-monomethod triangles exceed the corresponding validity diagonals. Researchers, however, must be aware that a substantial degree of variance in their measurements is attributable to the measurement method. In such situations, additional studies, such as the generalizability analyses, are needed to determine the magnitude of method variance components.

Criterion 4: The patterns of trait inter-relationships should be the same in all heterotrait triangles.

Following the recommendations made by certain authors (Schmitt e.a., 1986) advocating the use of factor analysis in the study of the MTMM-matrix, we studied this criterion in a two-fold factorial design.
First, the scores on the 6 traits are factor analyzed for each of the

four methods. Second, the six hetero-method blocks of the MTMM-matrix are factor analyzed. In these blocks, all combinations of the two sets of 6 traits, each measured with two methods, are factor analyzed.

We now restate this fourth criterion into two hypotheses, which should be confirmed after inspection of both sets of factor structures:

a) The factor structure of the 6 traits found within each of the four measurement methods should be similar when we measure the same interviews each of these methods.
b) The factor structure of the traits in the six hetero-method blocks of the MTMM-matrix should show such a picture that similar traits, measured by a different method, load on the same factor. It goes without saying that we measure the same interviews with each method. An illustration makes thus clearer: take, for instance, the block where the six traits are measured with the Global Expert-Rating and with the Global Self-Rating. When an important factor is found with high loadings of interpersonal and communicative skills measured by the expert ratings, the factor loadings of both traits measured by the self-rating should also be considerable on this same factor. The MTMM-matrix should be similar when we measure the same interviews with each of our four methods.

The first hypothesis has been tested by factor analyzing the trait scores of the MAAS-Self (N=80), Global Self-Ratings (N=80) and of the Global Expert-Ratings (N=160). In the standard program of SPSS (Nie et al., 1975), principal component analysis and then Varimax rotation of the factors with an eigenvalue exceeding 1 was carried out.

In table 12.6, we present the factor structures obtained by measuring 6 traits with each of the four methods. Factors with an eigenvalue of more than 1 and factor loadings higher than 0.6 are indicated. The factor loadings are indicated in order of decreasing magnitude.

Table 12.6: Overview of the factor structures obtained by measuring the 6 traits underlying interviewing with four methods of measurement (eigenvalue factors ≥.1; factor loadings ≥.6).

Factors	MAAS-PMHC	MAAS-Self	Global Self-Ratings	Global Expert-Ratings
first factor	IPS CS	EE HT CS	HT EE CS	CS EE PS ST
second factor	EE HT PS			

Legend: see table 12.2.

In the factor structure of the MAAS-PMHC, the traits "interpersonal and communicative skills" are most prominent in the main factor, whereas the traits "exploration of the reason for encounter", "history-taking" and "presenting solutions" load on the second factor. This means that in the MAAS-PMHC, measurement of the process aspects of interviewing predominate the measurement of the content of the interview.

The MAAS-SELF shows a strong method variance, accounted for by the single factor found in principal component analysis (Saal et al., 1980; Dielman et al., 1980). This factor shows a pattern different from the factor structure of the MAAS-PMHC. Measurement of patient- and physician-centered information collection is the most striking measurement property as witnessed by high factor loadings on the traits of "exploration of the reasons for encounter" and "history-taking".

The factor structure of the Global Self-Ratings takes a middleof-the-road position between the patterns of the MAAS-PMHC and the MAAS-SELF. Principal component analysis also yields one factor (method variance) that shows a mixed pattern of loadings on the traits

"exploration of the reason for encounter" and "history-taking", and, to a lesser extent, of "communicative skills". This pattern seems to represent a measurement property of this method, stressing the effective collection of patient- and physician-centered data pertaining to the presented mental health problem.

The factor structure of the Global Expert-Ratings reveals a completely different picture. The single factor found evidences a considerable method variance when we find a pattern of factor loadings different from the three other methods: the traits "communicative skills", "exploration for the reason for encounter", "presenting solutions" and "structuring the interview" load on this factor.

The interpretation of this factor structure may indicate the effective collection of patient-centered data by the physician. This interpretation is comparable with that of the factor structure of the MAAS-PMHC.

These findings permit the following conclusions to be drawn. First, according to the factor structures and patterns of factor loadings, the trait inter-relationships found in the four measurement methods show patterns that are different from each other. However, the MAAS-SELF and Global Self-Rating on the one hand, and MAAS-PMHC and Global Expert-Rating on the other hand, have some similarities in their pictures.

This finding implies that our first hypothesis, relating to similarities in the interrelationships of traits in the heterotrait-heteromethod triangle, is not confirmed.

The second hypothesis is studied by factor analyzing the six heteromethod blocks of the MTMM-matrix (see table 12.2). The standard program of SPSS as used in testing the previous hypothesis has again been used.

When we examine the resulting factor structures with their major factor loadings, then it is evident that our second hypothesis is not confirmed. In all six heteromethod blocks, the same patterns arise: the traits loading on one factor are clusters of different traits measured by one method instead of pairs of the same traits measured by two different methods. We take again our previously mentioned example of the heteromethod block where the six traits are measured with the Global Expert- and with the Global Self-Rating Scales (see table 12.7).

Table 12.7: Factor structures with their major (≥0.6) loadings on six interviewing traits obtained by factor analyzing the heteromethod block of the Global Self-Rating and the Global Expert-Rating scales in the MTMM-matrix.

	Factor structures with their major loadings on six interviewing traits.		
Combination of methods (hetero-method blocks)	Factor I	Factor II	Factor III
Global Self-Rating scale	-	EE	-
	-	HT	-
	-	-	PS
	-	-	-
	-	IPS	-
	-	CS	-
Global Expert-Rating Scale	EE	-	-
	HT	-	-
	PS	-	-
	-	-	-
	-	-	IPS
	CS	-	-

Legend: see table 12.2

We would expect factors with substantial loadings consisting of paired traits. For example, in factor I: EE, HT, CS, EE, HT, CS. We see, however, that EE, HT, CS (measured by the Global Expert-Rating Scale) and EE, HT, CS (measured by the Global Self-Rating Scale) load on two different factors in stead of one. In all six heteromethod blocks, a similar pattern is notable.

This finding is due to the high method variance in our methods, resulting in the clusters of different traits measured by the same method in the factors we obtained. The next section is devoted to this important issue of method variance.

Nevertheless, we have to draw the conclusion that we cannot meet this fourth criterion for convergent and divergent validity.

12.4.2 Influence of method variance on the convergent and divergent validity study of the MAAS-PMHC.

Method variance is not of mere theoretical importance, but can be attributed to effects that are very common in measurement practice: leniency and halo-effects, the latter being the most significant. Halo-effects are consistently conceptualized as an observer's failure to discriminate among conceptually distinct aspects ofsubject's behavior. Halo-effects are larger when variables have a moral connotation (such as our items pertaining to "interpersonal and communicative skills") or when single variables are not easily observed and/or are ill-defined (Streiner, 1985, citing Alport, 1937). Observers seem to rate an overall impression concerning the subjects under research conditions that are susceptible to halo-effects. As a result, observers are barely able to assess more than one or two dimensions accurately and all items are consequently associated with each other (Thorndike,1920; Guilford, 1954; Saal et al., 1980). Therefore, global rating scales are particularly susceptible to halo-effects (Streiner, 1985).

The problem is to assess the magnitude of method variance in our four instruments. Our design does not allow an exact assessment, but we study in greater depth the manifestations of method variance which we have already encountered in this chapter.

First, in the previous section we have cited Saal et al. (1980) and Dielman et al. (1980) who state that the findings of one single factor in principal component factor analysis suggest a high method variance. Principal components analysis of the trait scores of MAAS-SELF, Global Self-Ratings and Global Expert-Ratings all yield one factor, with an eigenvalue higher than one (cf table 12.6). By contrast in the MAAS-PMHC, two factors arise after principal component analysis of the traits.

Second, the height of the averaged correlations between the traits in the monomethod triangles is an indication of the amount of method variance. This evidence is even strong when it is assumed that the intercorrelation of theoretically different traits is zero to low. In the MTMM-matrix, the averaged correlations between the traits in the monomethod triangles is considerable in the MAAS-Self, Global Self-Ratings and Global Expert-Ratings (resp. 0.38, 0.32, 0.39), whereas the

averaged intercorrelations between the traits in the corresponding triangle of the MAAS-PMHC are lower (0.25).

Third, we notice in the factors of the heteromethod blocks a clustering of different traits measured with the same method, instead of clustering of similar traits measured with two different methods. This indicates that the clustering of different traits in one factor is due to the method variance, which these trait measures share.

These findings suggest considerable method variance mainly caused by halo-effects in the methods but least present in the MAAS-PMHC. This characteristic of the MAAS-PMHC is the consequence of the constructors' efforts to define and to operationalize the interviewing skills behaviorally by expressing criteria for scoring in single or multiple behavioral acts. Nevertheless, this conclusion is rather contradicted by the considerable method variance in the MAAS-SELF, especially when we consider that the MAAS-PMHC and the MAAS-SELF have almost similar items. We explain this higher method variance of the MAAS-SELF by the time lag of about 10-20 minutes between the interview and the rating of the checklist, causing a deficient recall of the performed interviewing behavior and leading to the induction of halo-effects. This deficient recall may cause an impaired discrimination between the questions asked by the physician-self and the topics spontaneously raised by the patient.

Global Self- and Global Expert-Ratings suffer from method-variance because global rating scales are vulnerable to halo-effects (see 4.8.1). It is striking, however, that the Global Expert-Ratings shows the highest halo-effect (witnessed by the correlations in the corresponding monomethod triangles). These findings are rather disappointing as we expected the experts to be very familiar with the concepts of interviewing in primary mental health care because of their own teaching and health-care experience in this domain.

An important consequence of high method variances may be their confounding effects in validity research. When methods are compared that share a high method variance of the same type, their intercorrelations may be too high, spuriously boosting convergent validity and attenuating divergent validity. In the case of high method variance which is not shared in the compared measurement methods, a

reversed picture of inflated divergent validity and repressed convergent validity may ensue.

Turning to our four methods, it is plausible on theoretical grounds that some methods will share method variance. MAAS-PMHC and MAAS-SELF show a similarity in item number and format, whereas MAAS-SELF and Global Self-Ratings have the self-evaluation aspect in common. Global Expert-Ratings and Global Self-Ratings have a comparative item format and are both rating scales. This suggests the probability of method co-variance artificially heightening the validity coefficients in the above-mentioned combinations of methods. The amount of method co-variance is difficult to assess exactly, but the averaged correlations in the common heterotrait-heteromethod may serve as an indication. These are low for the combinations MAAS-PMHC/MAAS-SELF (0.04) and Global Expert- Rating/Global Self-Rating (0.08), but considerable for MAAS-SELF/Global Self-Rating (0.30).

On combining these three figures, we may conclude that the self-evaluation dimension which the MAAS-SELF and the Global Self-Rating have in common explains the high method co-variance. This conclusion makes an inflation of the convergent validity between both methods probable.

It is surprising that method co-variance between the Global Expert-Ratings and Global Self-Ratings seems to be low, not withstanding their high method variance. Both method variances apparently are of a different nature. The consequence may be an inflated divergent and a deflated convergent validity between both methods. The findings in table 12.4 and 12.5 support this hypothesis.

12.5 Summary and concluding remarks about the convergent and divergent validity of the MAAS-PMHC and related methods.

The convergent and divergent validity of the MAAS-Primary Mental Health Care has been studied by means of the multitrait-multimethod matrix. In this matrix, six traits underlying the interviewing skills necessary in primary mental health care have been measured simultaneously by four methods.

The six traits measured by scales of the MAAS-PMHC are:

- exploration of the reasons for encounter
- history-taking (including psychiatric examination and socio-emotional exploration)
- presenting solutions
- structuring the interview
- interpersonal skills
- communicative skills

The four methods involved are the MAAS-PMHC, a method to observe 104 single and complex interviewing skills: the MAAS-SELF, a self-rating method for 92 single and complex interviewing skills: the Global Self-Rating Scale, Likert-type, self-rating scales for medical interviewing; the Global Expert-Rating Scale, Likert-type evaluative scales for 6 medical interviewing to be rated by experts. The items in these methods are all operationalizations of the 6 previously-mentioned traits.

The scores used in this study were obtained from an experiment in which 40 residents in general practice each interviewed two simulated patients, one with major depression and one with panic disorder. In order to remove "case influence" from correlations of the matrix, this effect has been partialized out by means of a dummy variable representing "case influence". The resulting first order correlations have been taken as MTMM-matrix for study.

To study the convergent and discriminant validity, four criteria developed by Campbell and Fiske (1959) have been applied to the matrix.

Convergent validity of the MAAS-PMHC has been reasonably supported by the MAAS-SELF and the Global Expert-Rating Scale. All six traits of the MAAS-PMHC have been supported by the Global Expert-Ratings except for the "exploration of the reasons for encounter" and "presenting solutions". Surprisingly, these traits, in combination with "history-taking", are well corrobated by the MAAS-SELF. This self-evaluation variant of the MAAS shows a better convergent validity regarding the "history-taking" ("problem-solving skills") than regarding the "interpersonal and communicative skills". The Global Self-Rating Scale does not contribute to convergent validity because of the insufficient opertionalization of theoretical concepts in items leaving too much room for substantial halo-effects.

Divergent validity regarding the traits (the second criterion of Campbell and Fiske) is reasonably met: "history-taking", "presenting solutions" and, to a lesser extent, "structuring the interview" and "interpersonal skills", are distinguishable traits in all the methods except for the Global Self-Rating Scale. The validity here of the trait "exploration of the reasons for encounter" is also questionable. From chapter 11, we already know that this phenomenon is due to confusion about the underlying theoretical concepts of this scale.
It is striking that the MAAS-SELF has a slightly better ability to distinguish this trait than the MAAS-PMHC and the Global Expert-Ratings. The Global Self-Rating Scale performs poorly in discriminating the traits because of its insufficient measurement properties.

Divergent validity in terms of separating trait from method variance (the third criterion), is best with the MAAS-PMHC. Method variance is discriminated from trait variance in the traits underlying "history-taking" and, to a lesser extent, but still surprisingly high, in the "communicative skills". Nevertheless, the "performance" of all four methods on this criterion in an absolute sense is moderate (MAAS-PMHC) to poor.

The fourth criterion of divergent and convergent validity claims a similarity in the patterns of trait inter-relationships in all heterotrait triangles. This criterion has not been met because of two reasons. First, factor analysis reveals that the traits of interviewing skills measured by the four different methods each shows a different pattern of intercorrelation. Secondly, the high method variance, mainly due to halo-effects, has a disturbing influence. This method variance is the lowest in the MAAS-PMHC becasue of its behaviorally defined items.

The MAAS-PMHC turns out to be comparatively the best measure of physicians' interviewing skills in Primary Mental Health Care, evidenced by reasonable convergent and divergent validity and by a relatively small method variance.

The factor analyses of the scores on the interviews with the four methods also give more insight into their measurement properties. In the MAAS-PMHC, a factor pattern arises which indicates that its measurement properties stress, in particular, the "interpersonal and

communicative skills" and, to a lesser extent, the content aspects of the three phases of the initial interview: exploration of the reason for encounter, history-taking and presenting solutions. This pattern suggests that it assesses good medical interviewing as a balanced combination of process and content aspects. In both self-evaluation methods, the MAAS-SELF and the Global Self-Ratings, the factor patterns resemble each other. They reveal an accent on the measurement of the accurate collection of patient- and physician-centered data. The factor pattern of the Global Expert-Ratings presents a quite different picture, stressing the "communicative skills" in gathering patient-centered information and in "presenting solutions". It is clear that each method stresses its own priorities in measurement.

REFERENCES

Alport GW. Personality: a psychological interpretation. Holt Co., New York, 1937.

Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol. Bull., 1959; 56: 81-105.

Dielman TE, Hull A, Davis WK. Psychometric properties of clinical performance ratings. Evaluation and the Health Professions, 1980; 3: 103-117.

Fiske D. Measuring the concepts of personality. Chicago, 1971.

Guilford JP. Psychometric methods. McGraw-Hill, New York, 1954 (2nd ed.).

Guilford JP, Fruchter B. Fundamental statistics in psychology and education. McGraw-Hill, Auckland, 1982 (6th ed.).

Nie NH, Hadlai Hull C, Jenkins JG, Steinbrenner K, Bent DH. Statistical package for the social sciences (SPSS). McGraw-Hill, New York, 1975 (2nd ed.).

Saal FE, Downey FG, Lahey NA. Rating the ratings: assessing the psychometric quality of rating data. Psychol. Bull., 1980; 88: 413-428.

Schmitt N, Stults DM. Mathodology review: analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 1986; 10: 1-22.

Streiner DL. Global rating scales. In: Neufeld VR, Norman GR (Eds.). Assessing clinical competence. Springer Publ., New York, 1985.

Thorndike EL. A constant error in psychological ratings. J. Appl. Psychol., 1920; 4: 25-29.

CHAPTER 13 CONSTRUCT VALIDITY STUDIES WITH THE MAAS-PMHC

H.F. Kraan and A.A.M. Crijnen

13.1 Introduction.

In general, the construct validity of a measurement method is supported when theoretical constructs underlying the method can be confirmed empirically. In this respect, the construct validity of the MAAS-PMHC is supported when this method can measure interviewing skills that contribute to the achievement of the three main functions of the medical interview in primary mental health care (Schouten, 1982). We restate once more these functions which have already been discussed in chapter 1:

a. collection of information from the patient necessary for diagnostics and clinical problem-solving;
b. conveyance of information in order to inform the patient and to enhance his insight and compliance;
c. establishment and maintenance of a physician-patient relationship of trust and acceptance for the achievement of both previous functions.

In this chapter, an exploratory correlative study is carried out to investigate how the physician's interviewing skills as measured with the MAAS-PMHC are related to the achievement of these functions.

From these functions of the medical interview, we derive three research questions:

1) Are there significant validity coefficients between the physician's interviewing skills measured with MAAS-PMHC variables and the quality of diagnosis and clinical problem-solving?
2) Are there significant validity coefficients between the physician's interviewing skills measured with MAAS-PMHC variables and the degree of insight the patient has into his problems and his intention to comply?

3) Are there significant validity coefficients between the physician's interviewing skills measured with MAAS-PMHC variables and the quality of the physician-patient relationship as experienced by the patient?

In the literature, only one study is found reporting construct validity research that relates measured interviewing skills with outcome variables (Inui et al., 1982; Carter et al., 1982). The authors compare three interaction analysis instruments of medical interviewing (Bales, 1950; Roter, 1977; Stiles, 1978; see chapter 4) to assess their capability for prediction of variance in several outcome measures: insight of patients in their problems, patients' compliance and satisfaction. In general, 35-40% of the variance in these outcome variables can be predicted from interviewing skills. Later in this chapter, we refer to some findings of this study.

The first research question is studied by determining validity coefficients between measurements of diagnostics and clinical problem-solving and interviewing skills as measured with the MAAS-PMHC. The relationship between these two aspects of the physician's competency is studied in 13.2.

The second and third research questions are investigated by determining validity coefficients between measurements of the MAAS-PMHC variables and variables measuring the effect of the medical interview on the patient. These patient variables are the recall of information conveyed by the physician; the degree to which the patient feels facilitated to tell his story; the degree to which the patient feels disrupted in the communication by the physician; the degree to which the patient feels directed towards ideas and solutions of the physician; his insight into his own problems and his intention to comply with medical advice. These research questions are studied in 13.3.

Both validity studies were carried out during the simulated consultation hour which has been elaborated extensively in chapter 8, but which we briefly recapitulate in 13.2.1.

13.2 Interviewing skills and the quality of the diagnostic and clinical problem-solving process.

The importance of the medical interview for diagnostics and, in a broader sense, for clinical problem-solving, has repeatedly been reported. Strong associations between the history-taking sections of the initial interview and clinical problem-solving have been established in somatic problems (Elstein et al., 1978; Kassirer et al., 1978) as well as for minor psychiatric disorders (Goldberg et al., 1980), as we have elaborated in chapter 2.

We expect significant validity coefficients between the physician's scores on the MAAS-PMHC scales "history-taking", "psychiatric examination" and "socio-emotional exploration" with measurements of diagnostics and clinical problem-solving. High scores on these MAAS-PMHC scales tend to be indicative of the quality of data collection for hypotheses generation and testing.

Furthermore, it is expected that the physician's scores on the scale "communicative skills" also has a significant validity coefficient with measures of diagnostics and clinical problem-solving. The physician's communicative skills are to establish an effective exchange of data between physician and patient. Consequently, physician and patient will be mutually aware of the meaning that one attaches to the messages sent by the other (Schouten, 1982).

