Maastricht University

Teaching and assessing diagnostic interviewing skills : an application to the mental health field

Citation for published version (APA):

Bögels, S. M. (1994). *Teaching and assessing diagnostic interviewing skills : an application to the mental health field*. [Doctoral Thesis, Maastricht University]. Datawyse / Universitaire Pers Maastricht. https://doi.org/10.26481/dis.19941007sb

Document status and date:
Published: 01/01/1994

DOI:
10.26481/dis.19941007sb

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: 17 Apr. 2024

TEACHING AND ASSESSING DIAGNOSTIC INTERVIEWING SKILLS

AN APPLICATION TO THE MENTAL HEALTH FIELD

Lay-out:
Marja Hortulanus

Druk:
Datawyse Maastricht / Krips Repro Meppel

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Bögels, Susan Maria

Teaching and assessing diagnostic interviewing skills : an application to the mental health field / Susan Maria Bögels - Maastricht : Universitaire Pers Maastricht. - Ill.
Thesis Maastricht - With ref. - With summary in Dutch.
ISBN 90-5278-154-0
Subject headings: skills-training ; mental health care / diagnostic interviewing / authentic assessment

TEACHING AND ASSESSING DIAGNOSTIC INTERVIEWING SKILLS

AN APPLICATION TO THE MENTAL HEALTH FIELD

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Rijksuniversiteit Limburg te Maastricht,
op gezag van de Rector Magnificus, Prof. Dr. H. Philipsen,
volgens het besluit van het College van Dekanen,
in het openbaar te verdedigen op
vrijdag 7 oktober 1994 om 14.00 uur

door

Susan Maria Bögels

★
UPM
UNIVERSITAIRE PERS MAASTRICHT

Promotor:

Prof.dr. H.G. Schmidt

Co-promotor:

Dr. C.P.M. van der Vleuten

Beoordelingscommissie:

Prof.dr. H.M. van Praag (voorzitter)
Prof.dr. M.A. van den Hout
Dr. H.F. Kraan
Prof.dr. H.T. van der Molen (Open Universiteit, Heerlen)
Prof.dr. W.H.F.W. Wijnen

CONTENTS

Chapter 1.	Introduction	1
Chapter 2.	A structured training approach to teaching diagnostic interviewing	7
Chapter 3.	Assessment and validation of diagnostic interviewing skills	19
Chapter 4.	Measuring diagnostic interviewing performance: Reproducibility of scores	31
Chapter 5.	Authentic assessment of interviewing and counselling skills: Effect of testing time per station on reproducibility and validity	39
Chapter 6.	Effects of a training in interviewing and counselling skills	51
Chapter 7.	Training diagnostic interviewing to students of Mental Health Sciences: Immediate and long-term effects	65
Chapter 8.	Diagnostic expertise in the mental health professions: Differences between novices and experts	81
Chapter 9.	Discussion	95
Summary		105
Samenvatting		109
References		113
Appendix I	Handleiding beoordeling van het anamnestisch interview in de geestelijke gezondheidszorg	123
Appendix II	Probleemoplossingsvragenlijst	140
Appendix III	Patientsatisfactie-vragenlijst	141
Dankwoord		143
Curriculum vitae		145

CONTENTS

Chapter 1
Introduction

In mental health care, the diagnostic interview[1] constitutes the main tool for assessment and, as such, for initiating appropriate treatment of mental disorders. Basic interviewing, counselling, and diagnostic interviewing, are, therefore, important skills for mental health professionals; skills students should master during their formative years. However, although basic interviewing and counselling skills tend to be emphasized in many training programmes, the teaching of diagnostic interviewing to future workers in mental health care, such as clinical psychologists, mental health scientists, counsellors, and psychiatrists, is often ignored during education (McGuire, 1982; McCready & Waring, 1986). Descriptions of training programmes to teach diagnostic interviewing skills to future mental health professionals are scarce. Little is known about effective ways to teach these skills to novices. The assumption may be that students automatically acquire these skills by means of practice in clinical settings. Research in the medical field showed that this is not the case, on the contrary, students' skills on certain measures actually decline from the first year to the fourth (Carrol & Monroe, 1980).

In 1985 a new academic programme was established at the University of Limburg, Faculty of Health Sciences. This program provides undergraduate training in what is called Mental Health Sciences. From the start, the importance of developing a substantial professional skills curriculum was recognized. In a fruitful collaboration between staff of at that time the Skills Laboratory[2] and the Department of Experimental and Abnormal Psychology, such a curriculum was developed. It aims to teach students a wide range of interpersonal professional

[1] Often, the term "psychiatric interview" is used to refer to the initial interview of patients with mental health problems.

[2] Part of the staff, responsible for the development of professional skills training for the Faculty of Health Sciences, joined the Department of Educational Development and Research in 1989.

skills, of which diagnostic interviewing forms the core. The training takes place in the context of a problem-based four-year curriculum. Theoretical instruction is organized in units of 6 weeks, in which integrated, multidisciplinary teaching takes place around a specific theme. Skills-training is closely linked to the theme of the unit, in order to integrate theory and practice. Innovative training methods, like the use of simulated patients and staff-independent training approaches are applied. In addition, behavioural tests have been developed to assess students' emerging competencies. The development and execution of this curriculum was associated with research to evaluate the effects of training and the appropriateness of skills assessments, such as the studies reported in this thesis.

Two main issues were studied: 1) the development and evaluation of a training programme to teach diagnostic interviewing to students in the mental health sciences, and 2) the assessment of diagnostic interviewing performance of (future) mental health workers. Before we proceed to describe the studies in more detail, we briefly review the main areas of research that provided a background for the present studies.

The first area of research stems from behaviouristic conceptions of skills and skills-training. In this approach, a complex task is divided into discrete, small behaviour entities, based on a precise analysis of the complex task. By means of learning principles stemming from behaviouristic theories, including step-by-step learning, modelling, reinforcement, and feedback, these entities are trained.

The microtraining method, developed by Ivey (1971), is an example of the behaviouristic approach. Microtraining is a model of instruction that subdivides complex interpersonal human behaviour into discrete behavioural units and then teaches these units through didactic instruction (including modelling), behavioural practice, self-observation, and immediate feedback about the correctness of performance. Microtraining has had, without doubt, a large influence on the development and recognition of interpersonal skills training as a distinguishable and essential part of education in the health professions. Microtraining not only stimulated the development of effective training programmes, but also encouraged empirical evaluation of –observable– training effects with regard to the acquisition of interpersonal skills (Baker & Daniels, 1989).

A second area of investigation, relevant for skills conception and skills training, is the cognitive approach. In contrast to the exclusive attention to observable behaviour in microtraining research, in the cognitive approach cognitive processes during task performance are investigated. The cognitive strategies which a person uses in manipulating information associated with a task are analyzed. Various cognitive measures, like thinking aloud, recall, and problem solving, are used to study differences in skilfullness between experts and novices. This line of research has resulted in a better understanding of what characterises expertise

or skillfulness. Expertise, for example, appears to be highly case-specific, that is: a clinician's actual experience with a particular problem has the greatest influence on clinical reasoning and problem-solving (Elstein, Shulman, & Sprafka, 1978). Experts, furthermore, possess more knowledge about the application of what they know than novices (Glaser, 1984). Experts appear to diagnose by means of rapid pattern recognition. They diagnose a new patient on the basis of similarities to previously encountered patients, and reason about a patient's signs and symptoms, applying knowledge in encapsulated form (Schmidt & Boshuizen, 1992).

The cognitive approach to skills acquisition has important consequences for training. The aim of training, from a cognitive perspective, is not so much to model and reinforce skilled behaviour, but to ensure that relevant and correct conceptual models are available to and selected by the trainee, while performing a task. Skills of on-the-spot information processing are further emphasized during training. Furthermore, the integration of the acquisition of knowledge and the acquisition of skills becomes extremely important from a cognitive perspective.

Methods of assessment of professional skills have been studied extensively in medical education. Clinical skills tests in medical schools usually consist of a number of "stations" at which students are required to perform a variety of clinical tasks, including history-taking, physical examination and counselling, using simulated patients. This way of testing is now termed "authentic assessment", referring to the "authenticity" of the test[3]. There is a wealth of literature on authentic assessment of physicians' skills (Van der Vleuten & Swanson, 1990). Swanson (1992) has summarized the most important problems in the assessment of these skills. First, the validity of authentic assessment has been questioned, as beginners often outperform experts on such tests. Secondly, for assessment procedures involving direct observation of performance, it is difficult to achieve consistency from one rater to the next. Thirdly, performance on one case is a poor predictor of performance on other cases, generally referred to as "the case-specificity of performance". Finally, practical problems of costs, logistics and security of authentic assessment can be insurmountable.

In spite of the psychometric and practical problems, authentic assessment is widely used in medical education, as it possesses many desirable characteristics. First, authentic assessment has high face validity. The test reflects the same tasks and activities that have been the subject of training. It is also representative

[3] In the past decade, "performance-based testing" was mostly used to refer to this method of skills assessment. In this thesis, the term authentic assessment is used to refer to tests to measure professional skills by means of encounters with simulated patients. These encounters are observed by raters, and skills are judged with the help of a rating form or checklist.

of those tasks and activities of the future profession, for which the students are prepared. Secondly, authentic assessment possesses high "consequential validity" (Rettie, 1992), as it has a desirable impact on teaching and learning. It is well known that tests and examinations strongly influence students' studying behaviour (Newble & Jaeger, 1983) and teachers' teaching behaviour (Frederiksen, 1984). Authentic assessment stimulates teachers to train their students until mastery is achieved, and motivate trainees to exercise their skills. Thirdly, authentic assessment can provide students with precise feedback regarding their level of skills enhancement. Skills-based tests are especially suitable for the assessment of artistic, procedural and interpersonal skills, since traditional evaluation methods, such as multiple-choice tests, cannot directly assess the performance on these types of skills. As a result of both the desirable characteristics of authentic assessment and the emerging recognition of the importance of skills-training in health education, there is a growing demand in health professions education for skills-based tests. Authentic assessment of diagnostic interviewing in mental health education has, however, received little attention so far.

In this thesis, diagnostic interviewing ability with regard to mental health problems (or psychiatric interviewing ability) was measured under authentic assessment conditions. Two types of skills were distinguished in the *overt* behaviour of a mental health worker interviewing a patient: basic interviewing or process skills, reflecting the interpersonal and communication skills, and content interviewing skills, referring to the information-gathering ability of the interviewer. Cognitive skills of the interviewers, like the mental representation of a patient case and diagnostic accuracy, were also measured in the different studies. The assessment usually consisted of a number of encounters with simulated patients, portraying different mental health problems. The interviewers under study were instructed to clarify the problem and gather the necessary information, in order to make a diagnosis. After each interview, subjects were asked to write down their problem analysis and diagnostic formulation. All interviews were recorded on videotape and judged by a number of raters who were trained in using a rating scale to establish the quality of process and content interviewing skills.

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) was used to analyze the data. By means of generalizability analysis the magnitude of multiple sources of measurement variance can be estimated. Variance resulting from differences between interviewers, raters, patient cases, and their interactions, was taken into account in the studies. The "case-specificity" issue (the interaction between interviewer and patient case) was addressed by means of this type of analysis. The overall reliability (or reproducibility) of the obtained results was estimated, based on the variance components found.

The following questions were investigated in this thesis:

1. *How should undergraduate students be trained in diagnostic interviewing?*

A review of existing training methods to teach diagnostic interviewing, empirical evidence on effective interview styles, and an analysis of the tasks of the diagnostic interviewer, is provided in Chapter 2, followed by a description of the diagnostic interviewing curriculum, developed at the University of Limburg.

2. *Can diagnostic interviewing performance reliably and validly be assessed under authentic assessment conditions?*

In Chapter 3, the inter-rater reliability, case-specificity, and the resulting reproducibility of diagnostic interviewing performance scores is investigated. Furthermore, the construct validity of diagnostic interviewing performance is examined.

3. *How can authentic assessment of diagnostic interviewing performance be made more efficient (feasible)?*

In Chapter 4, the reproducibility of diagnostic interviewing performance scores is investigated using another sample of interviewers, now with a restriction on the diagnostic categories of the simulated patient cases, as a means to increase reproducibility of scores. In Chapter 5, effects of shortening interviewing time per case on reproducibility of basic interviewing skills scores are examined, suggesting another alternative to shorten the testing time of authentic assessment without losing reproducibility.

4. *What do students learn from training in diagnostic interviewing?*

Effects of a training course in basic interviewing and counselling skills on various outcome measures are studied in Chapter 6. In Chapter 7, effects of two training courses in diagnostic interviewing on students' interview performance, recall of relevant patient information, and diagnostic accuracy, are investigated. Special attention is paid to possible case-specificity in learning.