13.2.1 Method.

13.2.1.1 Research setting.

Forty residents in general practice each interviewed two simulated patients, the first representing a major depression, the second a panic disorder. In this simulated consultation hour, each role was played by two patients who were randomly assigned to the residents. The interviews were rated "live" by trained observers and were videotaped. After the interview, the residents filled in a questionnaire measuring diagnostics and clinical problem-solving (see below), whereas the simulated patients filled in the Patient Satisfaction with Communication Checklist and responded to 6 open questions pertaining to the information conveyed by the physician (see chapter 6 and 13.3, below). After 4 months, the videotaped interviews were rescored by the same pool of trained observers.

13.2.1.2 Instruments.

The residents' interviewing skills were rated twice with the MAAS-PMHC; the first time, live and the second time, from the videotaped interviews to which the same pool of observers were randomly assigned. In this study, both sets of scores were summated to improve the reliability. The summated scores of the items per scale were used as indices for the residents' interviewing skills.

Diagnostics and clinical problem-solving were measured by an instrument called "Problem-Solving in Primary Mental Health Care" (PS-PMHC), see appendix H.

On the PS-PMHC, a semi-structured questionnaire, the physician has to respond to open-ended questions about differential diagnosis, explanatory hypotheses and hypotheses about further management. Scoring takes place by comparing the physicians' responses on the questions with criteria predefined by the same panel of experts who also designed the roles of both simulated patients. Two raters scored the PS-PMHC according to the criteria and attained high reliability (Kappa=0.88, Cohen, 1960). According to a calculation rule with preset weight factors, the correct answers have been summated into a numerical score on two variables, "diagnosis and aetiology" and "patient management plan".

The variable "diagnosis and aetiology" (DIAG) is a summation score of the correctness of the diagnosis, the elaboration of the differential diagnosis and correctness of explanatory hypotheses.
The variable "patient management plan" (HELP) is a summation of the correctness of the hypotheses about further patient management and the agreement between the request for help and the proposed patient management plan.

13.2.1.3 Data analysis.

First, the descriptive statistics of both variables DIAG and HELP are given.

Secondly, the relationship of the residents' interviewing skills to his diagnostics and clinical problem-solving is determined by Pearson's correlations of both variables DIAG and HELP with the 8 MAAS-PMHC scale scores. To estimate the "case influence", a dummy variable, taking the

value 1 of the "depression case" and the value 2 for the "anxiety case", are included in the correlation matrix.

In order to calculate validity coefficients from these correlations, a correction for attenuation is applied to the MAAS-scores and to the scores of diagnostics and clinical problem-solving (Guilford et al., 1982).

The validity coefficients are calculated from the Pearson correlations of 8 MAAS-PMHC scale scores with scores on both variables DIAG and HELP. In the formula of correction for attenuation, the inter-rater reliability (Kappa = 0.88, Cohen, 1960) of the variables DIAG and HELP are taken. As reliabilities for the MAAS-scales, the generalizability coefficients (random physician, observers; fixed items) for 2 observers are used (table 10.4), because a summated set of scores over 2 observers are taken for these validity studies.

13.2.2 Results.

Descriptive statistics of the variables DIAG and HELP over 40 residents and 2 cases are given in table 13.1.

Table 13.1: Descriptive statistics of the variables "diagnosis and aetiology" (DIAG) and "patient management plan" (HELP) of 40 residents, interviewing two simulated patients.

	mean	SD	kurtosis	skewness
DIAG (N=80)	30.78	11.45	-0.45	-0.35
HELP (N=80)	20.23	6.51	0.26	-0.73

The distribution of these scores is almost normal whereas their variances are sufficient to allow further analyses.

The matrix of the validity coefficients of 8 MAAS-PMHC scale scores with the scores on both variables DIAG and HELP is given in table 13.2. Moreover, the Pearson correlation of case influence with the other variables is included in the matrix.

Table 13.2: Matrix of validity coefficients of scores on MAAS-PMHC scales with measurements of diagnostics and problem-solving (after correction for attenuation).

	EE	HT	PE	SE	PS	ST	IPS	CS	CAS
DIAG	.26	.31*	.68*	.38*	-.10	-.04	-.06	-.68*	-.58*
HELP	.02	.21	.42*	.04	.37*	.23	-.04	-.48*	-.41*
CAS	-.36*	-.42*	-.30*	-.28*	-.19	.08	.09	.24	1.00

* $p \leq .01$ (N=80; 2-tailed test)

Legend:

EE = exploration for the reason for encounter
HT = history-taking
PE = psychiatric examination
SE = socio-emotional exploration
PS = presenting solutions
ST = structuring the interview
IPS = interpersonal skills
CS = communicative skills
CAS = case influence (dummy variable)
DIAG = correctness of (differential) diagnosis and of explanatory hypotheses
HELP = correctness of hypotheses concerning further patient management and agreement between request for help and proposal patient management plan

We find moderate validity coefficients between DIAG with the scales "history-taking" and "socio-emotional exploration". They are fairly high between DIAG and "psychiatric examination" (positive) and "communicative skills" (negative).

The variable HELP has moderate validity coefficients with the scale "psychiatric examination" and "presenting solutions" and (again) negatively with communicative skills. The variable DIAG and HELP both have considerable correlations with the dummy variable "case influence".

Finally, we also encounter significant and moderate correlations of "case influence" with all MAAS-PMHC scales, except for "structuring the interview" and "interpersonal skills".

13.2.3 Discussion.

Although already noted in chapter 10, the "case influence" in the scores on the MAAS-PMHC scales is again striking. Not only scales with many content categories, but also scales measuring process skills ("presenting solutions" and "communicative skills") suffer from case specificity. The "case influence" on the variables DIAG and HELP is also very notable as in all measures of medical problem-solving up until now (a.o. Elstein et al., 1978).

According to our expectation, the validity of the scales "history-taking", "psychiatric examination" and "socio-emotional exploration" is supported by the variable DIAG. Salient is the lack of validity of "exploration of the reason for encounter" and the magnitude of the validity coefficient of "psychiatric examination".

A closer look at the variable DIAG provides further insight into this validity issue. DIAG combines the quality of diagnosis on the symptom level (see 2.4.2.2.1) with that on an aetiological explanatory level. Diagnosis on the symptom level is brought about by information from "psychiatric examination". Diagnosis on the aetiological level is achieved by information-gathering during "history-taking" and "socio-emotional exploration". Apparently, in our study, the latter relationship cannot be so firmly established which may indicate a lack of construct validity in the scales "history-taking" and "socio-emotional exploration" and its absence in the scale "exploration of the reason for encounter".

The validity coefficient between "communicative skills" and the variable DIAG, being negative and fairly high, is a conspicious, but difficult-to-explain finding. It might point to deficient validity of this scale, probably based on a lack of reliability.

With measurements of hypotheses concerning the patient-management plan and of their accordance with the request for help (HELP), the MAAS-PMHC scales show two moderate validity coefficients. In accordance with expectations, the scale "psychiatric examination" has a significant validity coefficient but at least similar findings should be expected from the scales "history-taking" and "socio-emotional exploration" which purport to measure the data collection necessary for treatment hypotheses. A second significant validity coefficient is

found between "presenting solutions" and the variable HELP. It might support the validity of the negotiation aspect in the scale "presenting solutions". Negotiation is necessary to attain accord between physician and patient on the treatment proposal. For this negotiation process, the skills to "explore the reason for encounter" are also considered to be necessary to attune the treatment proposal to the patient's desires and needs (Tuckett, 1985). Therefore, a significant validity coefficient between the scale "exploration of the reason for encounter" with the variable HELP is missing.

13.2.4 Conclusion.

The construct validity with respect to diagnostics and clinical problem-solving turned out to be supported for the MAAS-PMHC scales "history-taking", "socio-emotional exploration" and "psychiatric examination". Moreover, there is slight support in validity of the MAAS-scale "presenting solutions" with respect to measurements of patient management hypotheses and their accordance to the patient's request for help. Also in this context, the MAAS-scale "psychiatric examination" takes the predominant role. Conspicious is the lack of validity of the MAAS-scales measuring process skills of interviewing in respect to diagnostics and clinical problem-solving.

13.2 Interviewing skills and patients' satisfaction with communication, comprehension, intention to comply and recall of conveyed information.

In this section, validity of the MAAS-PMHC in terms of relationships with variables measuring the effects of the interview on patients is assessed. In this respect, we heavily lean on the theory, described in chapter 3, from which we derive hypotheses about significant validity coefficients with MAAS-variables. We recapitulate the following six sets of theoretical relationships:

- Interviewing skills pertaining to the scales "exploration of the reasons for encounter", "interpersonal and communicative skills" and, to a lesser extent, "socio-emotional exploration", enhance the patient's feeling of being facilitated by the physician. These feelings of "facilitation" are elicited when the patient is

encouraged to describe his problems and co-existing emotions and attributions and to give additional information in an interviewing climate of trust and acceptance (Korsch et al., 1972; Eisenthal et al., 1976; Wolff et al., 1978; Stiles et al., 1979; Putnam et al., 1985).

- By contrast, the patient's perception of disrupted communication might be induced when he has not been able to tell "the story" from his own frame of reference. The variable "disrupted communication" is expected to act contrary to "facilitation". "Disrupted communication" is thus considered as the counterpart of "facilitation" or "affective satisfaction" (Wolff et al., 1978).
- The patient's perception of being directed towards ideas and solutions in the physician's frame of reference ("directivity") may be influenced by interviewing skills as measured by the scales "history-taking", "psychiatric examination" and "socio-emotional exploration". Moreover, process skills as measured by the scale "structuring the interview" may have a relation to this patient variable.
- The patient's recall of conveyed information is considered to be enhanced by a combined influence of the physician's skills measured by the scales "present solutions", and "interpersonal and communicative skills".

  Recall of information conveyed by the physician is improved by the use of appropriate methods to inform the patient (Ley, 1983; see also 2.3.2.2), by physicians addressing the patient's concerns and expectations (Bartlett, 1981) and by an explanation suited to the patients's frame of reference (Tuckett et al., 1985).
- The patient's insight is considered to increase by means of the information conveyed by the physician, by means of the "communicative and interpersonal skills" and the skills to "present solutions" (a.o. Wolff et al., 1978; Putnam et al., 1985; Pendleton, 1983).
- The patient's intention to comply with advice is influenced by the physician's "interpersonal and communicative skills" (Bartlett, 1981; Davis, 1968) and by the physician's skill to "present solutions" (Eisenthal et al., 1976).

These theoretical relationships are investigated in an explorative study with validity coefficients.

13.3.1 Method.

In this section, we discuss the instruments for measuring the variables and, subsequently, the analysis of the data. The research setting has already been described in 13.2.1.

Instruments.

Besides the MAAS-PMHC, two other instruments are used: the Patient Satisfaction with Communication Checklist (PSCC)and an instrument to measure the patient's recall of information conveyed by the physician.

With the 19-item PSCC, extensively described and studied in chapter 6, the following variables are measured: "facilitation" (3 items), "disrupted communication" (4 items), "directivity" (3 items), "insight" (5 items) and "intention to comply" (4 items).

After the completion of the PSCC, the simulated patients were asked to respond on paper to 6 open questions pertaining to the information conveyed by the physician. These questions are stated in table 13.3.

Table 13.3: Questions asked the patient about the information conveyed by the physician.

1. What is wrong with you, according to the physician?
2. What are, according to your physician, the cause(s) of your complaint(s)?
3. What advice has the physician given?
4. What further examination(s)/test(s) will take place?
5. What treatment has been advised?
6. To whom have you been referred?

The written responses are taken as the information conveyed by the physician and recalled by the patient. To measure the quality of recall, this written recalled information is compared with the information the physician actually conveyed during the interview. This comparison is carried out with a specially constructed instrument, the Recall of Conveyed Information (RCI). The most important part of this instrument is given in table 13.4.

Table 13.4: Recall of Conveyed Information (RCI), a rating scale to test the patient's recall of information conveyed by the physician and its inter-rater reliability.

RECALL OF CONVEYED INFORMATION	(1)	(2)	(3)	(4)	(5)	inter-rater reliability (Pearson correlation)
information conveyed about:						
diagnosis	o	o	o	o	o	.90
aetiological conditions	o	o	o	o	o	.86
treatment related advice	o	o	o	o	o	.80
further examination or tests	o	o	o	o	o	.97
referral	o	o	o	o	o	.92
total recall of conveyed information						.89

Scoring:
1 = no information written down by the patient
2 = information totally incorrectly reported
3 = less than 50% of the information correctly reported
4 = less than 90% and more than 50% of the information correctly reported
5 = more than 90% of the information correctly reported

The RCI enables the researcher to split the total amount of conveyed information into 5 categories: information conveyed about diagnosis, aetiological conditions, treatment-related advice, further examination/tests and referral. Two observers independently compared the information conveyed by the physician by reviewing the videotaped interviews with the responses given by the simulated patient concerning the information conveyed to them. Both observers rated the degree of agreement on the RCI scales.

Data analysis.

Firstly, descriptive statistics and reliabilities of the 5 PSCC variables and of the variable "recall of conveyed information" are calculated.

Secondly, the effects of the residents' interviewing skills on the simulated patients are explored by Pearson correlations. The 8 scores on the MAAS-PMHC scales are correlated with each of the 5 PSCC variables "facilitation", "disruptive communication", "directivity", "insight" and "intention to comply". From these correlations, validity coefficients are calculated after correction for attenuation (see also 13.2.1).

As reliability to be used in the formula for correction for attenuation, the alphas of the 5 (composed) variables of the PSCC are taken. For the same purpose, we take the averaged inter-rater reliability of the (composed) variables "recall of conveyed information" for the correction for attenuation.

To estimate the "case influence", a dummy variable, taking the value 1 for the "depression case" and the value 2 for the "anxiety case", is included in the correlation matrix.

13.3.2 Results.

Descriptive statistics and reliabilities of "recall of conveyed information" (see table 13.4) the PSCC variables "facilitation", "disruptive communication", "directivity", "insight" and "intention to comply" (see table 13.5) permit further research, although the distributions of "disrupted communication" and "intention to comply" are somewhat skewed.

The validity coefficients between the scores on the MAAS-PMHC scales with the five PSCC variables and recall of conveyed information are presented in table 13.6.

From general inspection of the matrix, it appears that case influence significantly ($p \leq .01$) and sometimes substantially correlates with variables such as directivity (.59), insight (.55) and recall of conveyed information (.80).

Table 13.5: Descriptive statistics of the composed PSCC variables "facilitation", "disrupted communication", "directivity", "insight" and "intention to comply" (N=80).

	mean	SD	range	kurtosis	skewness	alpha
"facilitation"	1.93	0.98	0-3	-0.30	-0.76	.80
"disrupted communication"	3.67	0.71	1-4	2.98	-2.01	.70
"directivity"	1.75	1.20	0-3	-1.41	-0.36	.85
"insight"	0.79	1.16	0-5	2.72	1.74	.79
"intention to comply"	3.18	0.81	1-4	-0.35	-0.63	.65

Furthermore, there is a considerable amount of moderate to reasonable validity coefficients between MAAS-scales and PSCC variables, except with the variable "intention to comply". We discuss these validity coefficients in the next section.

13.2.3 Discussion.

In this study we examine the validity of the MAAS-scales by means of their validity coefficients with variables measuring the effect of the interview on the patient. In 13.2.1, we postulated 6 groups of hypotheses concerning these validity coefficients. These are subsequently discussed.

The feelings of the patient of being facilitated to tell his own story ("facilitation") has a significant validity coefficient with all MAAS-scales except "communicative skills". The coefficients with the scales "socio-emotional exploration" and "interpersonal skills" are even considerable to fairly high. This finding agrees more or less with our expectations because these latter scales pertain to the interviewing skills leading to facilitation and fostering the expression of patient-centered information. Nevertheless, we expected the validity coefficient with the scale "exploration of the reason for encounter" (.33) to be higher.

Table 13.6: Matrix of validity coefficients between scores on MAAS-scales with patient variables measuring facilitation and disruption of communication, perceived directivity, insight, intention to comply and recall of conveyed information.

	EE	HT	PE	SE	PS	ST	IPS	CS	CAS
FAC	.33*	.28*	.29*	.49*	.30*	.30*	.59*	.13	-.15
DCO	-.50*	-.17	-.23	-.33*	.06	-.46*	-.15	-.03	.23
DIR	-.42*	-.58*	-.55*	-.36*	.12	.05	.19	-.10	.59*
INS	-.12	-.36*	-.16	-.06	.63*	.15	.08	-.10	.55*
COM	-.12	-.25	.02	.13	.18	-.02	.15	-.31*	.05
REC	-.27	-.61*	-.27	-.07	.13	.32*	.27	.37*	.80*

*) $p \leq .01$ (N=80; 2-tailed test)

Legend:

EE = exploration of the reason for encounter
HT = history-taking
PE = psychiatric examination
SE = socio-emotional exploration
PS = presenting solutions
ST = structuring the interview
IPS = interpersonal skills
CS = communicative skills
CAS = case influence
FAC = the patient's feelings of being facilitated to tell his own story
DCO = the patient's feelings of being disrupted in the narration of his own story
DIR = the patient's opinion about the degree to which the physician makes his mark on the course of the consultation
INS = the patient's insight into his problem by means of the information conveyed by the physician
COM = the patient's intention to comply with proposed advice
REC = the patient's recall of information conveyed by the physician

Significant validity coefficients with the scales "history-taking" and "psychiatric examination" point to the conclusion that patients may consider periods of systematic and directive questioning by the physician as facilitative. However, these findings also may be due to the validity problem of the scale "facilitation" in which patients tend to evaluate their relationship with the physician as overly satisfactory (see 3.2). Nevertheless, a significant validity coefficient with "communicative skills" is missing.

The patient's feeling of being disrupted in the narration of his story ("disrupted communication") shows a correlative pattern with MAAS-variables which is globally the reverse of that of "facilitation". This is to be expected because the patient-variable "disrupted communication" is theoretically the opposite of "facilitation", the measure of the patient's affective satisfaction (see 3.2). We find negative and significant validity coefficients with the MAAS-scales "exploration of the reason for encounter", "structuring the interview" (both moderate) and with "socio-emotional exploration" (modest), according to expectation. We miss (negative) correlations with "interpersonal and communicative skills". This may indicate that the patients feel more disrupted when hampered in the expression of certain factual or emotional information, than by the way they are allowed to express it. The negative correlation with "structuring the interview" may indicate that this MAAS-scale measures a way of structuring the interview in phases that is generally considered as facilitative and satisfactory by the patient.

All these findings agree with related reports that systematic questioning must neither suppress the expression of feelings nor must it be experienced by informants as induly intrusive or lacking in understanding (Cox et al., 1981).

The patient's perception of being directed towards ideas and solutions in the physician's frame of reference ("directivity") has moderate to fairly high validity coefficients with the MAAS-scales "exploration of the reason for encounter", "history-taking", "psychiatric examination" and "socio-emotional exploration". These negative validity coefficients may indicate that the physician's skills pertaining to these scales are experienced by the patient as "not

directive". This is also in accordance with a previous remark that the influence of systematic questioning and structuring does not influence patient satisfaction negatively. The variable "directivity", not really described in the literature, does not indicate negative effects on the patient. In this context, Carter et al. (1982) has also reported positive effects of "physician's imperative direction" on the degree to which patients feel informed about their problems.

Furthermore, it is conspicious that "directivity" has similar negative validity coefficients to "exploration of the reason for encounter" and to the scales "history-taking", "psychiatric examination" and "socio-emotional exploration". This finding reconfirms a conclusion drawn in the previous chapter: it is difficult to draw a distinction between the variables "exploration of the reason for encounter" and "history-taking", both pertaining to patient-centered information.

In "directivity", a strong case influence is finally notable (r=.59). This finding means that the directivity perceived by the patient is very dependent on the nature of the problem and the way the patient presents his problem. Certain problems, or modes of their presentation, may apparently elicit more "directivity" from the physician than other problems.

The patient's recall of information conveyed by the physician ("recall"), has a modest, positive validity coefficient with the MAAS-scale "structuring the interview" and a fairly high, negative validity coefficient with "history-taking".

The modest validity coefficient with "structuring the interview" may point to a positive effect on the recall and retention of information when the interview is structured according to the criteria implied in the MAAS-PMHC. However, we lack validity coefficients with "presenting solutions" and "interpersonal and communicative skills" and neither does the negative validity coefficient with "history-taking" contribute much to construct validity. This relationship is understandable from the assumption that extensive "history-taking" has been carried out at the cost of the physician's conveyance of information and its subsequent recall by the patient. In summary, we are not able to confirm with the MAAS-PMHC the recommendations about effective

conveyance of information (see 2.3.1.1): for example, simplification and explicitation of information (Ley, 1983), checking of comprehension (Tuckett, 1985) and coping with defensive mechanisms after conveyance of bad news (a.o. Schouten, 1982).