5. *What are the differences between novices and experts with regard to diagnostic interviewing performance?*

In Chapter 8, differences in interviewing behaviour and diagnostic accuracy of experts and novices are investigated. The purpose of this study was twofold. The study first attempts to obtain insight in what constitutes psychiatric expertise. Secondly, it provides an independent test of the discriminant validity of the proposed authentic assessment procedure to measure diagnostic interview performance.

In conclusion, as the chapters of this thesis were submitted as separate articles, some overlap between the different chapters was inevitable.

Chapter 2

A Structured Training Approach to Teaching Diagnostic Interviewing[1]

Summary

In this article, recent developments in the teaching of diagnostic interviewing skills in the mental health professions are discussed. First, the tasks and skills of the diagnostic interviewer are analyzed. Secondly, empirical evidence on effective interviewing styles is outlined. Thirdly, training methods for teaching diagnostic interviewing are reviewed. A training programme, developed to teach diagnostic interviewing to undergraduate students, is then described. The programme is highly structured and simulated patients are used to introduce complex clinical problems and to evaluate students' emerging competencies. Student and trainer satisfaction with the programme has been high over the past 6 years.

[1] Bögels, S. M. (1994). A structured training approach to teaching diagnostic interviewing. *Teaching of Psychology* (in press). A modified version of this chapter has been published as: Bögels, S. M. (1992). Anamnestische interviewvaardigheden bij geestelijke gezondheidsproblemen. In C. P. M. Van der Vleuten, A. J. J. A. Scherpbier, M. C. Pollemans (Red.) Gezond Onderwijs. Houten: Bohn Stafleu Van Loghum.

Introduction

The diagnostic interview is the main assessment tool in mental health care. The goals of the diagnostic interview are to arrive at a diagnosis and a rational treatment plan. The diagnostic interview traditionally has been a task of the psychiatrist. However, because of the increasing involvement of professionals from other disciplines (psychologists, social workers, doctors, and psychiatric nurses) in the provision of mental health care, diagnostic interviewing has become an important skill for all mental health professionals.

Most educators agree that diagnostic interviewing is an important skill that mental health professionals should master during their formative years. Although basic interviewing and counselling skills tend to be included in training (Baker & Daniels, 1989), the teaching of *diagnostic* interviewing skills is often ignored, in clinical psychology education, counsellor training, and psychiatric residency training (McCready and Waring, 1986; McGuire, 1982). A review of the literature on teaching diagnostic interviewing reveals that most articles focus on medical students and family medical residents (mostly termed history-taking). Studies on teaching diagnostic interviewing for mental health problems (often referred to as psychiatric interviewing) to psychiatric residents, psychology students, or counsellor trainees are scarce. Little is known about how to teach these skills effectively to novices in this field.

The purpose of this article is to review recent developments in the teaching of diagnostic interviewing for the mental health professions. First, the tasks and skills of the interviewer are discussed. Secondly, empirical evidence on effective diagnostic interviewing styles is reviewed. Thirdly, training methods related to the teaching of diagnostic interviewing are outlined. Finally, a training programme is presented as an example of an integrated approach to the teaching of diagnostic interviewing skills to undergraduates.

Tasks and Skills of the Diagnostic Interviewer

The major goals of the diagnostic interview, as outlined earlier, are to arrive at a diagnosis and treatment plan. Usually, a diagnosis is formulated in terms of psychopathology (the descriptive diagnosis), generally by means of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III, American Psychiatric Association, 1987). In addition, the interviewer seeks to reformulate patients' problems by reconstructing patients' life history and analyzing their present circumstances and the organization of their personality (Siassi, 1984). For this etiologic formulation, or explanatory diagnosis, no universally used language is available. The interviewing task thus consists of gathering relevant information from the patient for descriptive and explanatory diagnosis, as well as other

information necessary for planning treatment. Establishing a working alliance with the patient is another goal of the diagnostic interview (Reiser, 1984). Whether or not the interviewer becomes engaged in the treatment of the patient from the initial diagnostic interview, agreement between interviewer and patient, based on mutual rapport and comprehension, is important for compliance.

To teach a complex professional skill such as diagnostic interviewing, it is necessary to divide the skill into behavioural units (Authier & Gustafson, 1982). Mumford, Schlesinger, Cuerdon, and Scully (1987) distinguished two types of skills in the overt behaviour of a psychiatrist interviewing a patient: skills focusing on the relational and communicative aspects of the encounter (process skills) and skills focusing on information gathering (content skills). Process skills reflect the ability of interviewers to show interest in the problems of the patient, to elicit information in an open and natural way, to communicate their understanding of the problem to the patient, and to provide the patient with as much information as possible about the purpose of the interview, the therapy, and so forth. Process skills can be subdivided into two main categories: listening skills and regulating skills (Lang & Van der Molen, 1990). Listening skills are required to encourage patients to tell their story. By means of regulating skills (or meta-communicative skills) interviewers regulate the course of the interview and give insight in their working method. We separate the following listening skills: attending, asking questions, concreteness, paraphrasing of content, reflection of feelings, summarizing. The following regulating skills are distinguished: giving information, asking feedback, thinking aloud, structuring.

Content skills reflect the way interviewers handle their agenda (e.g., the topics important to cover during the interview). The literature on psychiatric interviewing suggests that there is consensus on the topics to be covered during the initial diagnostic interview. The interviewer's agenda should contain the following: identification of the patient, clarification of the present complaint, history of main problem, personal history, family history, medical history, history of prior treatments, mental status examination, assessment of the premorbid personality, present circumstances, and request for help (Gelder, Gath, & Mayou, 1989; Ginsberg, 1985; Mackinnon & Michels, 1971; Nicholi, 1978; Nurcombe & Gallagher, 1986; Reiser, 1984; Rosenthal & Akiskal, 1985; Shea, 1988; Siassi, 1984; Thompson, 1979).

So far we have concentrated our discussion on the overt behaviour of interviewers. We now turn to the mental activities during the interview. Interviewers have to process substantial information from verbal and nonverbal sources simultaneously, select diagnostic hypotheses based on this information, and develop questions to test and progressively refine their hypotheses (Nurcombe & Gallagher, 1986). To select diagnostic hypotheses, they may compare the case to other cases stored in memory. Furthermore, they may use their knowledge of diagnostic classifications to select and test hypotheses. These cognitive activities are referred to as cognitive skills or problem-solving skills.

Effective Diagnostic Interviewing Styles

Research in the medical domain showed that good interpersonal and communication skills affect the adequacy of the medical interview. These skills have a positive effect on the working relationship (Pendleton, Schofield, Tate, & Havelock, 1984), increase patient satisfaction (Kraan & Crijnen, 1987), promote compliance with treatment (DiMatteo & DiNicola, 1982), and improve diagnostic accuracy (Evans, Stanley, Mestrovic, & Rose, 1991). Goldberg and Huxley (1980) found that interviewing skills by general practitioners associated with a high detection of mental health problems include a) good eye contact b) clarification of patients' presenting complaints, c) picking up verbal and non-verbal cues, d) using open-ended questions, and e) being able to direct the course of the interview.

Surprisingly little empirical evidence is available on what constitutes effective diagnostic interviewing in mental health care. Cox and his colleagues (Cox, Holbrook, & Rutter, 1981; Cox, Hopkinson, & Rutter, 1981; Cox, Rutter, & Holbrook, 1981) conducted a series of studies to evaluate the efficiency of different diagnostic interviewing styles. Their subjects were psychiatrists interviewing parents of children referred to a clinic. In both a naturalistic study (Cox, Hopkinson, & Rutter, 1981) and an experimental study (Cox, Rutter, & Holbrook, 1981), they found that directive and nondirective approaches produced equal numbers of reported problems. However, directive interview styles, consisting of specific probes and inquiry for detailed information, produced fuller and more detailed information on each problem and more information on topics where there was no abnormality. Furthermore, systematic questioning had no negative effects on the expression of emotions by the patient (Cox, Holbrook, & Rutter, 1981).

Since the introduction of the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 1968), structured interview schedules, based on the DSM-III classification, have been developed (see Friedman, 1989, for an overview of the frequently used types). Although there is general agreement that structured interviews improve the reliability of psychiatric diagnoses (Siassi, 1984), they have some important disadvantages. First, they are time consuming. Secondly, they do not allow experienced diagnosticians to take shortcuts, whereas experts, as a result of their more elaborated knowledge, generally gather far less data than novices to accurately diagnose cases (Schmidt, Norman, & Boshuizen, 1990). Thirdly, they do not pay attention to the interpersonal and communication skills of the helper, and, as a result, patients may not feel understood by the interviewer. Furthermore, structured interviews do not cover themes such as request for help, life history, and premorbid personality.

The two interviewing styles that are generally distinguished, a nondirective, counselling approach and a directive, structured approach, have not been well

integrated into training courses (Shea & Mezzich, 1988). McCready and Waring (1986), for example, noticed that, as a result of education that focused only on data collection, diagnostic formulation, and management, psychiatric residents appear to be preoccupied with these issues at the expense of empathic listening. Interviewers using a counselling approach, by contrast, may not cover all important issues for accurate diagnostics. Shea and Mezzich (1988) observed that psychiatric residents tend to err in one of two ways; performing the interview in an overly structured way or allowing the patient to ramble unproductively.

To conclude, the few empirical studies directed at effective diagnostic interviewing techniques and styles in mental health care point to the importance of both nondirective, counselling and directive, structured techniques. Clinical impressions of educators are that trainees have difficulty integrating these two techniques, suggesting that training should focus explicitly on using nondirective and directive interviewing techniques in conjunction.

Training Methods

Research on the teaching medical interviewing suggests that highly structured training programmes in which specific skills are identified, demonstrated, practised and evaluated are more effective than less structured programmes in influencing behaviour (Carrol & Monroe, 1979). Moreover, highly structured programmes are more positively evaluated by students (Carroll & Monroe, 1979). Ivey's (1971) microcounselling approach is an example of such a highly structured training programme. Microcounselling is a model of instruction that subdivides complex human behaviours into discrete behavioural units and then teaches those units through didactic instruction, behavioural practice, observation, and immediate feedback. In microcounselling, skills are practised through role-playing. Patient roles are impersonated by fellow trainees. Although substantial evidence indicates that microcounselling is an effective training approach, the method also has been criticized. Robinson and Halliday (1987), for instance, argued that microcounselling concentrates too much on the overt behaviour of the counsellor and that the complex cognitive skill of reaching a deeper understanding of a client's problem needs to be given far more emphasis in training. Robinson and Halliday (1988) studied the interview behaviour of counsellors interviewing two simulated patients. They found that the frequency of reasoning out loud was highly predictive of problem-analysis quality, suggesting that the skills of on-the-spot information processing need to be emphasized more in counsellor training.

A second training approach, applied in psychiatric education, relies on direct supervision of residents interviewing "real" patients (Salvendy, 1987). In the traditional method of indirect supervision, the interviewer informs the supervisor

of the interview process and the patient's problems. The supervisor, therefore, must rely on secondhand information that may be unreliable. In the direct supervisory approach, by contrast, the supervisor observes the initial interview directly (by means of a one-way mirror or by sitting in with the interviewer) and gives feedback at the end of the interview. An alternative direct supervision approach, described by Lovett, Cox, and Abou-Saleh (1990), is to interview a patient in front of a small group of peers and a supervisor. Peer feedback is given afterward, and group discussion is facilitated by the supervisor. Jaynes, Charles, Kass, and Holzman (1979) found that, compared to patients whose therapists had indirect supervision, more than twice as many patients of therapists with direct supervision remained in active treatment or successfully completed treatment.

Shea (1988) developed a method to analyze the flow of the psychiatric interview, the facilics system (*facilis* is Latin for "ease of movement"). Residents learn to recognize the various clinical topics (called regions) explored in interview situations and the various methods of making transitions (called gates) between these regions. Two types of regions –process and content– are distinguished. Process regions focus on such issues as engagement, patient opinion, and patient perspective. Content regions refer to areas of conversation in which the focus is on data gathering. Facilics analysis, thus, provides a means for structuring the interview and making it more manageable for trainees.

A third approach, mainly practised in medical education, is to use simulated patients in teaching diagnostic skills. A simulated patient is a person who has been trained to completely simulate a patient or any aspect of a patient's illness, depending on the educational need (Barrows, 1971). Simulated patients have important advantages over real patients. First, the trainee can practise without the risk of hurting or exhausting the patient. Secondly, a simulation allows for controlled learning. For example, a complex diagnostic problem can be presented to the trainee in more manageable units. Thirdly, simulated patients are in a better position to give objective feedback about the quality of the care they have received, because their perspective on the interviewer is not influenced by real troubles and need for help. Compared to role-playing in which a fellow trainee simulates a patient, standardization and life-likeness of simulated patients is high (Norman, Neufeld, Walsh, Woodward, & McConvey, 1985; Rethans & Van Boven, 1987; Sanson-Fisher & Poole, 1980). As a result, skills learned during interactions with simulated patients may generalize easily to real practice.