The underlying reason for the poor contribution of "recall" to the construct validity of the MAAS-PMHC might be deficiencies in the method to measure the recall of information conveyed by the physician to the simulated patient (RCI). It is questionable whether the information simulated patients write down in response to the open questions (see table 13.3) reflects in a valid way the information the physician has conveyed. The response may be confounded with pre-existent knowledge about illness conditions and treatment from his simulated role. Furthermore, the response may be influenced by the simulated patient's ability to express himself in written language.

In addition, the strong case influence (r=.80) in "recall" might act as a confounder acting in suppressing potentially valid correlative relationships.

The patient's "insight" resulting from the information conveyed by the physician has significant validity coefficients with the MAAS-PMHC scale "presenting solutions" (r=.63). This fairly high validity coefficient is a logical consequence of the nature of this scale which measures the conveyance, the discussion and the bargaining of relevant information. When "presenting solutions" has been well achieved then, of course, it entails increased insight for the patient into his problems. The strong influence of case specificity is also here notable (r=.55), which is self-evident. Patients differ in their knowledge of mental health and disorder, often disposing of very idiosyncratic "fabrics" of knowledge. Tuckett (1985) therefore argues that the patient's insight can only be fostered when the physician himself has some insight into the patient's frame of reference pertaining to his case.

We here also lack positive validity coefficients with "interpersonal and communicative skills". The reliability of these scales may be too low to form a necessary condition for validity.

The negative validity coefficient of "insight" with "history-taking" may be due to the counteractive effects of extensive "history-

taking" on improving insight into the problems by the patient. In this respect, we have already assumed, in the previous subsection, a similar counteractive influence of "history-taking" on the "recall of conveyed information".

Disappointingly, the patient's "intention to comply" with the proposed medical advice ("intention to comply") does not show any positive validity coefficient with one of the MAAS-scale scores: or, at any events, we lack positive validity coefficients with "presenting solutions", "interpersonal and communicative skills". Surprisingly, a modest negative validity coefficient is found with "communicative skills". This finding, which is rather inexplicable, is probably again due to the low reliability of this scale.

We do not confirm the often-claimed relationship of "interpersonal and communicative skills" with "intention to comply" in this study (a.o. Davis, 1968; Bartlett, 1981; DiMatteo and DiNicola, 1982; Ley, 1983). Again, the lack of reliability is probably the explanation for the insufficient validity with respect to "intention to comply".

13.2.4 Conclusions.

In table 13.7, an overview of the theoretically relevant validity coefficients of the MAAS-PMHC scales with Patient Satisfaction with Communication variables, as well as with the variable recall of conveyed information is, presented.

In summary, the construct validity of the MAAS-scales "exploration of the reason for encounter", "history-taking", "psychiatric examination" and "socio-emotional exploration" is supported by the Patient Satisfaction with Communication variables as a criterion. To a lesser extent, the same holds for the scales "presenting solutions" and "structuring the interview". The construct validity of the scales "interpersonal and communicative skills" seems modest to low.

Some important validity coefficients with the MAAS-PMHC scales could not be established such as "recall of conveyed information" and, in particular, "intention to comply". The former may be partially due to a lack in validity in the way the recall of conveyed information is measured. The main cause, however, seems to be a deficit reliability and therefore a lack of validity of the scales "interpersonal and communicative skills".

Table 13.7: Overview of theoretically relevant validity coefficients of MAAS-PMHC scales with Patient Satisfaction with Communication (PSCC) variables and the variable recall of conveyed information.

PSCC variables	MAAS-scales with significant validity coefficients of theoretical importance	MAAS-scales with expected validity but not confirmed by validity coefficients
facilitation	- exploration of the reason for encounter (.33) - socio-emotional exploration (.49) - interpersonal skills (.59)	- communicative skills
disrupted communication	- exploration of the reason for encounter (-.50) - structuring the interview (-.46) - socio-emotional exploration (-.33)	- interpersonal skills - communicative skills
directivity	- exploration of the reason for encounter (-.42) - history-taking (-.58) - psychiatric examination (-.55) - socio-emotional exploration (-.36)	- presenting solutions - structuring the interview
insight	- presenting solutions (.63)	- communicative skills
intention to comply	none	- presenting - interpersonal skills - communicative skills (pos)
recall of conveyed information	- structuring the interview (.32) - communicative skills (.37) - history-taking (-.61)	- interpersonal skills - presenting solutions

Furthermore, the case influence in both MAAS-scales and PSCC variables may also play a confounding role in causing spurious significant and suppressed non-significant correlations (Guilford et al., 1982).

Finally, our statements about the validity coefficients should be treated with caution and are definitely not to be taken as causal relationships. Within the framework of this study, the validity coefficients are no more than inductive and hypothesis-generating in character. They are insufficient to confirm or refute hypotheses about construct validity.

Restrictions in the research situation also attenuate our validity statements.

The simulated consultation hour might be a threat to the "ecological validity" of our findings. In particular, the PSCC-scores of simulated patients may be questionable, especially in the more affective variables such as "facilitation". Furthermore, the validity of the information recalled by simulated patients is questionable, irrespective of the measurement method itself.

13.4 Summary and final conclusions.

In an explorative, correlative study, some issues of construct validity of the MAAS-PMHC have been investigated. Underlying theoretical relationships for construct validity have been derived from the three functions of initial medical interviews, as stated by Schouten et al. (1982).

Medical interviewing skills are intended to collect information from the patient enabling diagnostics and clinical problem-solving. Furthermore, the physician should convey information on problems and possible solutions, resulting in increased insight and compliance. Finally, the physician should build and maintain a relationship of trust and acceptance in which the previously mentioned functions of the medical interview can be performed. It has been stated that the construct validity of the MAAS-PMHC is supported if this method is able to measure interviewing skills which have validity coefficients with measured variables such as quality of diagnostics and clinical problem-solving, patient's feelings of being facilitated or disrupted in their

communication, recall of conveyed information, insight into their problems and intention to comply with advice.

With respect to diagnostics and problem-solving, the MAAS-scales "history-taking", "socio-emotional exploration" and, in particular, "psychiatric examination", show significant validity coefficients indicative of their construct validity. Measurements of the quality of the patient-management hypotheses and their accordance with the patient's request for help give some support to the scales "psychiatric examination" and "presenting solutions".

The construct validity of the MAAS-scales "exploration of the reason for encounter", "history-taking", "psychiatric examination" and "socio-emotional exploration" is supported by the Patient Satisfaction with Communication variables as a criterion. To a lesser extent, the same holds for the scales "presenting solutions" and "structuring the interview". The construct validity of the scales "interpersonal and communicative skills" seems modest to low.
The variable "recall of conveyed information" fails to act as a criterion in construct validity.

The methodology used in this study only permits cautious statements about construct validity. The validity coefficients found in this exploratory study are indicative of construct validity, but are not confirmative. Our results therefore have only a hypotheses generating character. It is necessary to carry out further research to establish whether confirmation or refutation of these hypotheses is possible. This will require methodology with regression and path analysis and the opportunity for cross-validation in more samples.

Furthermore, it is necessary to point to other restrictive considerations concerning our findings.

Firstly, the simulated consultation hour procures standardization and comparability at the cost of ecological validity. Secondly, the low to moderate reliability of some scales such as "presenting solutions" and "communicative and interpersonal skills", has put restraints on these validity studies, the correction for attenuation. Thirdly, "case influence" contributes much variance to the MAAS-PMHC scales and to some variables such as "insight", "recall of information" and "directivity". It is plausible that this variable may provoke spurious correlations or suppress potentially valid correlations.

REFERENCES

Bales RF. Interaction process analysis. Addison Wesley, Cambridge, 1950.

Bartlett EE. The contribution of consumers health education to primary care practice: a review. Medical Care, 1980; 18: 862-871.

Carter WB, Inui TS, Kukull WA, Haigh VH. Outcome-based doctor-patient interaction analysis. II. Identifying effective provider and patient behavior. Medical Care, 1982; 20: 550-566.

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960; 20: 37-46.

Cox A, Rutter M, Holbrook D. Psychiatric interviewing techniques V. Experimental study: eliciting factual information. British Journal of Psychiatry, 1981; 139: 29-37.

Davis MS. Variations in patients' compliance with doctor's advice: an empirical analysis of patterns of communication. American Journal of Public Health, 1968; 58: 274-288.

DiMatteo MR, DiNicola DD. Achieving patient compliance. The psychology of the medical practitioner's role. Pergamon Press, New York, 1982.

Eisenthal S, Koopman C, Lazare A. Process analysis of two dimensions of the negotiated approach in relation to satisfaction in the initial interview. Journal of Nervous and Mental Diseases, 1983; 171: 49-54.

Elstein AS, Shulman LS, Sprafka SA. Medical problem solving. An analysis of clinical reasoning. Harvard Univ. Press, Cambridge Mass., 1978.

Goldberg D, Huxley P. Mental illness in the community; the pathway to psychiatric care. Tavistock Publ., London/New York, 1980.

Guilford JP, Fruchter B. Fundamental statistics in psychology and education. McGraw-Hill, Auckland, 1982 (6th ed.).

Inui TS, Carter WB, Kukull WA, Haigh VH. Outcome-based doctor-patient interaction analysis. I. Comparison of techniques. Medical Care, 1982; 20: 537-549.

Kassirer JP, Gorry GA. Clinical problem solving: a behavioral analysis. Annals of Internal Medicine, 1978; 89: 245-255.

Korsch BM, Negrete VF. Doctor-patient communication. Scientific American, 1972: 66-74.

Ley P. Patients' understanding and recall in clinical communication failure. In: Pendleton D, Hasler J (Eds.). Doctor-patient communication. Academic Press, London, 1983.

Pendleton D. Doctor-patient communication; a review. In: Pendleton D, Hasler J (Eds.). Doctor-patient communication. Academic Press, London, 1983.

Putnam SM, Stiles WB, Jacob MC, James SA. Patient exposition and physician explanation in initial medical interviews and outcomes of clinic visits. Medical Care, 1985; 23: 74-83.

Roter DL. Patient participation in the patient-provider interaction: the effects of patient question asking on the quality of interaction, satisfaction and compliance. Health Education Monographies, 1977; 5: 281-315.

Schouten JAM. Anamnese en advies. Stafleu, Alphen a/d Rijn/Brussel, 1982.

Stiles WB. Verbal response modes and dimensions of interpersonal roles: a method of discourse analysis. Journal of Personality and Social Psychology, 1978; 36: 693-703.

Stiles WB, Putnam SM, Wolf MH, et al. Interaction exchange structure and patient satisfaction with medical interviews. Medical Care, 1979; 17: 667-679.

Tuckett D. Meeting between experts: an approach to sharing ideas in medical consultations. Tavistock, London, 1985.

Tuckett D, Williams A. An approach to the measurement of explanation and information giving in medical consultations: a review of empirical studies. Social Science in Medicine, 1984; 18: 571-580.

Wolf MH, Putnam SM, James SA, Stiles WB. The medical interview satisfaction scale: development of a scale to measure patient perceptions of physician behavior. Journal of Behavioral Medicine, 1978; 1: 391-401.

CHAPTER 14 SUMMARY AND CONCLUSIONS

14.1 Medical interviewing and the construction of the MAAS.

In medical interviewing there is a typical constrast between its generally agreed upon importance and the relative neglect of this competency by the medical profession. This is remarkable because physicians spend about 60% of their time devoted to patient care in talking to patients.

Fortunately, the teaching of interviewing skills in medical schools has been expanding increasingly over the last two decades. Evidence of its clinical importance has gradually become clear: medical interviewing skills enhance accurate diagnosis and clinical problem-solving; they establish an effective and trusting physician-patient communication which results in satisfaction and compliance with medical advice. Furthermore, they fosters early detection of mental health problems, minimizing somatic fixation and medicalization.

To serve educational objectives, we constructed two observation instruments to evaluate medical interviewing skills: the Maastricht History-taking and Advice Checklist (MAAS). The MAAS-GP is the version applicable in General Practice, whereas the MAAS-PMHC is to be used in Primary Mental Health Care.

The 68 items of the MAAS-GP are spread over 6 scales. The MAAS-PMHC has a comparable content and format but is extended by a further two scales and contains 104 items.

- Scale I, "exploring reasons for encounter", measures the ability to clarify the patient's complaint and to explore the request for help leading to the visit to the physician. This is the patient-centered part of the medical interview.
- Scale II consists of "history-taking skills". By this type of questioning the physician is able to generate hypotheses and to explain the patient's complaint in medical terms.
  In the MAAS-Primary Mental Health Care, this scale is divided into three parts. Besides the scale "history-taking", two further scales have been added: "psychiatric exploration" measures the skills to explore possible psychiatric symptoms in order to make a psychiatric

assessment of the patient; "socio-emotional exploration" measures the extent to which the physician explores aetiological conditions or consequences of mental health problems.

- Scale III deals with the interviewing skills during the "presentation of solutions". It pertains to the exchange of information on the medical problem (cause, prognosis) and to the negotiation between physician and patient about the problem definition and solutions to the problems (advice, referral, etc.).
- By scale IV, the way physicians "structure the interview" (introduction, balance of patient- and physician-centered styles of interviewing, sequence of the different phases, closing) is judged.
- Scale V consists of "interpersonal skills", the establishment of rapport with the patient. Moreover, the quality of the physician's response to emotions is assessed.
- Scale VI pertains to the "communicative skills", intended to promote the exchange of information between physician and patient.

The MAAS is based on an educational model of initial medical consultations. The first 3 scales pertain to the three characteristic phases of initial medical interviews. The remaining scales intend to measure skills pertaining to process aspects of interviewing.

The core question in this thesis pertains to the issue of whether the MAAS is an observational method for initial interviewing, which meets generally agreed upon criteria of scalability, reliability, validity and practicability.

14.2 Experimental situation.

The research questions of this thesis have been studied in a so-called simulated consultation hour. In the consultation room forty residents in general practice interviewed each 4 "simulated patients". A group of simulated patients, who are normal persons trained to present a clinical case, consulted the residents for the first time with major depression, sudden chest pain, inception of diabetes mellitus (detected during a medical examination for a driving license) and a panic disorder. The resulting 160 consultations were videotaped. Groups of observers rated the interviews with both versions of the

MAAS. Furthermore, the 40 residents made self-ratings about their interviewing skills and about their diagnoses. Finally, groups of experts in general practice and primary mental health care also rate the interviews.

All these ratings served as the data base for the forthcoming studies.

14.3 MAAS-General Practice.

A.A.M. Crijnen

The process of scale construction of the MAAS-GP was secured by applying so-called Rasch analyses to the original scales. When these scales fit in this demanding Rasch model, a homogeneous scale originates which measures one important characteristic of the physician's interviewing skills. Fortunately, only a small number of items had to be excluded before the scales appeared to fit the assumptions of the Rasch-model. The scales seemed to measure only one underlying dimension which we called "medical interviewing skills in general practice", since each of them forms an indispensable element in initial interviews in General Practice.

The reliability of MAAS-GP was studied both on item and scale level by means of a generalizability study, an application of multi-variate analysis. In such studies the influences of observers disagreement, of the way the patient presents his complaint, of the type of complaint and of characteristics of the measurement method itself are disentangled. Our study clearly reveals that items worded in behavioral terms display high levels of inter-observer reliability whereas items pertaining to larger units of interview behavior, or items requiring interpretation by the observers, generate more difficulties with regard to inter-observer reliability. High reliability is obtained for the scale "history-taking", moderate reliability for the scales "exploring reasons for encounter", "presenting solutions" and "structuring the interview", whereas low reliability is observed for the scales "interpersonal skills" and "communicative skills". Inter-observer reliability can be enhanced significantly when two extra observers are used in the process of measurement.

Generalizability studies on the scales of the MAAS-GP reveal that influences of the method of measurement, namely halo, leniency and differences in interpretation between observers, impair the quality of measurement differently for each of the scales. Since the MAAS-GP requires human observers in the process of measurement, these influences have to be accepted as a feature of the measurement process. Training of observers and sharper delineation of the items are expected to increase reliability to some extent, although one must be aware that validity can be hampered by formulations, which are too sharp.

The generalizability studies also reveal an interesting but complex interaction between physicians' interviewing skills and patients' behavior during the interview. This is especially the case when patients insist on discussing certain topics during the consultation and when they wish to obtain information about specific issues their behavior will interfere with the physician's interviweiing style. This ultimately yields less information from a psychometric point of view about physicians' interviewing skills during the "exploration of the reason for encounter" and the "presentation of solutions". We conclude, therefore, that the reliability of interactional instruments measuring only one side of the interaction has to be studied cautiously and that the results should be interpreted comparatively rather than absolutely.

The convergent and divergent validity of the MAAS-GP was further analyzed by means of a so-called multitrait-multimethod matrix. By this methodology the "performance" of different measurement methods in measuring similar theoretical constructs ("traits") is compared (convergent validity). Also the power of measurement methods to distinguish between different "traits" can be assessed (divergent validity). Convergent validity of the MAAS-scales "history-taking", "presenting solutions", "structuring the interview" and "interpersonal skills" is clearly warranted by the strength of the correlations, whereas the scales "exploring reasons for encounter" and "communicative skills" fail to provide evidence of convergent validity. The support for MAAS operationalizations of medical interviewing skills provided by experienced General Practitioners is particularly very encouraging. Essentially, identical conclusions can be drawn for a self-evaluation variant of the MAAS (MAAS-SELF) and the Global Expert-Rating Scale,

whereas there is insufficient evidence to support convergent validity of the Global Self-Rating Scale.

Divergent validity of dimensions in medical interviewing skills was established for "history-taking", "presenting solutions", "structuring the interview" and "interpersonal skills". Difficulties arose in distinguishing dimensions referring to the "exploration of reasons for encounter" and "communicative skills". In addition, MAAS-GP appeared to be most effective for discerning dimensions: more so than the Global Expert-Rating Scale and the MAAS-SELF: the Global Self-Rating Scale was not able to distinguish dimensions.

We concluded therefore, that the MAAS-GP displays the best measurement properties when compared to three different methods of measuring medical interviewing skills.

The validity of the MAAS-GP was finally examined by correlating it with other measurements of medical competency. The validity coefficients between the MAAS-GP and related medical competencies such as physician's interpersonal skills, care and concern for the patient, and the exchange of information necessary to establish a diagnosis, confirmed the validity of the MAAS-GP. The validity coefficients with unrelated medical competencies such as medical knowledge and the quality of a plan for further treatment, showed the distinct character of the MAAS-GP and supported therefore its validity.

14.4 The Patient Satisfaction with Communication Checklist (PSCC). A.A.M. Crijnen

In addition to the MAAS-General Practice and MAAS-Primary Mental Health Care, the Patient Satisfaction with Communication Checklist was constructed. It measures the patient's evaluation of distinct dimensions in the medical interview. Patents turned out to be able to reliably discern five dimensions: the feeling of being facilitated to tell their own story and express their concerns, increased insight in their medical problems, intention to comply with medical advice, the feeling of being disrupted in their communication and the feeling of being directed by the physician to certain topics or to certain solutions.

The process of scale construction was supported by a sequence of factor analyses and Rasch-analyses to secure the uni-dimensionality of the scales and the quality of measurement.

The development of the PSCC enabled us to study the contruct validity of the MAAS-GP and MAAS-PMHC. Measurements with the MAAS can be related to outcome measurements of the medical interview, carried out with the PSCC.

Furthermore this instrument can be used in patient satisfaction studies in general practice and primary mental health care.

14.5 MAAS-Primary Mental Health Care.

H.F. Kraan

The scale construction of the MAAS-PMHC was also carried out with the so-called Rasch-model. Eighty eight of its 104 items fitted in this strict probabalistic scaling model. All 8 MAAS-PMHC scales turned out to be Rasch homogeneous. Moreover, all measure one general underlying dimension: "skills, necessary for initial interviewing in Primary Mental Health Care".

As described in the previous section the reliability of the MAAS-PMHC is again studied by means of generalizability analysis. The MAAS-PMHC turned out to be moderately hampered by observers' influences, most notable in the scales "structuring the interview, interpersonal and communicative skills". In these scales also leniency- and halo-effects play a marked role.

Susceptibility to case influences (differences in the nature of the case and in the way the patient presents them) appears substantial in all scales. This inter-case reliability is impaired in particular in the scales "history-taking" and "interpersonal skills". The causes for this are the liability of the MAAS for differences in mental health problems and especially for differences in their presentation by patients.

In general, the reliability on the scale level is proportionally higher than on the item level.

Content validity suffers from low item reliabilities because of resulting losses in theoretical content. The theoretical "strength" of

the scales "structuring the interview, interpersonal and communicative skills" is therefore restricted.

Experts support the representativeness of the MAAS-PMHC, except for the scales "presenting solutions" and "exploration of the reason for encounter". The latter scale is conceived more as a measurement of extensive collection of patient-centered information than as measurement of how the patient's request for help is elicited. Furthermore, experts support a general non-directive, patient-centered interviewing style which is periodically interrupted by physician-centered, directive or systematic data-gathering.