To summarize, diagnostic interviewing ability consists of at least three types of skills and their interactions: process skills, content skills, and cognitive skills. Thus, the published literature suggests that effective training in diagnostic interviewing should a) focus on the acquisition and the integrated application of these three skills, b) be highly structured, following microteaching principles, and c) use simulated patients or direct supervision.

Teaching Diagnostic Interviewing: A Structured Approach

Training Programme

At the University of Limburg, Maastricht, The Netherlands, a training programme has been developed to teach Mental Health Sciences students to conduct a diagnostic interview. It is designed around the constraints of limited time in the curriculum and lack of access to a clinical population (clinical clerkships are organized at the end of the curriculum, outside the university in institutes for mental health care). The training courses take place in the context of a problem-based curriculum (Barrows, 1980). Education is organized in units of 6 weeks, in which integrated, multidisciplinary teaching takes place by means of a specific theme. The training programme aims at the acquisition of process, content, and cognitive skills by means of structured training methods, that use simulated patients[2].

In the first year of their study, students receive training in process skills, based mainly on the microcounselling model of Ivey (1971). The training in process skills consists of five 3-hr sessions, with groups of 8 to 10 students, guided by a faculty member. During each session, one or two process skills (e.g., the use of open-ended questions or paraphrasing information) are taught by means of video-modelling, role-playing, and immediate feedback. Role-playing is organized in triads, with one trainee playing a patient, another being the helper, and the third being the observer. Trainees impersonating a patient are provided with written scripts of patients with psychosocial problems. During the training, no emphasis is put on diagnostic issues. Feedback is directed only to the proper and functional use of process skills.

Introductory training in diagnostic interviewing occurs in the second year of the curriculum. The training course consists of six 3-hr sessions. During this training, students learn to conduct a diagnostic interview. A clear model of the diagnostic interview is provided, in which four stages are distinguished: a) problem clarification and history of the present problem, b) psychiatric examination, c) personal history, and d) present functioning. Content skills, process skills, and cognitive skills to be used in each stage are outlined (see Table 2.1).

---

[2] Apart from diagnostic interviewing students are also trained in other professional skills, like psychological testing and therapeutic interventions, by means of the same training principles. These training courses are however beyond the scope of this article.

Table 2.1.
The diagnostic interview in mental health care: content, process, and cognitive skills

Content skills	Process skills	Cognitive skills
1. Problem clarification and history of the main problem	Giving information Asking for feedback	
Introduction		
1a. Problem clarification		
Clarification of present problem	Attending	
Exploration of other problems	Open-ended questions	
Reason for encounter	Reflection of feelings	
Request for help	Paraphrasing of content	
1b. History of present problem	Open- and close-ended questions	Formulating hypotheses about axis-I diagnoses Selecting questions to test the hypotheses
Analysing main problem	Concreteness	
Causal factors	Thinking aloud (giving insight in questions from interviewers' perspective)	
Duration and course		
Consequences for daily life	Summarizing	
Prior treatments	Structuring	
Medical history		
2. Psychiatric examination	Giving information (introducing the second stage)	
Orientation in time, place, person	Specific open- and close-ended questions Thinking aloud Structuring	Formulating (differential) diagnosis axes I and III (if any)
Concentration and memory		
Perception		
Thinking and insight		
Mood and affect		
3. Personal history	Giving information (introducing the third stage) All listening skills Thinking aloud Structuring	Formulating precipitating, predisposing, and maintaining factors Axis II diagnosis (if any)
Childhood		
Adolescence		
Adulthood		
Intrapsychic, social, intellectual and sexual development		
Premorbid personality		
4. Present functioning	Giving information (introducing the fourth stage) All listening skills	Diagnosis axes IV and V
Quantity and quality of social relations		
Occupational functioning		
Work satisfaction		
Financial situation		
Closing the interview	Summarizing Giving information Asking for feedback	

The training format consists of the following components:

1. Orientation on personal base-line level. Prior knowledge about the diagnostic interview is activated by means of group discussion and individual learning goals are formulated.
2. Written information about the skills. Trainees receive an extensive training manual that operationally defines the skills to be learned during the training. Furthermore, relevant literature (e.g., Hersen & Turner, 1985 (Chapter 1 and 2)) about the psychiatric interview is supplied.
3. Modelling. Each training session starts with a 10- to 20-minute videotaped demonstration of one of the stages of the diagnostic interview that is discussed afterward.
4. Practising the skill. The skills are practised during 10- to 20-minute videotaped role-plays. The trainees interview fellow trainees simulating a patient with a psychiatric disorder, based on written patient roles. Role-playing is organized in triads or in plenary sessions.
5. Feedback. Role-playing is followed by immediate feedback by the "patient", the student-observers, and the trainer. Feedback is constantly focused on the single skill being taught. Videotaped feedback is also provided.
6. Diagnosis. Based on the information gathered during the role-play, trainees make diagnostic formulations and a plan for further management.
7. Group discussions. At the end of the session, the group evaluates what has been learned during the session, and individual learning goals with respect to process, content, and cognitive skills are formulated.

The training starts with an introduction to the four stages of the diagnostic interview. Next, students practise each stage separately. During the practice of the content skills, little attention is directed to process skills, following Ivey's principle of one skill at a time (Ivey, 1971). At the end of this training, students are trained to use process and content skills in conjunction. The case material used for role-playing during the training is restricted to affective disorders. The reason for this restriction is twofold. First, the training course is scheduled during an educational unit in which mood is the central theme, and equalizing training and educational content promotes integration of theory and practice. Secondly, research on clinical competence revealed that clinical competence is highly content-specific (Elstein, Shulman, & Sprafka, 1978), which implies that skills are acquired separately for different contents. Thereby, the learning principle of simplicity (one skill at a time) of Ivey (1971) may also be important for content (one category of disorder at a time).

Advanced training in diagnostic interviewing takes place in the third year. The size of the training course is also six 3-hr sessions. The purpose of the training is to learn to conduct diagnostic interviews with psychotic patients. The case material consists of acute psychoses, chronic psychoses, and prepsychotic disorders. Role-playing is organized with the help of simulated patients, in order

to make it as naturalistic as possible. Furthermore, two students in each training class interview a psychotic patient from a mental health institute (a psychiatric hospital or an ambulant mental health centre). The interview occurs at the mental health institute. The videotape of the interview is observed and discussed in the class afterward. Feedback is given with regard to the execution of both process and content skills.

Diagnostic reasoning is taught by showing a large number of short videotaped vignettes of actual and simulated patients with mental disorders. Students formulate hypotheses about the diagnosis in DSM-III-R terms. They motivate their (differential) diagnostic hypotheses on the basis of observed signs and symptoms, and formulate questions they would ask in order to test their hypotheses. Diagnostic reasoning is also encouraged during role-playing. After a "patient" has been interviewed, the trainer (or student-observer) asks the interviewer which diagnostic hypotheses went through his or her mind during the interview. It is then evaluated whether these diagnostic hypotheses have been tested during the interview. Other interventions are suggested to gather information for further verification of the hypotheses. The role-playing is then continued to try out the suggestions.

Use of Simulated Patients

Simulated patient encounters are organized in two ways: during skill training sessions or at the end of a training course (Blok & Bögels, 1990). By means of patient contacts during plenary training sessions problems from the practice are introduced "live" in the training group. During the training in diagnostic interviewing of psychotic patients, a simulated patient is invited to imitate a psychotic disorder in every session. The simulations vary depending on the setting (emergency centre, ambulant mental health institute, private practice, or psychiatric hospital for long-term stay) and kind of disorder (delusional disorder, schizophrenia, or reactive psychosis). Plenary sessions with a simulated patient in the group allow for discussion before and during the encounter on which interview strategy to follow. These encounters are occasions to confront trainees with new and difficult problems, such as dealing with hostile or paranoid patients, in a supporting learning environment.

For each student, an individual encounter with a simulated patient is organized at the end of every training course. The room is equipped with a one-way screen, and the encounter is videotaped. The trainer and a fellow trainee observe the encounter from behind a one-way screen. After the interview, the trainee receives feedback from the patient, fellowstudent, and trainer. The latter two provide feedback using a structured rating form. These encounters are meant to evaluate the level of skills a student has attained after training.

To faithfully simulate mental disorders, such as psychotic or manic disorders, it is helpful to have acquired experience in communication with such patients. Therefore, psychiatric nurses are used as simulated patients.

Evaluation of Students' Emerging Competencies

To measure diagnostic interviewing performance, two instruments are used: the Diagnostic Interviewing Rating Scale in Mental health care (DIRSM; Bögels et al., 1994), and a problem-solving questionnaire (Kraan & Crijnen, 1987).

The DIRSM measures two dimensions of diagnostic interviewing: process and content skills. Process scores reflect ratings on 10 aspects of interpersonal and communication skills (e.g., asking questions, summarizing, thinking aloud). Content scores reflect ratings on 13 aspects of information gathering ability (e.g., problem clarification, psychiatric examination, analysing personal history). Aspects are defined in behavioural terms and criteria of inadequate and adequate behaviour are described. Each aspect is scored on a 5-point scale ranging from unacceptable (1) to advanced (5).

Research (Bögels et al., 1994) has indicated that the process and content interviewing skills can be reliably measured, as far as inter-rater reliability is concerned. The intraclass correlation was .79 for process skills, and .72 for content skills ($N=20$). However, interviewer performance on one case proved to be a poor predictor of performance on other cases. Therefore, at least four simulated patient encounters with different categories of patients are necessary to obtain reliable scores of general diagnostic interviewing ability. Evidence supporting the validity of the DIRSM was also produced; process skills were strongly related to patient satisfaction (Pearson's $r=.68$), whereas content skills correlated with the amount of relevant information given by the patient ($r=.57$) and diagnostic accuracy ($r=.45$).

Cognitive skills of the trainees are measured by means of a problem-solving questionnaire. Using the information gathered during the simulated patient contact, students write a (differential) diagnosis in DSM-III-R terms. They also formulate precipitating, predisposing, and maintaining factors that can explain the present problem. Based on the descriptive and explanatory diagnoses, a plan for further management is made.

To evaluate the learning process of the trainees, we ask students to write a learner report (De Groot, 1974). Trainees evaluate the simulated patient contact based on the individual learning goals that were formulated during the training. They reflect on their own behaviour, (diagnostic) thoughts, feelings, and attitudes during the encounter.

Concluding Comments

The training programme outlined inhere attempts to provide trainees with a clear model for the diagnostic interview, in which process, content, and cognitive skills are well integrated. Also, trainers are provided with a structured method to teach diagnostic interviewing skills to students.

The programme has been received positively during the past 6 years, based on student evaluations as well as teachers' enthusiasm and participation. Students (N=240) judged the training program as instructive and effective (mean scores vary between 4.2 and 4.7 on a 5-point scale) and as relevant to their profession (mean scores range between 4.7 and 4.9). Research on the training in process skills (Bögels & Kreutzkamp, 1990) showed that the approach described is highly effective in the acquisition of process interviewing skills. An important question remaining is what students learn from diagnostic interview training, with regard to content and cognitive skills. We are now studying this issue.

A number of issues with regard to diagnostic interviewing need further clarification. First, more research is desired on what constitutes effective diagnostic interviewing in the mental health field. As a result of extensive research on diagnostic expertise of physicians, utilizing the novice-expert paradigm (see Schmidt, Norman, & Boshuizen, 1990), a fairly detailed understanding now exists of what goes on during the medical interview and how expertise in this field develops. In the same way, by comparing performance of mental health workers of different levels of expertise, insight can be obtained in the nature of diagnostic expertise in the mental health field. Furthermore, the development of effective methods to teach complex skills, such as diagnostic interviewing, needs further attention. Cognitive training strategies, in which students are specifically trained to formulate, test, and refine their growing comprehension of a case, may be especially important in the acquisition of diagnostic interviewing. Finally, it is important to study how diagnostic interviewing ability develops through training in the context of the curriculum and through clinical clerkships, in order to establish the relative effects of both approaches to skills acquisition.

Research on medical interviewing and counselling skills has led to improvement of medical interviewing and counselling styles, and to effective training strategies. The teaching of diagnostic interviewing skills for the mental health professions deserves the same attention.

Chapter 3

Assessment and Validation of Diagnostic Interviewing Skills[1]

Summary

A behavioural test was developed to assess the quality of diagnostic interviewing skills of (future) mental health professionals. Two aspects of diagnostic interviewing ability are distinguished; process skills, reflecting the interpersonal and communication skills, and content skills, referring to the information-gathering ability of the interviewer. It was found that diagnostic interviewing skills can be reliably measured with respect to inter-rater reliability. However, interviewer performance on one case proved to be a poor predictor of performance on other cases. It was concluded that a large number of cases is required to obtain reliable scores of general diagnostic interviewing ability. Validity of the test was supported by the correlational analyses. Process skills were strongly related to patient satisfaction, whereas content skills were related to the amount of relevant information given by the patient and the accuracy of the diagnostic formulation and treatment plan.