Parallel to the MAAS-GP convergent and divergent validity have been studied with the multitrait-multimethod matrix. The MAAS-PMHC has a reasonable convergent validity: the Global Expert-Rating Scale and the MAAS-Self measure approximately the same "traits" (expressed in the 8 scales of the MAAS-PMHC). Furthermore it is striking, that the MAAS-Self, an extensive self-assessment method of interviewing shows good characteristics, in particular, the scales measuring content aspects of initial interviewing such as "history-taking" or "socio-emotional exploration".

Divergent validity (the power of a measurement method to distinguish between different "traits") parallels the picture of the convergent validity of the MAAS-PMHC. Striking is the good performance of the MAAS-Self and the bad qualities of the Global Self-Rating Scale in this type of validity.

In general, the MAAS-PMHC performs best compared to the other methods concerning convergent and divergent validity. In measurement, the MAAS-PMHC takes as "good interviewing" a balanced combination of "interpersonal and communicative skills" with effective "history-taking" in its broad sense. The self-evaluation methods restrict somehow their valid measurement to the accurate collection of patient- and physician-centered data. In the Global Expert-Rating Scale, the validity of measurement of "presenting solutions" and "communicative skills" is notable. Method variance, partly attributable to leniency- and halo-effects, confounds the measurement of interviewing skills, especially in the global-rating scales and, to a lesser extent, in the MAAS-Self.

In further construct validity research, significant correlations were shown between the physician's competency in diagnostics and clinical problem-solving with the MAAS-scales "psychiatric examination" and "socio-emotional exploration".

Construct validity of the MAAS-PMHC is supported by means of the patient's judgments of the physician's interviewing skills as criteria. These judgements are measured with the Patient Satisfaction with Communication Checklist, described in the previous section. More detailed the MAAS-PMHC-scale are supported as follows by the PSCC-variables.

The patient's feeling of being facilitated to tell his own story has significant validity coefficients with the scales "exploration of the reason for encounter", "socio-emotional exploration" and "interpersonal skills". Its counterpart, the patient's feeling of being disrupted in his communication by the physician, has negative validity coefficients with the scales "exploration of the reason for encounter" and "structuring the interview". Furthermore, increased insight of the patient into his problems correlates with the scale "presenting solutions". Expected significant positive validity coefficients between the scales "presenting solutions" and "interpersonal and communicative skills" on the one hand, with the variables "recall of conveyed information" and "intention to comply" on the other hand could not be found.

These results support the MAAS-scales pertaining to skills necessary for the three phases of initial interviews better than the scales measuring process skills. The findings of this latter study, being explorative and hypotheses-generating in character, should be appraised with caution.

14.6 Theoretical significance of this thesis.

Medical interviewing has long been situated in the realm of art. Intuition and other rather vaguely decribed aptitudes were held as determinants of good interviewing, which, of course, were barely measurable. However, advances in behavioral assessment and the urge to include physician-patient communication in medical education has lead to the selction of a set of teachable and measurable interviewing skills from the "art" of physician-patient communication.

In our study, we have also encountered the limits into this development: "interpersonal and communicative skills" appears difficult to define and to operationalize in measurable interviewing behavior. Emotionally-loaded, non-verbals with strong communicative impact remain beyond reliable and behaviorally orientated measurement.

Another signifcant feature of this thesis is the issue of the most appropriate method of measurement for interviewing skills. From divergent and convergent validity studies, we have noticed the superiority of behavioral assessment by observers to global ratings. The same picture is revealed on comparing behavioral self-assessment with global self-rating scales. Although behavioral self-assessment has definite merits, it appears to be inferior to behavioral assessment where observers can use clearly-defined and unambiguous criteria for scoring.

The gains of further-improved measurement of interviewing skills are the more precise determination of their clinical effectiveness to patients. For instance, the effects of a directive and physician-centered interview style versus a non-directive and patient-centered interview style is gradually becoming more clear resulting in a differentiated picture of the advantages and disadvantages of both styles.

Reductionism may be an unfavorable side-effect of our research. We have mainly focused on the teachable physician skills in physician-patient communication, sometimes at the cost of the patient's contribution to the interview. Furthermore, we separated medical interviewing from the intricacies of the various aspects of medical competency. Finally, we limited our study to initial interviewing in general practice and primary mental health care.

When reliability is improved at the price of validity, reductionism also occurs. We noticed this phenomenon in the measurement of complex interviewing skills such as facilitation, confrontation, concretization and aspects of "structuring the interview". When we increase the reliability of such items by the use of more behavioral definitions, we do not value at its true worth the skills experienced interviewers show in such interventions. Nevertheless, we have put much effort into increasing reliability but we have also recognized its limits.

For instance, we had to acknowledge in our generalizability analyses that in measurement of physicians' interviewing skills, this "true" variance component is only a relatively low part of the total variance.

Finally, the use of simulated consultations as a research setting has their pros and cons compared with a naturalistic study of interviewing skills. Our experimental situation is advantageous in manipulating variables such as "cases", "case presentation" and "observers". Moreover, all subjects can be submitted to similarly standardized consultation situations. Although simulated patients undergo the human experience of being a patient, their playing of the patient-role is a violation of ecological validity. However, in the study of the Patient Satisfaction with Communication Checklist, we noticed great similarities in the reactions of both real and simulated patients to medical interviewers.

14.7 Practical significance.

Although medical interviewing skills are volatile and difficult to measure, it is a challenge to construct an instrument, which procures reliable and valid measurement of their process and content. Moreover, practicability is a main objective when the method is used to assess great numbers of students, residents and physicians within educational contexts. The MAAS largely fulfils these objectives with the burden of its scoring remaining bearable. Half of the items can be scored during the interview, including the items pertinent to the patient's contribution to the interview. The remaining part, mainly the process skills, requires a furhter 5-10 minutes of scoring time. This is far less time-consuming than methods of the interaction analysis type, which take up two for six hours of scoring per interview.

However, measurement with the MAAS also has its restrictions. Considerable case influences impair its (inter-case) reliability. Ideally, about 20 cases are necessary to completely remove these effects. Observers' effects are also notable, particularly as halo- and leniency-effects as well as differences in interpretation.

In order to improve reliability, it is better to use scores on the scale level which are generally higher in proportion to measures on the item level. Case influences can be restricted substantially by the use of roles for simulated patients, which are less demanding as to case

presentation. For research purposes, summated scores of at least two well-trained observers should be used.

The influence of improved measurement on the teaching of interviewing skills seems to be very productive.

Firstly, the operationalization and measurement of interviewing skills can improve educational programmes because it provides clear objectives. Moreover, students can use the MAAS or its versions in self- and formative evaluation.

Secondly, the Rasch analysis of the MAAS has provided scales with a hierarchy of difficulties of skills. Such a hierarchy may be of use in education.

Thirdly, the MAAS is appropriate for summative evaluation in medical education because its derivation from the educational model of initial interviewing provides the MAAS with a base for criteria-setting. In this way, the MAAS permits criterion- and group-referenced measurement in summative evaluation. Tests may be assembled from the Rasch homogeneous items of MAAS-GP and -PMHC with known item difficulties for student and resident populations. This enables constructors to tailor tests to the ability levels of students and residents. Moreover, the Rasch homogeneity of the scales allows comparisons between curricula of different institutions on an international level.

14.7 Further research.

Further research should first focus on improving item reliability by sharper definitions and criteria for scoring, especially in items with a considerable observer-variance component. Secondly, after amelioration of reliability, further construct validity research should be carried out to assess the impact of interviewing skills on the patient in greater detail. More sophisticated methodologies such as path analysis of nomological networks should be used.

Finally, cross validation studies should be carried out with populations with different levels of interviewing skills in simulated settings as well as in naturalistic settings to replicate studies in scalability and validity.

CHAPTER 15: SAMENVATTING EN CONCLUSIES

15.1 Het medisch interview en de constructie van de MAAS.

De waarde van het medisch interview in het kader van de uitoefening van de geneeskunde wordt alom erkend. Toch is er in de beroepspraktijk weinig aandacht voor gespreksvoering. Dit is merkwaardig omdat artsen ongeveer 60% van hun tijd besteden aan het praten met patiënten.

Gelukkig wordt er de laatste 15 jaar steeds meer aandacht besteed aan het onderwijs in gespreksvaardigheid. Hoe langer hoe meer is er bewijs beschikbaar gekomen dat een goede interviewtechniek leidt tot een betere diagnostiek. Het tot stand brengen van een vruchtbare arts-patiënt relatie, die zich kenmerkt door vertrouwen, leidt tot tevredenheid bij de patiënt en therapie-trouw. Tevens levert een goede interviewstijl een bijdrage aan de vroegtijdige opsporing van geestelijke gezondheidsproblemen (GGP) en doet dit het optreden van somatische fixatie en medicalisering verminderen.

Voor onderwijsdoeleinden werd aan de medische faculteit van de Rijksuniversiteit Limburg een observatie-instrument ontwikkeld voor het evalueren van medische gespreksvaardigheden, namelijk de Maastrichtse Anamnese en Advies Scoringslijst (MAAS). Het onderzoek vond plaats in het kader van het Hoofdproject Onderzoek van Onderwijs, onder de titel: "Competentiemeting in het Psycho-medisch domein".

De MAAS kent twee versies, namelijk de MAAS-Huisarts (MAAS-HA), voor toepassing bij initiële consulten in de huisartsgeneeskunde en de MAAS-Geestelijke Gezondheidsproblemen (MAAS-GGP) voor toepassing in de eerstelijns geestelijke gezondheidszorg.

De MAAS-HA bestaat uit 68 items verdeeld over zes schalen. De MAAS-GGP bestaat uit dezelfde zes schalen uitgebreid met twee schalen die gericht zijn op veel voorkomende geestelijke gezondheidsproblemen, in totaal 104 items.

Schaal I "Vraagverheldering", meet het interviewgedrag van de arts, gericht op het verhelderen van de klacht en op het op tafel krijgen van de redenen voor het bezoek aan de arts. Dit is het patiëntgerichte gedeelte van het gesprek.

Schaal II "Anamnese" bestaat uit anamnesevragen. Met dit soort vragen kan de arts hypothesen genereren over aard, oorzaken en

beïnvloedingsmogelijkeden van de klachten. In de MAAS-GGP is deze schaal in drieën verdeeld. Naast de anamnese-schaal is de schaal "psychiatrisch onderzoek" en "socio-emotionele exploratie" toegevoegd. De schaal "psychiatrisch onderzoek" dient om eventuele psychiatrische symptomen op het spoor te komen. De schaal "socio-emotionele exploratie" bevat items die gaan over uitlokkende factoren en gevolgen van geestelijke gezondheidsproblemen.

Schaal III heet "hulpaanbod" en bevat items omtrent informatie-overdracht bij het meedelen van diagnose en hulpaanbod. Ook het onderhandelingsproces tussen arts en patiënt over de juiste omschrijving van het probleem en mogelijke oplossingen (therapie, verwijzingen etc.) komt aan de orde.

In schaal IV wordt de wijze waarop de arts het gesprek structureert vastgelegd (opening, volgorde van de onderdelen, aandacht voor patiëntgerichte interviewstijl, afsluiting).

Schaal V bevat items over interpersoonlijke vaardigheden, nodig om een goede arts-patiënt relatie tot stand te brengen. Daarnaast wordt het omgaan met emoties van de patiënt beoordeeld.

Schaal VI bevat items over communicatieve vaardigheden, nodig om de uitwisseling van informatie tussen patiënt en arts te optimaliseren.

De MAAS is gebaseerd op een onderwijsmodel voor initiële medische interviews, hetgeen wil zeggen dat de arts en de patiënt elkaar nog niet eerder hebben gesproken over het huidige probleem. De eerste drie schalen lopen parallel met de drie karakteristieke fasen van een consult waarin een nieuwe klacht wordt gepresenteerd (vraagverheldering, anamnese, hulpaanbod). De overige 3 schalen zijn bedoeld om proces-aspecten van de gespreksvoering te meten.

De centrale vraag in deze dissertatie is of de MAAS als observatie-instrument voor initiële consulten voldoet aan algemeen geaccepteerde criteria van betrouwbaarheid, schaalbaarheid, validiteit en bruikbaarheid in de praktijk.

15.2 Onderzoekssituatie.

De onderzoeksvragen in deze dissertatie zijn vooral onderzocht tijdens een zogenaamd simulatie spreekuur. Veertig artsen van de

Maastrichtse beroepsopleiding tot huisarts, spraken gedurende 15 minuten met vier simulatiepatiënten (gezonde personen die getraind zijn in het presenteren van klachten en ziektegeschiedenissen) in het skillslab.

De simulatiepatiënten presenteerden de volgende klachten: een patiënte leed aan een depressie met vitale kenmerken,een patient kwam met plots ontstane pijn op de borst, een patiënte wilde meer weten over haar toevallig ontdekte type II diabetes mellitus, en een angstige patiënt kwam met klachten van een paniek stoornis (2 somatische en 2 geestelijke gezondheidsproblemen). Deze 40 x 4 consulten werden op videoband opgenomen. De banden werden beoordeeld door observatoren met de beide versies van de MAAS. De arts-proefpersonen gaven hun eigen oordeel over hun interviewvaardigheden door middel van twee zelf-evaluatie schalen. Ook schreven ze de door hen gestelde diagnose op. Tenslotte werden de banden beoordeeld door huisarts-experts en door experts uit de geestelijke gezondheidszorg. Deze gegevens bij elkaar vormden de basis voor de verrichte psychometrische analyses naar de betrouwbaarheid en validiteit van de MAAS.

15.3 MAAS-HUISARTS.

A.A.M. Crijnen

De schaalconstructie van de MAAS-HA is gebaseerd op Rasch-analyses van de oorspronkelijke schalen. Deze analyses geven aan of elk item op zinvolle wijze bijdraagt aan de meting van de diverse interviewvaardigheden. De moeilijkheidsgraad van alle items wordt bepaald en vervolgens wordt een schaal samengesteld bestaande uit items met toenemende moeilijkheidsgraad. Slechts enkele items dienden te worden verwijderd voordat voldaan werd aan de strenge eisen van het Rasch-model. De schalen bleken samen één basisdimensie te meten. Deze dimensie werd "medische interviewvaardigheid in de huisartsgeneeskunde" gedoopt omdat de schalen allen een onmisbaar element in consulten met patiënten, die een nieuwe klacht presenteren vormen.

De betrouwbaarheid van de MAAS-HA werd zowel bekeken op item- als op schaalniveau door middel van een generaliseerbaarheidsanalyse, een soort multivariate analyse. Deze analyse laat ondermeer zien hoe de betrouwbaarheid beïnvloed wordt door de observatoren, de

klachtpresentatie door de patiënt, de casus en de gebruikte items. In deze analyse werd duidelijk dat items die in gedragsmatige termen zijn beschreven een hoge inter-observator betrouwbaarheid laten zien. Items die een groter deel van een gesprek beslaan of die voor meerdere uitleg vatbaar waren, vertoonden een lagere inter-observator betrouwbaarheid. De schaal "anamnese" is zeer betrouwbaar te scoren. De schalen "vraagverheldering", "hulpaanbod" en "structurering" zijn redelijk betrouwbaar. De schalen "interpersoonlijke vaardigheden" en "communicatieve skills" zijn niet erg betrouwbaar. De betrouwbaarheid kan belangrijk verbeterd worden als gebruik gemaakt wordt van twee observatoren die eenzelfde gesprek bekijken.

De generaliseerbaarheidsanalyse van de MAAS-HA laat zien dat storende effecten veroorzaakt door de observator op verschillende manieren optreden. Deze effecten (b.v. halo-efffect, verschil in interpretatie van items) zijn overmijdelijk bij het gebruik van de mens als observator, en we moeten ze dan ook als onvermijdelijk en inherent aan onze metingen beschouwen. Een goede training van observatoren en nauwkeurige afbakening van de reikwijdte van de items kunnen er toe leiden dat deze effekten geminimaliseerd worden en de betrouwbaarheid toeneemt. Aanbevelingen hiertoe zijn in het proefschrift beschreven.

De generaliseerbaarheidsanalyses laten verder zien dat er een ingewikkelde interactie bestaat tussen de patiënt en het interviewgedrag van de arts-proefpersonen. Vooral als de patiënt erop staat om bepaalde onderwerpen aan bod te laten komen en specifieke informatie wil krijgen, wordt de interviewstijl van de arts aanzienlijk beïnvloed. We denken daarom dat de betrouwbaarheid van meetinstrumenten, die de inbreng van de arts in de communicatie meten zorgvuldig onderzocht moet worden omdat tot op heden onbekende invloeden mede een rol lijken te spelen. De resultaten van generaliseerbaarheidsanalyses zullen veel meer vergelijkend dan absoluut bekeken moeten worden.

De convergente en divergente validiteit van de MAAS-HA is onderzocht met een multitrait-multimethod matrix. In deze matrix worden de prestaties van verschillende meetmethoden bij de meting van eenzelfde theoretische "trek" met elkaar vergeleken, ook wel genoemd convergente validiteit, ("psychometrisch vergelijkend waren-onderzoek").

Aan de andere kant levert het een beeld op van het vermogen van de diverse instrumenten om verschillende "trekken" ook inderdaad te onderscheiden, ook wel genoemd divergente validiteit. Bestudering van deze matrix, volgens een viertal criteria, maakt uitspraken over de convergente- en divergente validiteit mogelijk. De convergente validiteit van de MAAS-schalen "anamnese", "hulpaanbod", "structurering" en "interpersoonlijke vaardigheden" blijkt duidelijk uit de sterkte van de correlaties, terwijl de schalen "vraagverheldering" en "communicatieve vaardigheden" geen duidelijke convergente validiteit blijken te bezitten.

Divergente validiteit werd aangetoond voor "anamnese", "hulpaanbod", "structurering" en "interpersoonlijke vaardigheden". Bij "vraagverheldering" en "communicatieve vaardigheden" lukte dit niet. De MAAS-HA bleek het beste van de onderzochte methoden in staat te zijn om de diverse dimensies in interviewgedrag te onderscheiden. De MAAS-Zelfbeoordelingsschaal en de Globale-Expertbeoordelingsschaal komen op de tweede plaats, terwijl de Globale-Zelfbeoordelingsschaal niet geschikt is om medische interviewvaardigheden te meten. Tevens blijkt dat de MAAS-HA er meettechnisch het best uit komt, vooral vanwege zijn lage methodevariantie.

Als laatste validiteitsonderzoek werden MAAS-HA scores gecorreleerd aan metingen van verschillende aspecten van medische competentie. De validiteitscoëfficiënten, verkregen door vergelijking van MAAS-HA met metingen van soortgelijke medische competenties, zoals interpersoonlijke vaardigheid van artsen, het tonen van zorg en medeleven met patiënten, het kunnen verkrijgen van informatie nodig om een diagnose te kunnen stellen, waren hoog en ondersteunden zo ook de validiteit van de MAAS-HA. De validiteitscoëfficiënten voor niet vergelijkbare competenties zoals medische kennis en het kunnen opstellen van een goed behandelplan waren laag en ondersteunden zo de validiteit van de MAAS-HA.

15.4 De Patient-Satisfactie met de Communicatie Checklist (PSCC).

A.A.M. Crijnen

Naast de MAAS-HA en de MAAS-GGP werd de PSCC geconstrueerd. Dit instrument meet het oordeel van de patiënt over een aantal dimensies in

het medisch interview. Patiënten bleken in staat om vijf verschillende dimensies in de communicatie te onderscheiden, te weten:

- het gevoel gestimuleerd te worden door de arts om het eigen verhaal te vertellen en de eigen zorgen uit te spreken;
- een beter inzicht te hebben gekregen in het probleem;
- de geneigdheid om adviezen op te volgen;
- het gevoel dat de communicatie regelmatig op onplezierige wijze wordt onderbroken;
- het gevoel gestuurd te worden met betrekking tot onderwerpen in het gesprek.

Bij de constructie van dit instrument werd gebruik gemaakt van een aantal factor-analyses en Rasch analyses om de uni-dimensionaliteit van de schalen en de meetkwaliteit te garanderen. Door de ontwikkeling van de PSCC waren we in staat om de begripsvaliditeit van de MAAS-HA en MAAS-GGP te bestuderen. Metingen met de MAAS (het proces) kunnen namelijk vergeleken worden met metingen van hetzelfde medisch interview door middel van de PSCC (het produkt). Eenvoudiger gezegd: als men met de MAAS een consult goed oordeelt, dan dient men met de PSCC een hoge tevredenheid te meten.

Uiteraard kan de PSCC gebruikt gaan worden bij tevredenheidsonderzoek in de huisartspraktijk.

15.5 MAAS voor Geestelijke Gezondheidsproblemen (MAAS-GGP).

H.F. Kraan

De schaalbaarheid van de MAAS-GGP werd eveneens vastgesteld met het Rasch-model. Zoals eerder opgemerkt, wordt met een dergelijke analyse uit de items een schaal met toenemende moeilijkheidsgraad samengesteld, welke karakteristiek is voor de populatie van de betreffende proefpersonen. Van de 104 items uit de MAAS-GGP bleken er 88 te passen in dit strenge model.

Hoewel alle acht MAAS-GGP schalen Rasch homogeen bleken, meten ze samen één basisdimensie die omschreven kan worden als de vaardigheid nodig om een initieel gesprek in de eerste lijn over geestelijke gezondheidsproblemen te voeren.

De betrouwbaarheid van de MAAS-GGP wordt negatief beïnvloedt door observator invloeden, vooral in de schalen "structurering", "inter

persoonlijke vaardigheden" en "communicatieve vaardigheden". Waarschijnlijk spelen casus-invloeden een rol in alle schalen. De intercase betrouwbaarheid is vooral bij de schalen "anamnese" en "interpersoonlijke vaardigheden" laag. Dit wordt veroorzaakt door de gevoeligheid van de MAAS voor het verschil tussen diverse geestelijke gezondheidsproblemen en in het bijzonder door het verschil in presentatie van die problemen door de simulatie patiënten. Over het geheel genomen is de betrouwbaarheid op schaalniveau duidelijk hoger dan die op itemniveau.

De inhoudsvaliditeit van de MAAS-GGP wordt geschaad door de vrij lage betrouwbaarheid, als men de items afzonderlijk bekijkt. Lage item betrouwbaarheid leidt tot verlies van theoretische inhoud en vermindert derhalve de inhoudsvaliditeit. De schalen "structurering" en "interpersoonlijke vaardigheden" zijn hierdoor theoretisch enigszins verarmd. Hierdoor is de theoretische kracht van de schalen "structurering", "interpersoonlijke en communicatieve vaardigheden" beperkt.

De experts allen afkomstig uit de geestelijke gezondheidszorg ondersteunen de representativiteit van de MAAS-GGP items, behalve die van de schalen "hulpaanbod" en "vraagverheldering". Deze laatste schaal wordt door hen meer gezien als het uitvoerig verzamelen van patiënt gerichte informatie dan - zoals wij bedoelden - het verzamelen van informatie over de wijze, waarop de patiënt geholpen wil worden. Experts geven verder hun steun aan de manier van interviewen, zoals deze met de MAAS-GGP wordt gemeten. Deze stijl is in het algemeen nondirectief en op de patiënt gericht. De arts kan op sommige momenten deze wijze van interviewen onderbreken met directieve of systematische gegevensverzameling vanuit zijn medisch referentiekader.

De convergente en divergente validiteit van de MAAS-GGP wordt evenals de MAAS-HA onderzocht middels een multitrait-multimethode matrix. De MAAS-GGP bezit een behoorlijke convergente validiteit: zowel de Globale Expertbeoordelingsschaal als de MAAS Zelfbeoordelingsschaal meten ongeveer dezelfde trekken (zoals die in de acht schalen van de MAAS-GGP begrepen zijn). De divergente validiteit (het vermogen van een meetmethode om onderscheid te maken tussen verschillende trekken) vertoont hetzelfde beeld als dat van de convergente validiteit. Opvallend is de goede prestatie van de MAAS-Zelfbeoordelingslijst.

De Globale Zelfbeoordeling komt er slecht uit. Over het algemeen komt de MAAS-GGP er het best uit van de onderzochte methoden.

Uit dit validiteitsonderzoek komen ook de accenten, die deze instrumenten bij het meten leggen, naar voren. De MAAS-GGP benadrukt vooral het meten van interpersoonlijke en communicatieve vaardigheden in combinatie met een doelgerichte anamnese. Beide zelfevaluatie methoden beperken in hun bruikbaarheid tot de accurate vergaring van arts gerichte en patiënt gerichte gegevens. De Globale Expert-beoordelingsschaal laat een behoorlijke validiteit zien op de onderdelen hulpaanbod en communicatieve vaardigheden. Methode variantie bemoeilijkt de meting van interviewvaardigheden vooral in de beide globale beoordelingsschalen en - in mindere mate - in de MAAS-Zelfbeoordelingsschaal.

Onderzoek naar begripsvaliditeit toont significante correlaties tussen de diagnostische en probleem oplossende competentie van de arts en de MAAS schalen "psychiatrisch onderzoek" en "socio-emotionele exploratie". De begripsvaliditeit van de MAAS-GGP wordt ondersteund door het oordeel van de patiënt over de interviewvaardigheden van de arts. Dit oordeel werd gemeten met de PSCC. De PSCC schaal: "het gevoel gestimuleerd te worden door de arts om het eigen verhaal te vertellen" correleert significant met de schalen "vraagverheldering", "socio-emotionele exploratie" en "interpersoonlijke vaardigheden". De tegenovergestelde PSCC-schaal: "het gevoel steeds onderbroken te zijn in de communicatie" correleert negatief met de MAAS-schalen "vraag-verheldering" en "structurering". De schaal "een beter inzicht hebben gekregen in het probleem" correleert met de schaal "hulpaanbod".

De verwachte relatie tussen de MAAS-schalen "hulpaanbod" en "inter-persoonlijke en communicatieve vaardigheden" enerzijds en "herinnering van de verstrekte informatie" en "geneigdheid adviezen op te volgen" door de patiënt anderzijds kon echter niet worden gevonden.
Deze resultaten geven dus vooral steun aan die MAAS-schalen die betrekking hebben op de drie fasen van intitiële interviews en veel minder aan de schalen die proces aspecten van het interview meten. De resultaten uit het zojuist besproken gedeelte van deze begripsvaliditeitsstudie moeten voorzichtig bekeken worden vanwege hun exploratieve karakter.

15.6 Theoretische betekenis van deze dissertatie.

Medische gespreksvoering heeft lang gegolden als een kunst. Een goede gespreksvoering werd vooral geleid door intuïtie en andere vaag omschreven kundigheden, die uiteraard geen van allen goed te meten waren. Door de noodzaak om de arts-patiënt communicatie in het medisch onderwijs een plaats te geven werd, veelal op louter theoretische gronden, een selectie gemaakt van onderwijsbare interviewvaardigheden uit deze kunst van communicatie tussen arts en patiënt.

In ons onderzoek raakten we echter aan de grenzen van deze ontwikkeling van "kunst" tot "onderwijsbaar interviewgedrag": "interpersoonlijke en communicatieve vaardigheden" bleken moeilijk te omschrijven en te operationaliseren in meetbaar, "hard", interviewgedrag. Helemaal buiten het meetbereik van betrouwbare meetmethoden blijven emotioneel beladen, non-verbale signalen, die echter een grote invloed op het verloop van de communicatie lijken te hebben.

Een ander belangrijk aspect van deze dissertatie wordt gevormd door de vraag wat de meest geschikte methode van het meten van interviewvaardigheden is. Uit de validiteitsonderzoeken komt duidelijk naar voren dat gedragsmatige meting door observators superieur is aan beoordelingen op globaal niveau. Ditzelfde geldt voor de vergelijking van de gedetailleerde, gedragsmatige zelfbeoordeling met de globale zelfbeoordeling. Hoewel zelfbeoordeling zeker z'n verdienste heeft lijkt beoordeling door observatoren aan de hand van helder omschreven, eenduidige criteria veel beter bruikbaar.

De winst die behaald wordt met de verbetering van het meten van interviewvaardigheden ligt vooral in de mogelijkheid om de invloed van de kwaliteit van de interviewvaardigheden op de beleving en gezondheidstoestand van de patiënt nauwkeurig te gaan bestuderen. Zo worden bijvoorbeeld de effecten van een directieve, arts gerichte interviewstijl versus een non-directieve, op de patiënt gerichte stijl geleidelijk aan duidelijk. Dit levert een gedifferentieerd beeld op van de voor- en nadelen van beide stijlen.

Reductionisme is misschien een ongewenst neven-effect van ons onderzoek geweest. We hebben ons vooral verdiept in onderwijsbare vaardigheden van artsen in de arts-patiënt communicatie af en toe

misschien ten koste van het aandeel van de patiënt in het gesprek. Daarnaast hebben we ons best gedaan om medische gespreksvoering te isoleren uit het moeilijk te ontwarren web van de diverse aspecten van medische competentie.

We hebben ons beperkt tot initiële interviews in de huisartspraktijk. Reductionisme eist ook zijn tol als betrouwbaarheid wordt verhoogd ten koste van validiteit. Wij stuitten op dit verschijnsel bij het meten van ingewikkelde interviewvaardigheden als facilitatie, confrontatie, concretisering en aspecten van structurering. Als we de betrouwbaarheid van dergelijke items verhogen door het gebruik van meer gedragsmatige omschrijvingen doen we tekort aan de werkelijke waarde van de vaardigheid zoals een ervaren interviewer die laat zien bij dergelijke interventies. Desondanks hebben we ons veel moeite getroost om een hoge betrouwbaarheid te verkrijgen, maar we vonden daarin onze grenzen. In onze generaliseerbaarheidsanalyse kwamen we er bijvoorbeeld achter dat bij de meting van interviewvaardigheid van artsen de "ware" variantie component slechts een gering deel van de totale variantie uitmaakt.

Tenslotte heeft het simulatie contact ook zijn voor en tegen bij gebruik als onderzoeksinstrument vergeleken met een naturalistische setting. Onze experimentele situatie heeft voordelen wat betreft het kunnen manipuleren met variabelen als casus, klachtpresentatie en observatoren. Alle proefpersonen kunnen aan precies dezelfde spreekuur situaties worden blootgesteld. Hoewel simulatiepatiënten hun rol ondergaan als een menselijke belevenis, lijkt er toch een vermindering van de ecologische validiteit op te treden. Echter in ons onderzoek met de PSCC zagen we een opvallende overeenkomst in reactie tussen echte en simulatiepatiënten op medische interviewers.

15.7 Practische betekenis.

Omdat medische interviewvaardigheden ijl zijn en moeilijk te meten, vormde het een uitdaging een instrument te maken, dat een betrouwbare en valide meting van hun proces en inhoud mogelijk zou maken. Nog sterker: practische bruikbaarheid is een belangrijke eis als een meetinstrument wordt gebruikt bij het beoordelen van grote hoeveelheden studenten, arts-assistenten en practici in leersituaties. De MAAS

voldoet aan deze eisen terwijl het invullen van de lijsten nog draaglijk blijft. Ongeveer de helft van de items kan direct ingevuld worden tijdens het gesprek, inclusief de items die over het patiëntenaandeel in de informatie-uitwisseling gaan. De rest, vooral de proces-items, vragen ruim 5 minuten aan extra invultijd. Dit kost dus veel minder tijd dan de interactie analyse methode, die tussen de twee en zes uur per gesprek vergen.

Toch heeft het meten met de MAAS z'n beperkingen. De betrouwbaarheid wordt aanzienlijk verlaagd door casus invloeden. Idealiter zouden zo'n 20 casus nodig zijn om deze effecten volledig te verwijderen. Ook zijn er duidelijke observator effecten, vooral "leniency en halo effecten" (de neiging tot een (te) gunstige beoordeling ongeacht de prestatie resp. de neiging alle beoordelingen te laten beïnvloeden door een of enkele opvallende kenmerken van de proefpersoon) maar ook problemen bij de interpretatie van items. Om de betrouwbaarheid te verhogen is het beter om scores op schaal niveau te gebruiken. Casus invloeden kunnen fors ingeperkt worden door patiëntenrollen te gebruiken die weinig vergen op het punt van klachtpresentatie. Voor onderzoeksdoeleinden dienen gesummeerde scores van tenminste 2 goed getrainde observators gebruikt te worden.

De constructie van het instrument is erg stimulerend op het onderwijs in interviewvaardigheden geweest. Studenten kunnen de MAAS gebruiken in zelfevaluatie of formatieve evaluatie. De Rasch analyse van de MAAS heeft schalen opgeleverd die een in moeilijkheid oplopende hierarchie bevatten van vaardigheden. Zo'n hierarchie kan in het onderwijs nuttig zijn. De MAAS is geschikt voor summatieve evaluatie in het medisch onderwijs omdat door zijn afleiding uit het onderwijsmodel van het intiële interview de MAAS voorzien is van een basis voor normering. Met de MAAS kan summatief gemeten worden met gebruik van "group referenced" en "criterium referenced" normering. Toetsen kunnen worden samengesteld uit de Rasch homogene items van de MAAS-HA en MAAS-GGP, waarvan de item-moeilijkheid voor studenten en arts-assistenten populaties bekend is. Tenslotte laat de Rasch homogeniteit van de schalen toe dat er vergelijkingen worden getrokken tussen curricula in verschillende (inter)nationale instituten.

15.8 Verder onderzoek.

Verder onderzoek zou allereerst gericht moeten zijn op het verbeteren van de item betrouwbaarheid, vooral bij de items met een aanzienlijke observator variantie component. Vervolgens zou verder onderzoek naar de begripsvaliditeit moeten worden gedaan om het effect van interviewvaardigheden op de beleving en de gezondheidstoestand van de patiënt nauwkeuriger te bestuderen. Hierbij zou een meer sophisticated methode als pad analyse van een nomologisch netwerk gebruikt moeten worden. Tenslotte zou er cross-validatie onderzoek verricht kunnen worden bij populaties met uiteenlopende niveaus van interviewvaardigheden in zowel gesimuleerde alsook in naturalistische settingen om zo het onderzoek naar validiteit en schaalbaarheid te repliceren.

APPENDICES

A Manual for observers MAAS-GP
B Manual for observers MAAS-PMHC
C MAAS-Self-evaluation in GP
D MAAS-Self-evaluation in PMHC
E Global Self-Rating Scales
F Global Expert-Rating Scales
G Generalizability analysis of Rasch homogeneous scales
H Measurement of Problem-Solving in Primary Mental Health Care

APPENDIX A

THE MAASTRICHT HISTORY-TAKING AND ADVICE CHECKLIST (MAAS-GP)

AN OBSERVATION INSTRUMENT FOR THE MEASUREMENT OF THE PHYSICIAN'S INTERVIEWING SKILLS IN INITIAL MEDICAL CONSULTATIONS IN GENERAL PRACTICE

MANUAL FOR SCORING

BY

A.A.M. CRIJNEN, H.F. KRAAN, J. ZUIDWEG AND J. VAN DALEN

MAAS-GENERAL PRACTICE

LIST OF ITEMS

I. EXPLORATION OF THE REASON FOR ENCOUNTER

	YES	NO
1. Asks for the reason for encounter	0	0
2. Explores the emotional impact of the complaint/problem.	0	0
3. Asks the patient to clarify why he is presenting this problem at this particular moment.	0	0
4. Asks the patient to give his opinion on what are the causes of the problem.	0	0
5. Asks how the complaint or problem is discussed within the family or primary group.	0	0
6. Asks the patient to state what help (s)he desires.	0	0
7. Asks how the patient has tried to solve the problem by him/herself.	0	0
8. Explores the influence of the complaint on daily life.	0	0

II. HISTORY-TAKING

	YES	NO
9. Asks the patient to describe the complaint.	0	0
10. Explores the intensity of the complaint.	0	0
11. Asks about the localization of the complaint.	0	0
12. Asks about shifts/radiations of the complaint.	0	0
13. Asks about the course of the complaint during the day.	0	0
14. Asks about the history of the complaint.	0	0
15. Asks which factors or situations triggered the complaint.	0	0
16. Asks which factors or situations increase the complaint.	0	0
17. Asks which factors/situations maintain the complaint.	0	0
18. Asks which factors/situations decrease and/or eliminate the complaint.		
19. Asks which life circumstances or problems accompany the complaint.	0	0
20. Explores the gains of the complaint.	0	0
21. Explores both somatic and psychological determinants of the complaint.	0	0
22. Explores the quality of the relationships within the family/primary group.	0	0
23. Explores current professional functioning.	0	0
24. Explores functioning during leisure time.	0	0

25. Explores risk and vulnerability factors in the patient's biography.	0	0
26. Asks about illnesses and mental health problems in the past.	0	0
27. Asks about professional treatment and its effects in the past.	0	0
28. Asks about other current professional consultions.	0	0
29. Asks about (ab-)use of medication and substances.	0	0
30. Asks about hereditary or family aspects of the complaint.	0	0
31. Reviews the system pertaining to the main complaint.	0	0

III. PRESENTING SOLUTIONS

	YES	NO
32. Explains diagnosis or problem-definition understandably.	0	0
33. Explains causes of the complaint.	0	0
34. Gives information on prognosis of the complaint.	0	0
35. Explores the patient's expectations concerning solutions.	0	0
36. Proposes solutions.	0	0
37. Explains how the solution is appropriate to the problem.	0	0
38. Discusses the pros and cons of the proposed solutions.	0	0
39. Explores whether the patient has a different point of view on problem-definition and/or proposed solutions and discusses any different opinion.	0	0
40. Asks whether the patient is intending to comply.	0	0
41. Explains in concrete terms how the advice given should be carried out.	0	0
42. Checks whether the patient has understood the advice given.	0	0
43. Makes appointments on the follow-up.	0	0

IV. STRUCTURING THE INTERVIEW

	YES	NO
44. Introduces him/herself at the beginning of the interview and clarifies his functional relationship with the patient.	0	0
45. Offers an agenda for the consultation.	0	0

46. Concludes the exploration of the reason for encounter with a summary.	0	0
47. Concludes the "history-taking" with an ordering of the main results.	0	0
48. Explores the reason for encounter before history-taking.	0	0
49. Completes the exploration of the reason for encounter and the history-taking sufficiently before presenting solutions.	0	0
50. Begins presenting solutions with an explanation of the problem-definition.	0	0
51. Asks at the end of the interview if the main problems have been discussed satisfactorily.	0	0

V. INTERPERSONAL SKILLS

	YES	NO
52. Facilitates the communication.	0	0
53. Reflects emotions properly.	0	0
54. Reacts properly to emotions which are directed towards him/herself as a physician.	0	0
55. Asks the patient about his feelings during the interview.	0	0
56. Makes, when necessary, meta-communicative comments.	0	0
57. Performs the history-taking and the review of systems properly.	0	0
58. Puts the patient at ease when necessary.	0	0
59. Sets the proper pace during the interview.	0	0
60. Physician's non-verbal behavior agrees with his/her verbal behavior.	0	0
61. Makes proper eye-contact with the patient.	0	0

VI. COMMUNICATIVE SKILLS

	YES	NO
62. Uses closed-ended questions properly	0	0
63. Concretizes at the proper moment.	0	0
64. Makes proper summaries.	0	0
65. Provides information in small amounts.	0	0
66. Checks whether the patient has understood the information.	0	0
67. Makes, when necessary, proper confrontations.	0	0
68. Uses comprehensible language.	0	0

I. EXPLORING REASONS FOR ENCOUNTER

General remarks:

During the exploration of the reasons for encounter, the patient is invited to talk about the reason for the visit; the symptoms and/or complaints; the attributions and emotional impact of the complaints; the way of coping with these problems in interaction with important others. The physician tries to obtain information about the complaints and/or symptoms stated within the patient's frame of reference. This requires a facilitating and listening attitude of the physician who asks the questions in an open manner.

Item 1: **Asks for the reason for encounter.**
This item refers to questions about the reason for the visit to the physician. Open questions like "What can I do for you?" are intended in this item. These opening questions are very general and the answers of the patient may be very divergent. Patients may mention some complaints; they may say that they have been sent by their family or by a colleague; they may ask for a prescription or some certificate; in case of a visit to a GP, they may ask for referral.
Scoring: "Yes" in the case of an open question concerning the reason for encounter.

Item 2: **Explores the emotional impact of the complaint/problem.**
This item refers to the physician's interviewing behavior by which he explores the emotional impact of the complaint or problem of the patient. Emotions, worries, anxiety, concerns, thoughts etc. of the patient about the complaint are intended. Questions like "How do you feel about this problem?" or reflections on the emotional dimension in the patient's information may be expected. Patients are often anxious about the prognosis of their complaint or problem and they may sometimes present feelings of guilt and shame, especially in the case of mental health problems.
Scoring: "Yes", if the physician explores the patient's emotions concerning his main complaint or problem.

Item 3: **Asks the patient to clarify why he is presenting this problem at this particular moment.**
This item asks which immediate motives have effectuated the decision to initiate the medical consultation. The answer to this question yields information about the factors which forced the patient to seek such help. It yields, moreover, an impression of the severity of suffering. If there has been a need for help for a long time and the patient or his important others have not asked for help, the physician can explore the factors which have delayed help-seeking. Feelings of guilt and anxiety may interfere with the decision to initiate a medical consultation.
Scoring: "Yes", when this subject is explored by means of an open question.