[1] Submitted for publication by Bögels, S. M., Van der Vleuten, C. P. M., Blok, G., Kreutzkamp, R., Melles, R., and Schmidt, H. G. Parts of the study have been published as: Bögels, S. M., Schmidt, H. G., & Van der Vleuten, C. P. M. (1991). Ontwikkeling van een vaardigheidstoets voor het anamnestisch interview in de geestelijke gezondheidszorg: Betrouwbaarheid en validiteit. In J. Hoogstraten & W. J. Van der Linden (Eds.), *Methodologie* (pp 37-47). Amsterdam: Stichting Centrum van Onderwijsonderzoek.

Introduction

The diagnostic interview in mental health care (mostly termed psychiatric interview) constitutes the main tool for assessment and, thereby, for treatment of mental health problems. Many practitioners have documented how to conduct a psychiatric interview (see, for example Hersen & Turner, 1985; Shea, 1988). Although it is widely recognized that the development of psychiatric interview skill should be one of the main objectives in psychiatric education (Borus & Jager, 1986; Maguire, 1982; Pollock, Shanley, & Byrne, 1985), there has been little systematic study on the assessment of psychiatric interviewing performance (McCready & Waring, 1986; Rutter & Cox, 1985; Shea & Mezzich, 1988).

In contrast with the lack of empirical studies on the assessment of psychiatric interviewing, a variety of studies has been directed at medical (interviewing) performance (van der Vleuten & Swanson, 1990). Worldwide medical schools have started to use authentic assessment to measure clinical skills. Authentic assessment in medical schools usually consists of a number of "stations" at which students are required to perform a variety of simulated patient-related tasks. Authentic assessment is most often used to measure history-taking, physical examination and counselling. One important finding, consistent in almost all research on medical performance, is the low correlation between doctors' or students' performance across cases. This has been termed *the case specificity of clinical competence* (Elstein, Shulman, & Sprafka, 1978). As a result, authentic assessments in medical schools usually consist of a large number of cases to derive a reproducible judgement of skills.

A few systematic studies were directed to the assessment of psychiatric interviewing. Mumford, Schlesinger, Cuerdon, and Scully (1987) investigated ratings of videotaped simulated patient interviews to evaluate a psychiatric clerkship. They identified two types of skills: process skills, necessary to establish a positive interpersonal relationship, and content skills, referring to an information-gathering ability. Ratings on process and content interviewing skills did not or only slightly correlate with preceptor ratings of data gathering-, interaction-, and problem-solving skills. Furthermore, no relation was found between performance during the psychiatric interview and a multiple-choice end-of-clerkship psychiatric examination. Kraan and Crijnen (1987) evaluated general practice residents' diagnostic interviewing skills, based on consults with two simulated patients, one portrayed an anxiety disorder, the other portrayed an affective disorder. It appeared that the skills were very susceptible to differences between the two cases. A negative relationship was found between process skills and diagnostic accuracy. A positive relation was found between process skills and patient satisfaction, and between content skills and diagnostic accuracy. Pollock, Shanley, and Byrne (1985) studied what outcome variables were associated with the performance of clinical clerks on a psychiatric interview. Contrary to their expecta-

tions, no relation was found between performance during the psychiatric interview and overall performance during the one month psychiatric rotation, a multiple choice psychiatric examination, and patients' evaluation of the empathy of the interviewer.

In summary, diagnostic interviewing tests in the mental health field so far seem to lack validity, as low correlations were found between interviewing scores and relevant concurrent and outcome measures. Low correlations could be the result of a low true relationship between interviewing skills and other measures. On the other hand, unreliability of scores caused by case specificity of diagnostic interviewing ability could also account for the low correlations found. Not only inter-rater reliability, but also inconsistencies of interviewing behaviour across different cases (i.e. mental disorders) may be an important source of variability. If this is true, validity studies need to control for case specificity.

The diagnostic interview has two goals (Reiser, 1984). The first goal is establishment and maintenance of a good working alliance, if only to promote compliance. The second goal is collection of information to arrive at a diagnostic formulation and treatment plan. Adequate construct validity of a test of diagnostic interviewing skills is indicated if this test measures interviewing skills that contribute to these goals. Based on the definition of process and content skills, process skills can be expected to contribute mainly to the goal of establishment of a good interviewer-patient relationship, whereas content skills can be expected to contribute mainly to the goal of diagnostic information collection and adequate diagnostic formulation.

The purpose of this study was to examine a test to assess diagnostic interviewing ability of (future) mental health professionals. We wished to explore inter-rater and inter-case reliability, as well as the construct validity of the test.

Material and methods

Subjects

Twenty-two graduated students of Mental Health Sciences (University of Limburg, The Netherlands) participated in the experiment. Two persons were excluded as a result of missing data. The mean age of the 20 Ss was 23.6 yr (range 20-42), 15 were female, 5 were male. During their study they were extensively trained in psychiatric interviewing skills. The training programme consists of a basic training in process skills, followed by training in interviewing and diagnosing patients with anxiety disorders, affective disorders, and finally psychotic disorders. Interviewing simulated patients is part of the training programme.

Cases

Four standardized case histories were developed, derived from real patients. The cases concerned a widow with depressive complaints, a middle-aged man with social anxiety and a history of avoidance, a young woman with panic attacks and obsessions, and a young man with delusional thoughts and an alcohol problem. The DSM-diagnosis and differential diagnosis of the cases are presented in Table 3.1. The cases were enacted by four experienced simulated patients, each patient portrayed one case. They were trained to simulate the case with high fidelity and consistency over time. The validity of simulated patients has been demonstrated in research on medical students', residents', and counsellors' skills (Gallagher & Hargie, 1989; Norman, Tugwell, & Feightner, 1982; Sanson-Fisher & Poole, 1980).