Item 4: **Asks the patient to give his opinion on what are the causes of the problem.**
Questions like: "What, in your opinion, are the causes of your problem?" are intended. The answer yields information about causal attributions by the patient. Since patients often lack a scientific understanding of the causes of their complaints, they will construct a theory based on lay information and prior experience. The verbalization of the patient's personal constructs provides additional insight and enables modification of the constructs towards a usually more realistic view of the complaints.
The exploration of the attributions of the patient contributes significantly to the favorable climate of the medical consultation. It enhances an atmosphere of trust and understanding.
Scoring: "Yes", when the physician asks, by means of an open question, about the patient's causal attributions of the complaints.

Item 5: **Asks how the complaint or problem is discussed within the family or primary group.**
By means of this item, is examined whether the problem is discussed with family members or other important others and how they react: this can be by means of reinforcement, defense, help, persuasion to initiate medical consultation, etc.
Scoring: "Yes", only when both of these aspects are examined otherwise "no" is scored.

Item 6: **Asks the patient to state what help (s)he desires.**
This item deals with the kind of help the patient wishes to receive. Although these wishes may have unrealistic aspects because too much regarding the solution of the problems is expected from the physician, the physician must have an insight into these wishes. The physician has to meet the wishes of the patient as much is possible during the management plan he offers the patient. In this respect, the difference between wishes and expectations concerning help is of importance. The patient may, for instance, wish for a management plan A but is expecting, on the grounds of previous experience with the physician, that not management plan A but management plan B will be offered.
Scoring: "Yes", when the wishes of the patient are asked for explicitly with regards to the help that is desired.

Item 7: **Asks how the patient has tried to solve the problem by him/herself.**
By means of this question, the physician explores what treatment has been adopted by the patient himself in order to get relief from his/her complaint, be it with or without success. The answer may be, for instance: self-medication, changing in life patterns or habits.
Item 28 concerns professional treatment.
Scoring: "Yes", when this question is posed in an open way.

Item 8: **Explores the influence of the complaint on daily life.**
This item deals with the concrete consequences of the complaint or problem for daily life. The behavior aspects intended by this item have a close relationship with the emotional aspects of item 3. The emotional impact and the behavioral consequences of the complaint for daily life may give an insight into the amount of subjective suffering of the patient.
Scoring: "Yes", when the physician inquires about these consequences by means of an open question.

II. HISTORY-TAKING

General remarks:

In this section of the interview, the physician explores the main complaint according to his/her medical frame of reference. These questions often have a closed character in order to provide the exact information necessary for the physician's diagnostic and problem-solving process.

Item 9: **Asks the patient to describe the complaint.**
This item is scored subjective when the physician asks, by means of an open question, for a description of the complaint that formed the incentive for the patient to visit the physician. The patient can have somatic complaints and/or mental health problems.
Scoring: "Yes", when the physician asks for a description of the complaints by means of an open question.

Item 10: **Explores the intensity of the complaint.**
The physician asks for a subjective description of the intensity of the complaint which provided the motive for the consultation. The intensity is an important aspect of the complaint: it may provide an estimate of the degree of suffering of the patient. Intensity often becomes evident from the impact of the complaint on the patient's behavior. For instance, a stabbing headache may hinder physical exertion; a depression may vary from a low mood after a disappointment with few implication for daily life to a psychotic depression which profoundly influences the emotional and thought processes.
Scoring: "Yes", when the physician asks about the intensity of the complaint.

Item 11: **Asks about the localization of the complaint.**
Scoring: "Yes", when the physician asks about the localization of the complaint.

Item 12: **Asks about shifts/radiations of the complaint.**
Scoring: "Yes", when the patient is asked about the localization, shifts and radiation of the complaint.

Item 13: **Asks about the course of the complaint during the day.**
Scoring: "Yes", when the physician inquires about the "time-intensity graphic" during the time cycle of one day.

Item 14: **Asks about the history of the complaint.**
By this item is meant the gathering of information about the start of the complaint; any fluctuations; any complaint-free intervals; any change in character and intensity of complaint during life-time.
Scoring: "Yes", when the physician asks about one or more of these four aspects of the history of the complaint/problem.

Item 15: **Asks which factors or situations triggered the complaint.**
This item scores the physician's questioning behavior in the search for internal or external factor(s) which elicited the complaint. Provoking factors from history and present time may be revealed. These questions form the "interviewing correlate" of the physician's clinical problem-solving process.
Scoring: "Yes", when the physician searches for provoking factors in past and present.
NB: The quality of the clinical problem-solving process and hypotheses is not judged; only the presence of the "search behavior" of the physician is judged. This remark is also valid for the following 3 items.

Item 16: **Asks which factors or situations increase the complaint.**
The physician, using open or closed (directive) questions, asks about factors that increase already existing complaints/problems.
Open questions will be asked when the physician has no clear hypotheses; closed or directive questions will be asked to test existing hypotheses.
Scoring: "Yes", when open or closed (directive) questions are put in order to analyze factors that increase the problems/complaints.
See also the final remark of item 15.

Item 17: **Asks which factors/situations maintain the complaint.**
Scoring: "Yes", when the physician asks for complaint maintaining factors by means of open or closed (directive) questions. See also the final remark of item 15.

Item 18: **Asks which factors/situations decrease and/or eliminate the complaint.**
Scoring: "Yes", when the physician, by means of open or closed (directive) questions, asks about factors that decrease or eliminate complaints.
Note also the final remark of item 15.

Item 19: **Asks which life circumstances or problems accompany the complaint.**
In this item, the patient is asked to talk about the problems, complaints or life-circumstances which accompany the main complaint or problem. The objective is to inquire about temporary relationships which are considered to exist between events and complaints from the patient's point of view. The answer to this open question may be: an other important complaint; a stressful life event that influences the complaint; a totally different problem which has no connection with the main complaint or problem etc.
Scoring: "Yes", when this question is posed in an open manner.

Item 20: **Explores the gains of the complaint.**
The physician checks whether the complaints have a function in the illness behavior of the patient in the sense of secondary gains of the illness. This is done in two steps:
Firstly, the physician asks how important others have reacted to the patient's illness/complaints. This issue is scored in item 5.
Secondly, the physician explores the function that this reaction can have for the patient. A possibility is an excuse function; diminished responsibility; diverting attention from other problems or the control of communication patterns within the patient's system (rigidity).
Scoring: "Yes", when the physician expores the function that the reaction of important others has to the patient.

Item 21: **Explores both somatic and psychological determinants of the complaint.**
This item is scores "yes", when the physician asks:
1) in case of a "pure" somatic problem, some open screening questions about its influence on psychosocial functioning, and
2) in case of a "pure" mental health problem, some (open) screening questions about the quality of physical functioning.

Scoring: "Yes", when one of these situations is present.

Item 22: **Explores the quality of the relationships within the family/primary group.**
As features of these relationships the following aspects may be considered: flexibility in the case of changing situations;

flexibility of roles and positions; differentiation of roles and tasks; possibilities of emotional and social support; flexibility and tolerance in norms and values.
Scoring: "Yes", when two or more of these features are explored.

Item 23: **Explores current professional functioning.**
By "professional functioning" is meant profession, household or study.
Scoring: "Yes", when the physician asks for the experienced quality of one of these three aspects.

Item 24: **Explores functioning during leisure time.**
Scoring: "Yes", when the physician explores satisfaction in functioning during leisure time.

Item 25: **Explores risk and vulnerability factors in the patient's biography.**
This item concerns long standing factors of vulnerability in relation to the main complaint/problem.
The following factors from the patient's biography may be considered:
- genetric or constitutional deficiencies or handicaps (mental and physical)
- periods of social dysfunctioning
- risky life styles that enhance vulnerability
- periods of emotional, cultural or material deprivation
- traumatic and/or stressful life events.

This item should be differentiated from the next item, in which information requested about past illnesses or mental health problems. Some overlap will be inevitable.
Scoring: "Yes", when 2 or more of these 5 categories are explored.

Item 26: **Asks about illnesses and mental health problems in the past.**
The physician asks for a "historical" picture of illnesses and mental health problems. A relationship with the present complaint/problem is not necessary.
Scoring: "Yes", when the physician asked about illnesses and mental health problems in the past.

Item 27: **Asks about professional treatment and its effects in the past.**
This item asks, in contrast to the self-care which is asked about in item 7, about the way the patient has presented his problems/complaints to other professionals in the past and the effects of their treatment.
Scoring: "Yes", when the physician pays attention to both the kind of treatment and the effects of treatment.

Item 28: **Asks about other current professional consultations.**
This item refers to the exploration of consultations, diagnostic investigations and treatment that are being carried out currently but that are (not) related to the main problem or complaints. By "professional" is meant (para)medical disciplines as well as professional "alternative healers". Overlap with item 7 pertaining to self-help may arise when the patient applies prescriptions or advice that have been given in the past to the current complaint on his own initiative.
Scoring: "Yes", when the patient is asked about current professional consultations (not) related to the main complaints.

Item 29: **Asks about (ab-)use of medication and substances.**
Medication and substances are:
- self-medication
- professionally-prescribed medication
- drugs, e.g. smoking, alcohol
- soft- and hard drugs

Scoring: "Yes", when the physician explores the 4 following aspects:
- what medication is used
- what drugs are used
- their quantity
- the degree of dependency

Item 30: **Asks for hereditary or family aspects of the complaint.**
Scoring: "Yes", when this question is asked.

Item 31: **Reviews the system pertaining to the main complaint.**
The physician should deal intensively with the system pertaining to the main complaint such as the cardio-vascular system in the respiratory-tract. When the MAAS is used in evaluation settings with simulated patients, an elaborated checklist of questions pertaining to a specific problem can replace this item. In this case, non-medically trained observers are not able to score this item.
Scoring: "Yes", when the physician asked 80% or more of the common questions pertaining to the tract of the main complaint.

III. PRESENTING SOLUTIONS

General remarks:

This section is defined as the answer to the request for help within the context of initial medical interviews. In the previous sections, the physician has clarified the reason for the visit and the request for help. Moreover, he has collected systematic medical information for his diagnostic and problem-solving hypotheses. Further evaluation of these hypotheses is based on data from the physical examination

which is not part of this instrument.
Presenting solutions is the primary reaction of the physician to the patient following the first two sections of the initial medical interview. This solution is based on data gathered during the work-up and on prior knowledge of the patient and his/her social orbit in the case of general practitioners with relationships with the patient. The "presenting solution" section consists of items concerning the quality of information exchange about diagnosis, prognosis, etiology and the negotiation about the further work-up (further investigation, treatment, referral etc.). This scale is intended to measure interview behavior performed during this process; the quality or adequateness of the given advice is not assessed.

Item 32: **Explains diagnosis or problem-definition understandably.**
The physician mentions a (probable) diagnosis or any other definition of the problem, sometimes an important "rule-out". The information is mainly on a descriptive level. For instance: "The symptoms are indicative of measles". Etiology is not considered here but in the next item (for instance: "Measles are a contagious infection caused by a virus"). Nevertheless, there may be some overlap with the next item. If the physician does not have any (probable) diagnosis but has made some important "rules-out" (e.g. "The chest pain probably does not mean a heart disease"), these statements are also valid in this item. It is important that the physician does not use any jargon or terms which do not match the intellectual or/and socio-cultural level of the patient.
Scoring: "Yes", when the physician gives descriptive information about what is wrong (or not wrong) in terms understandable to the patient.

Item 33: **Explains causes of the complaint.**
This item means an explanation of the complaint in terms of patho-physiological mechanisms. As in the former item, the patient must be able to understand the explanation.
Scoring: "Yes", if etiological explanation and understandability are present in the given information.

Item 34: **Gives information on prognosis of the complaint.**
The prognosis of a disease may be strongly influenced by medical treatment. The physician should include minimally some information about the severity of the illness consisting of a description of the natural course of the disease. In addition, information about the disease after treatment should be given.
Scoring: "Yes", when the physician gives information about the course of the disease with and without treatment.

Item 35: **Explores the patient's expectations concerning solutions.**
Expectations always have a factual and emotional aspect. We draw attention to the difference between wish and expectation as explained in item 6. In short, wishes tend to reflect unrealistic hopes, whereas expectations are hopes which have been moulded by reality. For example: "The patient wishes to obtain a thorough explanation from the physician but he expects to receive only a prescription which does not resolve all his worries or fears. This item must be distinguished from item 6 although some overlap may be inevitable.
Scoring: "Yes", when both factual and emotional aspects of the expectation concerning solutions for the problems are explored.

Item 36: **Proposes solutions.**
This proposal may consist of further history-taking, further investigations (with or without referral), treatment or preventive advice. The physician may offer possible alternatives, of which one is always "no further professional help". This gives the patient the opportunity to make a choice for which (s)he takes responsibility.
Scoring: "Yes", if the physician introduces a proposal for help with one or more alternatives.

Item 37: **Explains how the solution is appropriate to the problem.**
Scoring: "Yes", if the physician offers this explanation (related to the problem stated in item 32) in an understandable way.

Item 38: **Discusses the pros and cons of the proposed solutions.**
Information about pros and cons is intended to support the patient by choosing the most suitable solution to his problem. The pros and cons must be weighted against the patient's situation in order to result in a feasible plan. Pros and cons may be: adverse and beneficial effects *(medication); estimated* probability of success or failure; impact on daily life; social restrictions; cost; waiting list etc.
Scoring: "Yes", when the physician discusses at least one pro and con of the proposed help with the patient.

Item 39: **Explores whether the patient has a different point of view on problem-definition and/or proposed solutions and discusses any different opinion.**
While giving information, it may become clear that the patient does understand the physician, but holds a different opinion on some part. The physician should then check this difference in point of view explicitly. The physician should discuss the possible difference in view-point and clarify the exact point of difference. By "discusssing", we do not mean arguing with the patient about his point of view. This item refers only to a discussion of different points of view. When arguing takes place and the physician obliges the patient to adopt his opinion, this item will be scored "no".
Scoring: This item is scored "yes" when the physician explores the

presence of a different point of view on problem-definition or treatment plan and when possible differences in points of view are clarified.
Scoring: "No", when the physician does not check any possible differences in points of view or when he tries to persuade the patient to change his opinion.

Item 40: **Asks whether the patient is intending to comply.**
This ends the phase of negotiation while giving advice. The physician explores whether the patient intends to comply with the given advice.
Scoring: "Yes", when the physician asks this question.

Item 41: **Explains in concrete terms how the advice given should be carried out.**
After reaching a decision about the advice, the physician should explain how this advice has to be followed. If the explanation is of good quality, it will increase the patient's compliance and therapeutic results. Good advice must always be given in terms of concrete behavior so that it enables the patient to carry out the advice.
For example: When "rest" is prescribed, it should be clear whether the patient should sleep longer, should seek situations of relaxation, should avoid conflicts, or should stay in bed day and night with someone to look after him.
Scoring: "Yes", when, acccording to the observer's opinion, the advice given is formulated concretely enough for the patient to follow adequately.

Item 42: **Checks whether the patient has understood the advice given.**
After having made clear in terms of concrete behavior how the patient should follow the advice, the physician should check whether the patient has understood the advice. There are several ways to do this: the physician can ask the patient if he has understood the advice or he can ask the patient to repeat the advice.
Scoring: "Yes", when the physician makes sure that the advice given is understood by the patient.

Item 43: **Makes appointments for the follow-up.**
The following subjects can be arranged:
- what is going to happen
- who is doing what
- who takes initiatives
- at what time

Scoring "Yes", when all four subjects are treated concretely, otherwise "no".

IV. STRUCTURING THE INTERVIEW

General remarks:

In the scale "structuring the interview" is traced how the different phases of an initial consultation link up. Since prior studies have revealed that the reliability of this scale declines when different stages are being followed incompletely or chaotically, observers should keep to the scoring-instructions literally.

Item 44: **Introduces him/herself at the beginning of the interview and clarifies his functional relationship with the patient.**
This item relates to 2 subjects, namely, the introduction and the clarification of the relationship. Since the MAAS is used in a variety of situations, it can be the case that one or either subject is not applicable, because they are already known to the patient, e.g. in the practice of a general practitioner. Clarifying the functional relationship is not only important for physicians and students in training-situations but also for physicians communicating with patients via different functional connections such as physicians in examining advising relationships; physicians standing in temporarily; or physicians cooperating in different situations all of whom have to explain their functional relationship with the patient.
Scoring: "Yes", only if both introduction and clarification actually take place.

Item 45: **Offers an agenda for the consultation.**
Having clarified the reason for the encounter, the physician frames a plan in which he explains how he wishes to handle this request for help from the patient. This design for the rest of the interview includes the subjects the physician wishes to bring under discussion, the physical examination he wishes to perform and the intended sequence.
Scoring: "Yes", only if such a plan is offered to the patient.

Item 46: **Concludes the exploration of the reason for encounter with a summary.**
This item checks whether the reason for encounter is well understood by the interviewer and is done by means of a summary. We wish to emphasize that a summary always has a testing character. The physician invites the patient to react to the summary by means of the content of the summarized information and the interval afterwards. If the reason for encounter is understood sufficiently, the physician will continue by asking more directive questions stemming from the system of the present illness. Theoretically, the following possibilities can occur with regard to this item:
- The request for help is implicit (the patient comes for the results of investigations done before) or the request is being verbalized by the patient spontaneously. In both cases, a summary has still to

be made in order to test whether the physician has understood the request well. In the first case, a (closed-ended) question like "Did you come for the results" will suffice.
- The request for help has not (yet) been verbalized properly by the patient which means that the summary comes too quickly and is, by definition, incomplete. This becomes obvious because the patient adds further information to the summary.

When the request for help has not been explored sufficiently, the exploration has to be continued and has to be concluded with a summary again. This item measures the concluding function of the summary. The quality of the summary is assessed by item 58.
Scoring: "Yes", if a summary is made at the end of the exploration of the reason for encounter or if the summary is perfected after another try.

Item 47: **Concludes the "history-taking" with an ordering of the main results.**
The physician mentions the main problems which have come up during the exploration of the reasons for encounter and history-taking phases. This summary originates from the physician's frame of reference and differs in this respect from a "summary" which originates from the patient's frame of reference. With the ordering, the physician closes the first two phases of an initial interview.
Scoring: "Yes", when the information is ordered after the exploration of the reasons for encounter and the directive questions pertaining to the history-taking section.

Item 48: **Explores the reason for encounter before history-taking.**
In the medical interview, physicians frequently switch between different sections which appears to have a confounding effect on the patient. Ideally, the section with history-taking questions is preceded by the exploration of the request for help.
Scoring: "Yes", if, according to the observer's opinion, the request for help has been explored sufficiently before the physician continues with the history-taking section.

Item 49: **Completes the exploration of the reason for encounter and the history-taking sufficiently before presenting solutions.**
Scoring: "No", if one of the following cases is present:
- The physician goes back to items from former phases when parts of the presentation of solutions have already been under discussion.
- The phases of the exploration of reasons for encounter or history-taking have not been elaborated extensively enough in the observer's opinion.

In all other cases, score: "yes".

Item 50: **Begins presenting solutions with an explanation of the problem-definition.**
Scoring: "Yes", when the physician introduces this phase with information about a probable diagnosis, a problem-definition or an important rule-out of the problem/complaint.
"No", if this does not happen.

Item 51: **Asks at the end of the interview if the main problems have been discussed satisfactorily.**
Scoring: "Yes", when the physician asks this question.

V. INTERPERSONAL SKILLS

General remarks:

This section is scored on a 3-point-scales: yes - indifferent - no. The extension of the scale provides the opportunity of giving a more qualitative assessment of physician's interpersonal skills.

Item 52: **Facilitates communication.**
This item requires the observer to give a global judgment of the quality of the physician's facilitating behavior during the interview. Facilitating is necessary to stimulate the patient to speak from his own frame of reference and experience and to express emotions.
In addition to the exploration of emotions, it is also important to ask questions about facts in a facilitating way. Facilitation is given concrete form in the following ways:
- well-indicated open questions, especially during the exploration of the reason for encounter and the exploration of emotions during the presenting of solutions;
- stimulating questioning behavior within the patient's frame of reference;
- remarks stimulating to openness;
- a listening attitude which becomes apparent by means of well-timed, short periods of silence;
- physican's self-disclosure.

Scoring:
- "Yes", when at least 4 aspects are shown.
- "Indifferent", when 2 or 3 different aspects are shown.
- "No", when 1 or fewer aspects are shown.

Item 53: **Reflects on emotions properly.**
This item covers reflections about verbally or non-verbally expressed emotions by the patient. Reflections form the most important interview behavior for the physician to react to the patient's emotions. Reflections on emotions are used in the proper way, when:

- They are timed adequately, which means at the moment or directly after the emotions are expressed. The relation between the patient's emotion and the reflection has to be clear.
- Non-verbally expressed emotions are recognized and reflected upon.
- The right content of the emotion is reflected, which means congruence between emotion and reflection.

Scoring:
- "Yes", when reflections of emotions are used in 80% according to the criteria.
- "No", when less than 80% of the reflections of emotions are used according to the criteria or when the physician does not react to clearly expressed emotions.
- "Indifferent", when no reflections on emotions are used, and also when they are not necessary.

Item 54: **Reacts properly to emotions which are directed towards him/herself as a physician.**

This item refers to the physician's reactions to emotional expressions of the patient that are directed against the physician. When the patient expresses sadness, disappointment, anxiety, anger, blame or cynicism against the physician, (s)he has to try to keep the communication on-going. The communication can be disturbed when the physician does not handle the emotions well by presenting different defense mechanisms against the emotions such as
- denial, negotiation, minimizing, rationalization, shifting, reacting by the counterpart, etc.;
- using antagonistic behavior (discussion, quarrel, etc.).

Scoring:
- "Yes", when the physician handles emotions which are directed towards him in the appropriate way with the result that the communication keeps going.
- "No", when the physician uses defense mechanisms or antagonistic behavior.
- "Indifferent", when the patient does not express emotions which are directed towards the physician.

Item 55: **Asks the patient about his feelings during the interview.**

This item refers to the physician's questions about the feeling the patient has during the interview. The questions are most likely to occur during the presentation of solutions. The questions have the characteristics of open-ended questions and pertain to the momentary feelings and emotions of the patient. Open-ended questions are asked in a proper way when:
- the physician asks questions within the patient's frame of reference;
- the question does not rule-out any categories for answering;
- each question deals with one subject.

Scoring:
- "Yes", when these questions are asked in the appropriate way in 80% or more of the cases.

- "No", when these questions are not asked in the approporiate way in 50% or less of the cases.
- "Not/indifferent", when this interview behavior is not shown or handled appropriately in only 50-80% of the cases.

Item 56: **Makes, when necessary, meta-communicative comments.**
The physician makes meta-communicative comments to stimulate an inhibited communication. Inhibition of communication may have several causes but results mostly from inadequate interview behavior performed earlier in the interview and which was not corrected.
Examples:
- neglecting or minimizing strong emotions;
- inadequate reassurance;
- asking questions which have nothing to do with the case;

Inhibited communication is expressed and can be detected by several characteristics of the physician-patient communication:
- defensive behavior of the patient (negativism, denial, refusal);
- obstinate discussion;
- frequent misunderstanding;
- long periods of silence;
- repetition.

The result is that communication is hampered in the course of several phases of the medical interview. The communication can be stimulated by meta-communicative comments, like: "It seems that we are going round in circles here", or, "How can it be that we frequently misunderstand each other?".
Scoring:
- "Yes", when inhibited communication is stimulated by meta-communicative comments.
- "No", when in the case of inhibited communication, the physician does not make meta-communicative comments, or when he makes unnecessary meta-communicative comments which have an inhibiting influence on the communication.
- "Indifferent", when meta-communication is not shown, and is not necessary.

Item 57: **Performs the history-taking and the review of systems properly.**
Directive medical questions during history-taking should not lead to endless rows of questions as they can stimulate feelings of uncertainty and anxiety and are likely to be misunderstood.
Scoring:
- "Yes", when the physician briefly explains why (s)he wants to ask a number of directive questions and when these questions do not take up too much time and attention in the observer's opinion.
- "No", when directive medical questioning does not fulfil <u>both</u> criteria.
- "Indifferent", when there is no history-taking.

Item 58: **Puts the patient at ease when necessary.**
This item refers to specific, explicit behavior which is aimed at putting the patient at ease. It can be necessary to put the patient at ease:
- to make acquaintance with the physician;
- during physical examination;
- after the expression of strong emotions during the presentation of solutions;

Scoring:
- "Yes", when the physician shows explicit behavior which is meant to put the patient at ease.
- "No", when this behavior is necessary but the physician fails to perform it in the observer's opinion.
- "Indifferent", when such behavior is not necessary and is not shown.

Item 59: **Sets the proper pace during the interview.**
This item asks for a global judgment on an important quality of the interview. The pace of an interview is strongly related to facilitation behavior and to "directivity" of the physician, and is considered as such to be an important quality of an interview.
Scoring:
- "No", when:
  - there are periods of silence which disturb the pace of the interview;
  - the physician jumps too quickly from one subject to an other;
  - the physician interrupts the patient;
  - the physician allows the patient too much discussion of subjects which are not of evident importance for the presented complaint/problem.
- "Yes", when the physician regulates the pace of the interview smoothly.
- "Indifferent", when there is a mixture of "proper" and "improper" pace.

Item 60: **Physician's non-verbal behavior agrees with his/her verbal behavior.**
This item is best scored by first judging the non-verbal behavior of the physician and then comparing the nature of the verbal behavior with that of the non-verbal behavior. Afterwards, the observer judges whether or not they agree.
Cues for non-verbal behavior are:
- look/eye-contact;
- tone of voice;
- expression;
- body expression;
- gestures.

Scoring:
- "Yes", when the non-verbal behavior agrees with the verbal behavior.

- "No", when incongruent behavior is present in the interview.
- "Indifferent", when the observer finds it impossible to decide either yes or no.

Item 61: **Makes proper eye-contact with the patient.**
Scoring:
- "No", when:
  - the physician avoids eye contact or continues to gaze at his file or at some other object;
  - the physician gazes continuously at the patient.
- "Yes", when normal eye-contact is maintained.
- "Indifferent", when no judgment is possible (for instance, in case of an unsuitable camera position in videotaped consultations).

VI. COMMUNICATIVE SKILLS

General remarks:

This section is scored on a 3-point scale: yes - indifferent - no. The extension of the scale provides the opportunity of giving a qualitative assessment of physician's interpersonal skills.

Item 62: **Uses closed-ended questions properly.**
The physician asks closed-ended questions in a proper way when:
- the question does not contain a suggestion for an answer;
- the question deals with one subject only;
- this type of question is used on the proper indication;

Closed-ended questions are indicated when:
- the physician searches for factual information;
- the patient deviates from the subject;
- the patient resists the discussion of a subject.

Closed-ended questions are not indicated when:
- there is a chance that the physician will miss the relevant answer by limiting the answer categories;
- they are used instead of open questions during the clarification of the request for help or the exploration of emotions in general.

Scoring:
- "Yes", when 80% of all closed-ended questions are used in the proper way.
- "Indifferent", when 60-80% of all closed-ended questions are used in the proper way.
- "No", when less than 60% of all closed questions are used in the proper way.

While scoring the item, it can be helpful to use the scoring stave on the scoring list. Each closed-ended question can be scored right or wrong. At the end of the interview, the total item can be scored.

Item 63: **Concretizes at the proper moment.**
Concretization is necessary when the patient speaks in a vague, impersonal, general or unclear way about subjects related to the complaint. The physician invites the patient to reexpress himself in a more clear, personal and specific way. If one of these aspects is evident, then the intervention is done in the proper manner.
Scoring:
- "Yes", when the physician concretizes in the proper manner and in an appropriate situation.
- "No", when the physician does not concretize when it is necessary, or when he does not concretize in the proper manner, or when he concretizes too much.
- "Indifferent", when it is not necessary to concretize and it is not done.

Item 64: **Makes proper summaries.**
A summary is a restatement of important information given by the patient but verbalized in the physician's own words.
A summary is close to the patient's frame of reference, in contrast with ordering, which stems from the physician's frame of reference. In this item, the observer makes a judgment on the proper content of the summary.
Scoring:
- "Yes", when 80% or more of the summaries are an appropriate restating of the content of the patient's utterances.
- "No", when 60% or less of the summaries restate the content of the patient's utterances appropriately.
- "Indifferent", when this interview behavior is not shown or when 60-80% of the summaries are appropriate.

Item 65: **Provides information in small amounts.**
During the presentation of solutions, the physician provides the patient with information which has to be understood and remembered. Recall of information can be stimulated by providing information in small amounts. Small amounts are considered to be two or three sentences.
Scoring:
- "Yes", when 80% or more of the information is provided in small amounts.
- "No", when less than 80% of the information is provided in small amounts.
- "Indifferent", when no information is provided.

Item 66: **Checks whether the patient has understood the information.**
After providing information about diagnosis, causes, prognosis and treatment plan, the physician has to check whether the patient has understood the information.

Scoring:
- "Yes", when the physician checks whether the patient has understood the information 3 or more times.
- "Indifferent", when the physician checks whether the patient has understood the information once or twice.
- "No", when the physican does not check.

Item 67: **Makes, when necessary, proper confrontations.**
A physician's ability to make proper confrontations is measured in this item. "Proper" refers to situations in which confrontations are necessary because communication is inhibited by contradictions. This situation occurs when:
- there are contradictions in the patient's words;
- there are contradictions between the patient's words and his non-verbal behavior;
- there are contradictions between the past and present behavior of the patient.

Scoring:
- "Yes", when the physician makes proper confrontations which stimulate the communication.
- "No", when the physician fails to make proper confrontations and the communication remains hampered or when the physician makes unnecessary confrontations which inhibit the communication.
- "Indifferent", when the behavior is not shown and is not necessary.

Item 68: **Uses comprehensible language.**
Scoring:
- "Yes", when comprehensible language is used during the interview.
- "Indifferent", when this category is not applicable in this item.
- "No", when according to the observer's judgment, several difficult words (medical jargon, words from a different social class) are used, or when problems arise from using dialect or unsuitable dialects.

APPENDIX B

THE MAASTRICHT HISTORY-TAKING AND ADVICE CHECKLIST (MAAS-PMHC)

AN OBSERVATION INSTRUMENT FOR THE MEASUREMENT OF THE PHYSICIAN'S INTERVIEWING SKILLS IN INITIAL CONSULTATIONS IN PRIMARY MENTAL HEALTH CARE

MANUAL FOR SCORING

BY

H.F. KRAAN, A.A.M. CRIJNEN, J. ZUIDWEG AND J. VAN DALEN

MAAS-PRIMARY MENTAL HEALTH CARE

LIST OF ITEMS

I. EXPLORATION OF THE REASON FOR ENCOUNTER

	YES	NO
1. Asks for the reason for visit.	0	0
2. Asks the patient to decribe his complaint/problem.	0	0
3. Explores the emotional impact of the complaint/problem.	0	0
4. Asks the patient to clarify why (s)he is presenting this problem at this particular moment.	0	0
5. Asks the patient to give his opinion on what are the causes of the problem.	0	0
6. Asks how the complaint/problem is discussed with the family or primary group.	0	0
7. Asks how the patient has tried to solve the problem by him/herself.	0	0
8. Explores the consequences of the complaint/problem for daily life.	0	0
9. Asks what life circumstances or what other problem accompany these complaint/problem.	0	0
10. Asks for the ways in which the patient has usually resolved similar problem in the past.	0	0
11. Asks the patient whether the complaint/problem might be a burden to others.	0	0
12. Asks about recent life-events.	0	0
13. Asks the patient to state what help (s)he desires.	0	0

II. HISTORY-TAKING

	YES	NO
14. Explores the intensity of the complaint/problem.	0	0
15. Asks about the course of the complaint during the day.	0	0

16. Asks about the history of the complaint/ problem.	0	0
17. Analyses the (causal) factors which provoked the complaint/problem.	0	0
18. Analyses the factors which increase the complaint/problem.	0	0
19. Analyses the factors which maintain the complaint/problem.	0	0
20. Analyses the factors which decrease and/or eliminate the complaint/problem.	0	0
21. Explores the functionality (gains) of the complaint/problem.	0	0
22. Asks about illnesses and/or mental health problems in the past.	0	0
23. Asks about current professional treatment and its effect in the past.	0	0
24. Asks about other current professional consultations.	0	0
25. Asks about current (ab)use of medication.	0	0
26. Asks for (pseudo) hereditary aspects of the complaint/problem.	0	0

III. PSYCHIATRIC EXAMINATION

	YES	NO
Examines symptoms of affective disorders:		
27. disturbances in mood and affect	0	0
28. biological symptoms	0	0
29. disturbances of thought	0	0
30. suicidal ideation and behavior	0	0
Examines symptoms of anxiety disorders:		
31. character and intensity of anxiety	0	0
32. phobic symptoms	0	0
33. anxiety in-/decreasing factors	0	0
34. consequences of anxiety	0	0
Examines disturbances in consciousness and orientation:		
35. mild disturbances: drowsiness, concentration, disturbances	0	0
36. disturbed orientation (time, place, persons)	0	0

Examines disturbances in memory:

37. immediate retention and recall	0	0
38. recent memory	0	0
39. remote memory	0	0

Examines perceptual disturbances:

40. distinguishes hallucinations from illusions and psycho-hallucinations	0	0
41. character of hallucinations	0	0

Examines disturbances of thought:

42. disturbances in the stream of thought	0	0
43. disturbances in the content of thought	0	0
44. disturbances in the perception of the own processes of thought	0	0

IV. SOCIO-EMOTIONAL EXPLORATION

	YES	NO
45. Explores feelings of love/affection in interpersonal relations.	0	0
46. Explores aggressive feelings in interpersonal relations.	0	0
47. Explores perspectives and aspirations in life.	0	0
48. Explores care-giving.	0	0
49. Explores feelings of responsibility.	0	0
50. Asks about religious feelings.	0	0
51. Explores character traits and/or self-image.	0	0
52. Explores the quality of the relations with family/primary group.	0	0
53. Asks about social support.	0	0
54. Asks about cultural differences in social relationships.	0	0
55. Examines current professional functioning.	0	0
56. Examines functioning during leisure time.	0	0
57. Explores sexual functioning.	0	0
58. Explores sleep habits.	0	0
59. Explores eating habits.	0	0
60. Explores use of substance.	0	0

61. Explores (dis)satisfaction with housing conditions.	0	0
62. Explores (dis)satisfaction with financial situation.	0	0
63. Explores history of education and professional life.	0	0
64. Explores early developments up to adolescence (traumatic experiences, socio-emotional and psycho-motor development).	0	0

V. PRESENTING SOLUTIONS

	YES	NO
65. Provides a problem-definition or "diagnosis" in understandable terms.	0	0
66. Gives information about causal and maintaining factors of the complaint/ problem.	0	0
67. Gives information about the prognosis of the problem; without and with treatment.	0	0
68. Explores the patient's expectations concerning help.	0	0
69. Explores how much responsibility the patient is prepared to take for his/ her treatment.	0	0
70. Introduces a proposal for help (with alternatives).	0	0
71. Explains how the proposal for help goes with the problem.	0	0
72. Discusses the pros and cons of the proposed help.	0	0
73. Asks for the patient's opinion about the proposed help.	0	0
74. Explores how "important others" might influence the proposed help.	0	0
75. Checks whether the patient has a different point of view on problem-definition and/or proposal for help and discusses any different opinion.	0	0
76. Asks the patient to make a choice from several proposals for help.	0	0
77. Gives concrete information on how given advice should be followed.	0	0
78. Checks whether the patient has understood given advice.	0	0
79. Makes appointments for further follow-up.	0	0

VI. STRUCTURING THE INTERVIEW

	YES	NO
80. Introduces him(her)self at the beginning of the interview and makes his functional relationship clear to the patient.	0	0
81. Offers a plan for the consultation.	0	0
82. Concludes the exploration of the reason for encounter with a summary.	0	0
83. Concludes history-taking, psychiatric examination and socio-emotional exploration with an ordering of the main results.	0	0
84. The exploration of the reason for encounter section precedes history-taking, psychiatric examination and socio-emotional exploration.	0	0
85. Completes exploration for the reason for encounter, history-taking, psychiatric examination to such an extent that solutions can be presented.	0	0
86. Starts "presenting solutions" with information about the problem-definition/diagnosis.	0	0
87. Asks at the end of the interview whether the main problems have been discussed satisfactorily.	0	0

VII. INTERPERSONAL SKILLS

	YES	INDIFF	NO
88. Facilitates the communication with the patient.	0	0	0
91. Reflects in a proper way emotions which are non-verbally or covert-verbally expressed.	0	0	0
92. Reacts adequately to emotions which are directed towards him/herself as a physician.	0	0	0
93. Asks the patient for his feelings at that moment.	0	0	0
97. Makes, when necessary, meta-communicative comments.	0	0	0
99. Takes the medical history and reviews the systems in an appropriate way.	0	0	0
100. Puts the patient at his ease.	0	0	0
101. Sets the proper pace during the interview.	0	0	0

103. The physician's non-verbal behavior is in agreement with his verbal behavior.	0	0	0
104. Makes proper eye-contact with the patient.	0	0	0

VIII. COMMUNICATIVE SKILLS

	YES	INDIFF	NO
89. Uses closed-ended questions in a proper way.	0	0	0
90. Concretises at the proper moment.	0	0	0
94. Summarizes the content of the patient's statements properly.	0	0	0
95. Conveys information in small units.	0	0	0
96. Checks whether the conveyed information has been understood.	0	0	0
98. Makes, when necessary, proper confrontations.	0	0	0
102. Uses language which is understandable to the patient.	0	0	0

MAAS-PRIMARY MENTAL HEALTH CARE

Manual for observers

Guidelines for use.

To avoid useless repetitions in the description of the criteria for the scoring of the 104 MAAS-PMHC items, we refer to the manual of the MAAS-General Practice (Appendix 4.1).

In the table below is shown how the items of the MAAS-PMHC correspond with MAAS-General Practice items. The underlined item numbers in this table are specific to the MAAS-PMHC. Their criteria for scoring is found in this appendix.

ITEM CORRESPONDENCES BETWEEN MAAS-PMHC AND MAAS-GP

	ITEMS MAAS-PRIMARY MENTAL HEALTH CARE	ITEMS MAAS-GENERAL PRACTICE
I. EXPLORATION OF THE REASONS FOR ENCOUNTER	1	2
	2	9
	3	2
	4	3
	5	4
	6	5
	7	7
	8	8
	9	19
	10	-
	11	-
	12	-
	13	6
II. HISTORY TAKING	14	10
	15	13
	16	14
	17	5
	18	16
	19	17
	20	18
	21	20
	22	26
	23	27
	24	28
	25	-
	26	30

	ITEMS MAAS-PRIMARY MENTAL HEALTH CARE	ITEMS MAAS-GENERAL PRACTICE
III. PSYCHIATRIC EXAMINATION	27	-
	to	-
	44	-
IV. SOCIO-EMOTIONAL EXPLORATION	45	-
	to	-
	51	-
	52	22
	53	-
	54	-
	55	23
	56	24
	57	-
	to	-
	64	-
V. PRESENTING SOLUTIONS	65	32
	66	33
	67	-
	68	35
	69	-
	70	36
	71	37
	72	38
	73	-
	74	-
	75	39
	76	-
	77	41
	78	42
	79	43
VI. STRUCTURING THE INTERVIEW	80	44
	81	45
	82	46
	83	47
	84	48
	85	49
	86	50
	87	51
VII. INTERPERSONAL SKILLS	88	52
	91	55
	92	56
	93	57
	97	61
	99	63
	100	64
	101	65
	103	67
	104	68

	ITEMS MAAS-PRIMARY MENTAL HEALTH CARE	ITEMS MAAS-GENERAL PRACTICE
VIII. COMMUNICATIVE SKILLS	89	53
	90	54
	94	58
	95	59
	96	60
	98	62
	102	66

MAAS-PRIMARY MENTAL HEALTH CARE

Instruction for scoring of the items:

I. EXPLORATION OF THE REASON FOR ENCOUNTER

Item 10: Asks about the ways the patient has usually resolved similar problems in the past.
This item asks about the coping mechanisms by means of which the patient has solved similar problems in the past.
Scoring: "Yes", when such an open question is asked.

Item 11: Asks the patient whether the complaint/problem might be a burden to others.
This item is scored "Yes", when the interviewer explores, by means of open questions, how the patient estimates the burden he puts on his primary group.

Item 12: Asks about recent life-events.
This item is scored "yes", when acute traumatic circumstances or other heavily emotional events within the last 3 months are asked about.

II. HISTORY-TAKING

Item 25: Asks about current (ab)use of medication.
This question pertains to current use of medication, prescribed as well as self-medication. There may be some overlaps with item 23 (past treatment) or item 60 (substance (ab)use).
Scoring: "Yes", when the use and abuse of current medication is asked about.

III. PSYCHIATRIC EXAMINATION

Items 27-30: examines symptoms of affective disorders.

Item 27: Disturbances or mood and affect: depressed, low mood; crying; feelings of despair; irritated, aggressive, dysphoric moods; elated moods; hypo-or anaesthesia; loss of interest.

Item 28: Biological ("vital") symptoms of depression: anorexia; weight loss; typical disturbances of sleep; impaired concentration; mood-swings during day-time; psychomotor agitation or retardation; loss of energy; increased activity.

Item 29: Cognitions influenced by affective disturbances; depressive brooding and worrying; excessive feelings of insufficiency, worthlessness and guilt; elated, inflated self-esteem; grandiosity.

Item 30: Suicidal ideation and behavior: strong death-wishes; inappropriate feelings of guilt, despair and insufficiency with a strong, restrictive impact on the perception of reality; powerless aggression towards self and others; concrete, destructive fantasies concerning suicide; recent suicide attempts.

Scoring items 27-30: "Yes", when about 70% of the pertinent depressive and manic content has been asked about.

Items 31-34: examines symptoms of anxiety disorders.

Item 31: Character and intensity of anxiety: shakiness, tension, inability to relax, restlessness, sweating, heart pounding, paraesthesias, dysphorea, dizziness, hot and cold flushes, anxiety, chest discomfort, dry mouth, (generalized anxiety disorder, panic disorder).

Scoring: "Yes", when 70% of these topics are touched upon.

Item 32: Exploration of special phenomena of anxiety, phobic, obsessive and compulsive phenomena?

Item 33: Anxiety in-/decreasing factors?

Item 34: Consequences of anxiety? Avoidance, social isolation, restrictive behavior, etc.

Scoring items 32-34: "Yes", when pertinent questions are asked.

Items 35-36: examines disturbances in consciousness and orientation.

Item 35: Are there mild disturbances: drowsiness, impaired concentration, impaired immediate recall, reduction in clarity of awareness?

Scoring: "Yes", when 70% of these topics are raised.

Item 36: Disturbed orientation in time, place and person?

Scoring: "Yes", when this question is asked.

Items 37-39: examines disturbances in memory.

Item 37: Immediate retention and recall.

Item 38: Short-term memory.

Item 39: Long-term memory.

Scoring, items 37-39: "Yes", when examining questions concerning these topics are posed.

Items 40-41: examines perceptual disturbances.

Item 40: Distinguishes hallucinations from illustions or pseudo-hallucinations.
Scoring: "Yes", when, from the definitions of these phenomena questions are asked to draw a distinction.

Item 41: Character of hallucinations.
Scoring: "Yes", when, in case of hallucinations, the sensory quality is assessed and when, in case of acostic hallucinations, the interviewer asks for their imperative character.

Items 42-44: examines disturbances of thought.

Item 42: Disturbances in the stream of thought, (coherence; goal directedness; logicality).

Item 43: Disturbances in the context of thought (delusions; preoccupation; insight; judgement).

Item 44: Disturbances in the perception of own cognitive process.
Scoring items 42-44: "Yes", when the pertinent questions are asked from more than 70% per item.

IV. SOCIO-EMOTIONAL EXPLORATION

Item 45: Explores feelings of love/affection in interpersonal relations.
Scoring: "Yes", when such feelings originating from the patient have been asked about.

Item 46: Explores aggressive feelings in interpersonal relations.
Scoring: "Yes", when such question(s) are asked.

Item 47: Explores perspectives and aspirations in life.
This item pertains to life cycle problems.
Scoring: "Yes", when open/questions on perspective, aspirations and their fulfilment are asked.

Item 48: Explores care-giving.
Care-giving may be accompanied by satisfaction as well as by too heavy a physical or emotional burden.
Scoring: "Yes", when a question is asked about care-giving with the accompanying effects.

Item 49: Explores feelings of responsibility.
Scoring: "Yes", when open questions on this subject are posed.

Item 50: Asks about religious feelings.
Scoring: "Yes", when this topic is raised by means of an open question.

Item 51: Explores character traits or self-image.
The patient is asked to describe his character. Furthermore, the interviewer asks about personality traits which cause vulnerability to mental health problems such as dependent, paranoid, compulsive, histrionic, schizoid, anti-social, borderline, passive-aggressive, avoident traits.
Scoring: "Yes", when
- the patient is asked for a character self-description and/or
- the patient is asked two or more questions about the above-mentioned personality traits.

Item 53: Asks about social support.
This item pertains to the social support system of patients.
Scoring: "Yes", when the physician asks whether, in the patient's social orbit, there are persons who support him emotionally or materially.

Item 54: Asks about cultural differences in social relationships.
Confrontations with habits, values and norms of different (sub)cultures may entail adaptation problems.
Scoring: "Yes", when such a question is asked.

Item 57: Explores sexual functioning.
This item asks for the quality of sexual functioning. In case of dysfunctioning (e.g. inhibited desire, excitement and orgasm; dyspareunia; vaginismus; premature ejaculation), the nature of the disturbance should be explored.
Scoring: "Yes", when the physician asks about satisfaction with sexual functioning and - in the case of dysfunctioning - its nature.

Item 58: Explores sleep habits.
Scoring: "Yes", when the quality of sleep is asked about and - in the case of insomnia - the nature of it.

Item 59: Explores eating habits.
In this item, the physician explores the eating habits of the patient (appetite, dietary habits) and possible disturbances, such as fear of becoming obese, extreme weight loss, periods of buliminia, disgust etc.
Scoring: "Yes", when
- the physician asks about appetite and eating habits
- in the case of disturbances, their nature is explored.

Item 60: Explores substance use.
This item pertains to the use and abuse of substances: there may be some overlap with item 25.
The DSM-III criteria for abuse are: use of at least 1 month's duration during which period there is impairment in social and occupational functioning due to the substance abuse. Consequences of abuse may be dependency in physical and psychological sense.
Scoring: "Yes", when the patient is asked about use and possible abuse and, in the latter case, about phenomena of dependency.

Item 61: Explores (dis)satisfaction with housing conditions.
In cases of dissatisfaction, objective data should be collected about the housing situation and what changes the patient has in mind should be assessed.
Scoring: "Yes", when the patient is asked about his satisfaction with his housing situation and, in case of dissatisfaction, for objective data concerning the housing situation and his desire for change.

Item 62: Explores (dis)satisfaction with financial situation.
Scoring (analogous to the previous item): "Yes", when the patient is asked about his satisfaction with his financial situation and, in case of dissatisfaction, for objective data concerning his financial situation and his desires for change.

Item 63: Explores history of education and professional life.
In contrast with item 55, in this item the history of professional functioning and education is explored.
Scoring: "Yes", when the physician asks about:
- "historical" facts
- satisfaction or problems experienced

Item 64: Explores early development up to adolescence.
Scoring: "Yes", when at least two of the following three areas are explored:
- traumatic experiences
- socio-emotional development
- (psycho)motor development.

V. PRESENTING SOLUTIONS

Item 67: Gives information about the prognosis of the problem; without and with treatment.
Scoring: "Yes", when information about the prognosis of the disorder/problem is conveyed, concerning treated as well as untreated conditions.

Item 69: Explores how much responsibility the patient is prepared to take for his treatment.
Scoring: "Yes", when the physician explores how much responsibility the patient is prepared to take for the method of treatment (rules, obligations, efforts, time investment etc.) and for the determination of the objectives of treatment.

Item 73: Asks for the patient's opinion about the proposed help.
Scoring: "Yes", when the physician explores the patient's attitudes towards the proposed help.

Item 74: Explores how "important others" might influence the proposed help.
There may be a mutual influence on each other of "proposed help" and "the important others" of the patient. For example: the "proposed help" implicates involvement of "important others" in the treatment; the "proposed help" may be unacceptable for "important others"; considerable support may be asked for in the compliance of the "proposed help"; etc.
Scoring: "Yes", when mutual influence on each other of the "proposed help" and "important others" is explored.

Item 76: Asks the patient to make a choice out of several proposals of help.
After proposing alternatives of help, after discussion of their feasibility and after exploration of the patient's opinion, the physician asks the patient to make a choice between the alternatives.
Scoring: "Yes", when the physician asks the patient to make a selection from alternatives of help (of which one is, of course, no help!).

APPENDIX C AND D

THE MAAS-SELF EVALUATION IN GENERAL PRACTICE AND IN PRIMARY MENTAL HEALTH CARE (MAAS-SELF-GP, MAAS-SELF-PMHC).

This instrument has exactly the same format as the MAAS-GP and the MAAS-PMHC. The items are re-edited in an "I"-format. For instance, the item "Asks the patient what attempts he has made to solve the problem", has been re-edited to "I asked the patient what attempts he has made to solve the problem".

The transformation of the scale "psychiatric examination" of the MAAS-SELF-PMHC is an exception. Its items are transformed into the following for purposes of abreviation:

- I explored symptoms of affective disorders.
- I explored symptoms of anxiety.
- I explored disturbances in consciousness and orientation.
- I explored disturbances in memory.
- I explored disturbances of perception.
- I explored disturbances in thought and thinking.

These instruments are not printed here.

APPENDIX E

GLOBAL SELF-RATING SCALES FOR THE MEASUREMENT OF PHYSICIANS' INTERVIEWING SKILLS IN GENERAL PRACTICE AND IN PRIMARY MENTAL HEALTH CARE

These rating scales are scored on a five-point Likert scale (strongly agree, partially agree, indifferent, partially disagree, strongly disagree).

THE GLOBAL SELF-RATING SCALE FOR GENERAL PRACTICE

Item 1: I adequately explored the reason for encounter: in other words, the complaint and its significance to the patient.

Item 2: I collected all the data necessary for medical problem-solving.

Item 3: I adequately performed the phase of "presenting solutions": in other words, I conveyed information about causes and consequences of the problem and discussed their further management with the patient.

Item 4: I structured the interview: in other words, I discussed the agenda of the interview with the patient and announced the transitions of its different phases.

Item 5: I adequately used interpersonal and communicative skills: in other words, I facilitated communication, asked closed-ended questions in a proper way, summarized and concretized well, confronted adequately, reflected emotions in a proper way and made appropriate meta-communicative remarks.

Item 6: I tried to empathize with the patient.

Item 7: The exchange of information between the patient and me was effective.

Item 8: This interview proceeded satisfactorily.

THE GLOBAL SELF-RATING SCALE FOR PRIMARY MENTAL HEALTH CARE

This self-rating scale is similar to the General Practice version, except for Item 2. This item corresponds with two items in the Primary Mental Health Care version.

a. I collected all the data necessary to explain the mental health problem of the patient.
b. I collected all the data necessary to change the mental health problem of the patient.

APPENDIX F

GLOBAL EXPERT-RATING SCALES FOR THE MEASUREMENT OF PHYSICIANS' INTERVIEWING SKILLS IN GENERAL PRACTICE AND IN PRIMARY MENTAL HEALTH CARE

These rating scales are scored on a five-point Likert scale (strongly agree, partially agree, indifferent, partially disagree, strongly disagree).

THE GLOBAL EXPERT-RATING SCALES FOR GENERAL PRACTICE

Item 1: The physician adequately explored the reason for encounter: in other words, he clarified the complaint and its significance to the patient.

Item 2: The physician collected all the data necessary for medical problem-solving.

Item 3: The physician adequately performed the phase of "presenting solutions": in other words, he conveyed information about causes and consequences of the problem and discussed their further management with the patient.

Item 4: The physician structured the interview: in other words, he discussed the agenda of the interview with the patient and announced transitions of its different phases.

Item 5: The physician adequately used his interpersonal and communicative skills in other words, he facilitated the communication, asked closed-ended questions in a proper way, summarized, concretized well, reflected emotions in a proper way and made appropriate meta-communicative remarks.

Item 6: The physician tried to empathise with the patient.

Item 7: The exchange of information between physician and patient was effective.

Item 8: The physician conducted a good interview.

THE GLOBAL EXPERT-RATING SCALE FOR PRIMARY MENTAL HEALTH CARE

The rating scale is similar to the General Practice version, except for Item 2. This item corresponds with two items in the Primary Mental Health Care version:

a. The physician collected all the data necessary to explain the mental health problem of the patient.
b. The physician collected all the data necessary to change the mental health problem of the patient.

APPENDIX G

GENERALIZABILITY ANALYSIS OF RASCH HOMOGENEOUS MAAS-PMHC SCALES

I. EXPLORATION OF THE REASON FOR ENCOUNTER (MAAS-PMHC)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	0.567	3.66***	2.0
2. observer (o)	5	1.958	12.61***	2.6
3. items (i)	12	6.514		18.1
4. p.o.	95	0.155		4.5
5. p.i.	228	0.499	4.52***	24.7
6. o.i.	60	0.430	3.89***	6.1
7. o.i.p. + error	1140	0.110		42.0

II. HISTORY-TAKING (MAAS-PMHC)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	0.353	2.98***	1.7
2. observer (o)	5	0.924	7.79***	1.7
3. items (i)	12	4.767		19.9
4. p.o.	95	0.119		5.0
5. p.i.	228	0.330	3.87***	22.4
6. o.i.	60	0.176	2.07***	2.5
7. o.i.p. + error	1140	0.085		46.9

III. PSYCHIATRIC EXAMINATION (MAAS-PMHC)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	0.186	5.32***	3.6
2. observer (o)	5	0.101	2.90*	0.5
3. items (i)	8	1.185		10.4
4. p.o.	95	0.035		5.0
5. p.i.	152	0.193	5.62***	34.2
6. o.i.	40	0.062	1.80**	1.8
7. o.i.p. + error	760	0.034		44.5

IV. SOCIO-EMOTIONAL EXPLORATION (MAAS-PMHC)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	0.449	6.59***	2.2
2. observer (o)	5	1.331	19.57***	2.2
3. items (i)	17	5.643		27.5
4. p.o.	95	0.068		2.4
5. p.i.	323	0.325	5.84***	28.3
6. o.i.	40	0.134	2.40***	2.5
7. o.i.p. + error	1615	0.056		35.0

V. PRESENTING SOLUTIONS (MAAS-PMHC)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	0.868	4.19***	3.7
2. observer (o)	5	5.200	25.12***	8.3
3. items (i)	12	6.465		21.1
4. p.o.	95	0.207		6.9
5. p.i.	228	0.203	1.99***	7.3
6. o.i.	60	0.506	4.97***	8.7
7. o.i.p. + error	1140	0.102		44.0

VI. STRUCTURING THE INTERVIEW (MAAS-PMHC)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	0.945	3.65***	8.7
2. observer (o)	5	2.034	7.85***	6.7
3. items (i)	4	2.148		5.2
4. p.o.	95	0.259		19.7
5. p.i.	76	0.240	1.95***	7.4
6. o.i.	20	0.390	3.16***	5.0
7. o.i.p. + error	380	0.123		47.5

VII. INTERPERSONAL SKILLS (MAAS-PMHC)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	0.566	4.20***	2.6
2. observer (o)	5	1.953	14.51***	3.3
3. items (i)	9	15.042		43.4
4. p.o.	95	0.135		4.9
5. p.i.	171	0.146	1.70***	3.7
6. o.i.	45	0.680	7.89***	10.8
7. o.i.p. + error	855	0.086		31.4

VIII. COMMUNICATIVE SKILLS (MAAS-PMHC)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	0.385	2.24***	1.7
2. observer (o)	5	2.053	12.83***	4.9
3. items (i)	6	6.302		15.2
4. p.o.	95	0.160		8.3
5. p.i.	114	0.231	1.86***	6.4
6. o.i.	30	0.154	9.28**	18.6
7. o.i.p. + error	570	0.124		44.9

Note: - * = $p \leq .05$; ** = $p \leq .01$; *** = $p \leq .001$
- items are fixed

GENERALIZABILITY ANALYSIS OF RASCH HOMOGENEOUS MAAS-GP SCALES

I. EXPLORATION OF THE REASON FOR ENCOUNTER (MAAS-GP)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	1.00	11.35***	8.9
2. observer (o)	5	0.80	9.13***	2.1
3. item (i)	6	6.90		21.8
4. p.o	95	0.09		5.2
5. p.i	114	0.44	5.15***	24.5
6. o.i	30	0.18	2.15***	2.0
7. o.i.p + error	570	0.09		35.4

II. HISTORY-TAKING (MAAS-GP)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	2.17	46.21***	20.1
2. observer (o)	5	0.12	2.47*	0.2
3. item (i)	10	1.76		6.9
4. p.o	95	0.05		2.7
5. p.i	190	0.36	6.20***	31.5
6. o.i	50	0.13	2.22***	2.2
7. o.i.p	950	0.06		36.4

III. PRESENTING SOLUTIONS (MAAS-GP)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	1.44	7.48***	7.0
2. observer (o)	5	2.29	11.84***	3.5
3. item (i)	10	6.87		19.0
4. p.o	95	0.19		6.5
5. p.i	190	0.38	3.40***	16.4
6. o.i	50	0.47	4.21***	6.6
7. o.i.p + error	950	0.11		41.0

IV. STRUCTURING THE INTERVIEW (MAAS-GP)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	0.89	5.10***	7.1
2. observer (o)	5	1.13	6.50***	2.9
3. item (i)	5	5.50		14.4
4. p.o	95	0.17		10.4
5. p.i	95	0.40	3.33***	17.0
6. o.i	25	0.38	3.09***	4.6
7. o.i.p + error	475	0.12		43.7

V. INTERPERSONAL SKILLS-MAAS-GP

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	0.79	3.36***	4.1
2. observer (o)	5	0.61	2.62*	0.9
3. item (i)	7	9.45		26.4
4. p.o	95	0.24		10.5
5. p.i	133	0.22	1.67***	5.2
6. o.i	35	0.51	3.93***	6.8
7. p.o.i + error	665	0.13		46.2

VI. COMMUNICATIVE SKILLS-(MAAS-GP)

Source of variation	DF	MS	F	% OF TOTAL VARIANCE
1. physician (p)	19	0.49	1.93*	2.2
2. observer (o)	5	2.55	10.08***	6.6
3. item (i)	5	6.55		16.3
4. p.o	95	0.25		14.6
5. p.i	95	0.18	1.36*	2.7
6. o.i	25	0.86	6.69***	12.7
7. p.o.i + error	475	0.13		44.7

Note: - * = $p \leq .05$; ** = $p \leq .01$; *** = $p \leq .001$
- items are fixed

APPENDIX H

PROBLEM-SOLVING IN PRIMARY MENTAL HEALTH CARE

Instruction:

* please fill out the items as fully as possible
* first study the example of this instrument on the next page (not provided here)

1. Diagnostic classification pertaining to the main complaint. Write down a maximum of three possible diagnoses in order of decreasing probability.
   -
   -
   -

2. Write down the three most important recent factors or conditions which play a role in the origin of this mental health problem (e.g. psychosocial stressors, life-events, somatic factors, etc.).
   -
   -
   -

3. What long-standing factors precipitated and/or maintained the pertinent mental health problem?
   Mention two of them.
   -
   -

4. What factors in the patient diminish his/her psychologic capacities (e.g. developmental disturbances, genetic factors, unfavorable personality traits, oligophrenia, etc.)?
   -
   -

5. What patient management plan (further examination, treatment plan or both) do you propose, based on the collected data?
   -
   -
   -

6, Adduce arguments in support of each of the elements of this management plan.
   -
   -
   -

CURRICULA VITARUM

Herro Foeke Kraan was born on May 6th 1944 in Wymbritseradeel, The Netherlands. After graduating from high school, at the Thorbecke Lyceum (Gymnasium-B) in Arnhem, he studied in Groningen at the Medical Faculty and the Faculty of Arts (Slavic languages) from 1963 to 1971.
After military service he studied Psychiatry from 1972 to 1978 at the Ursula Kliniek, Wassenaar (Neurology Department; head: Dr. J. Tans sr.) and in the Wilhemina Gasthuis, Amsterdam (Psychiatry Department; head: Prof. Dr. P.C. Kuiper). From 1976 to 1978 he worked as a social psychiatrist in the Department of Mental Health at the Municipal Health Service and in the Institute of Multidisciplinary Psychotherapy, both in Amsterdam.
Since 1978 he has been a staff member of the Department of Social Psychiatry, University of Limburg, Maastricht, The Netherlands and works as a social psychiatrist and psychotherapist at the RIAGG, Maastricht.

Alfons, Arjen, Marie Crijnen was born on February 5th 1956, in Eindhoven, The Netherlands. He graduated in 1974 from the St. Maartenscollege in Maastricht (Atheneum-B). He attended Medical School from 1974 to 1980 and finished his General Practice Residency Program at the University of Limburg, Maastricht in 1981.
He worked from 1981 to 1987, at the Department of Social Psychiatry, University of Limburg, Maastricht (head: Prof. Dr. M.A.J. Romme and Prof. Dr. M.W. deVries) on the research project "Measurement of clinical competency in the psychomedical domain". From 1982 to 1983 he was a general practitioner at the Consultation Bureau for Alcohol and Drugs in Sittard and Maastricht and worked from 1983 to 1986, as a general practitioner in Heerlen.
Currently, he is a clinical assistant at the Department of Clinical Psychiatry, Academical Medical Centre, University of Amsterdam (head: Prof. Dr. F.E.R.E.R. de Jonghe).