Table 3.1. Diagnosis and differential diagnosis (DSM-III-R) of the four cases

Case A	Diagnosis	Major depression, single episode, moderate
	D.D.	Major depression, single episode, with psychotic features
		Major depression, single episode, severe without psychotic features
		Uncomplicated bereavement
Case B	Diagnosis	Social phobia, generalized
		Avoidant personality disorder
	D.D.	Personality disorder NOS
		Generalized anxiety disorder
Case C	Diagnosis	Panic disorder with agoraphobia
		Obsessive-compulsive disorder
	D.D.	Panic disorder without agoraphobia
		Anxiety disorder NOS
Case D	Diagnosis	Delusional disorder, persecutory type
		Alcohol abuse
	D.D.	Post-traumatic stress-disorder, delayed onset
		Schizophreniform disorder
		Schizophrenia, paranoid type
		Psychoactive substance-induced delusional disorder

Instruments

Diagnostic Interviewing Rating Scale in Mental Health Care (DIRSM)
A rating form was developed to measure the quality of diagnostic interviewing skills of mental health professionals (see Table 3.2). The DIRSM consists of two subscales called "process skills" and "content skills". Process scores reflected

ratings on 11 aspects[2] of interpersonal skills, such as asking questions, summarizing, thinking aloud and giving information. Content scores on the DIRSM reflected ratings on 13 aspects of information-gathering technique, including chief complaint, history of the present complaint, social and family history. Aspects were precisely defined in behavioural terms. Clear criteria for inadequate and adequate behaviour were described. Each aspect was scored on a 5-point scale from 1 (insufficient) to 5 (advanced). Finally, raters were asked to provide global judgements of process and content quality of the interviewer, both on a 10-point scale.

Table 3.2. Summary of the items of the Diagnostic Interviewing Rating Scale in Mental Health Care (DIRSM)

Process skills	Content skills
1. Attending	12. Clarification of presenting complaint
2. Open-ended questions	13. Exploration of other problems
4. Close-ended questions	14. Reason for encounter, help request
4. Concreteness	15. History of main problem
5. Reflection of feelings	16. Analyzing causal factors
6. Paraphrasing of content	17. Exploration of duration and course
7. Summarizing	18. Examination of consequences for daily life
8. Giving information	19. History of prior treatments
9. Asking for feedback	20. Psychiatric examination
10. Thinking aloud	21. Personal history
11. Structuring	22. Assessment of premorbid personality
	23. Exploration of social relations
	24. Exploration of occupational, financial, and housing conditions

Patient Satisfaction

The Patient Satisfaction with Communication Checklist (Crijnen & Kraan, 1987), originally developed to measure patient satisfaction related to physicians' communication, was modified for use in this context. Simulated patients indicated how strongly they agreed or disagreed with 18 statements, on a 5-point scale. An example of a statement is: "The interviewer encouraged me to talk about what's bothering me".

Amount of relevant information

Quantity of relevant information given by the patient was measured as follows. Of each case a list of case-specific information items was constructed, important

[2] The first version of the process scale consisted of 11 items, the later and definite version of the process scale consists of 10 items; the items "open-ended questions" and "close-ended questions" were combined to one item "asking questions".

for diagnosis and treatment planning. This list consisted of about 30 items per case. Each item was rated in one of two categories; "mentioned" or "not mentioned" by the patient during the interview.

Diagnostic accuracy
Diagnostic accuracy was measured by means of a semi-structured questionnaire (Problem Solving Questionnaire; Kraan & Crijnen, 1987). The interviewer responded to 6 open-ended questions about diagnosis, differential diagnosis, explanatory hypotheses and plans for further management.

Knowledge of the DSM-III-R
A test, consisting of 6 open-ended questions, was developed to measure knowledge of mental disorders, as classified by DSM-III-R (American Psychiatric Association, 1987). One question concerned knowledge of the DSM-III classification system, the other questions concerned general diagnostic knowledge of the disorders that were impersonated by the simulated patients.

Procedure

Subjects were instructed to conduct a diagnostic interview with the four simulated patients, each for a period of 30 minutes, in order to make a diagnosis and management plan. The interviews were videotaped. Immediately after each interview Ss were asked to complete the Problem Solving Questionnaire concerning the patient they has just seen. They were allowed to use the DSM-III-R to answer the questions about diagnosis. The patients simultaneously completed the Patient Satisfaction with Communication Checklist. After completion of the four interviews, Ss were tested on their DSM-III-R knowledge. The whole test session took 3.5 hours, and the Ss interviewed the patients in varying order.

Scoring

Four expert psychologists were asked to rate the videotaped skills of the interviewers. Raters were trained in using the DIRSM to a satisfactory inter-rater reliability level (.80). All videotapes were then independently scored by two raters. The amount of relevant information was rated by two Mental Health Sciences students. Preliminary research showed that one rater provided sufficient reliable ratings (inter-rater reliability between the two raters, based on their ratings of 9 videotapes of different Ss and different patients, was .97 (intraclass correlation)). Each videotape was, therefore, scored by only one rater. In order to judge the answers of the Ss on the Problem Solving Questionnaire a scoring key was developed by 8 psychiatric experts, consisting of accurate, semi-accurate and wrong answers. Two psychologists (others than those who scored the

videotapes) independently rated the completed Problem Solving Questionnaires of the Ss, based on the scoring key. Yet another psychologist rated the knowledge of the DSM-III-R, by means of a predetermined scoring key of accurate, semi-accurate and wrong answers.

Total scores across items were calculated for process and content skills of the DIRSM, diagnostic accuracy, patient satisfaction, and knowledge of DSM-III-R. Because the variable amount of relevant information had different total scores in different cases, a proportional score was calculated.

Statistical analysis

Reliability was analyzed by using generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson, Webb, & Rowley, 1989). Generalizability theory provides a framework for examining the dependability of behavioural measurements. In contrast to classical test-theory, that treats only one error source at a time (e.g., inter-rater reliability or test-retest reliability), generalizability theory recognizes that there may be multiple sources of error variance which define how accurately observed scores permit us to generalize about interviewers' behaviour in a universe of situations. By means of generalizability analysis (Brennan, 1983), the relative magnitude of variance caused by interviewers, cases, raters, and their interactions, was estimated. A raters-nested-within-cases-by-interviewers design was used to estimate variance components for process and content skills and diagnostic accuracy ("j:ixp" design in generalizability terminology). For the variables amount of relevant information and patient satisfaction, variance components were estimated according to a simple person-by-case design ("pxi" design). For all instruments generalizability coefficients were calculated for a variable number of cases and -where appropriate- of raters. These coefficients express the reproducibility (reliability) of the observed scores generalized to the universe of raters and cases, from which the sample of raters and cases participating in this study is taken. Inter-rater reliability of process skills, content skills, and diagnostic accuracy was estimated by calculating an intraclass correlation on all available scores of rater pairs ("i:p" design in generalizability terminology). This coefficient provides an index of inter-rater reliability that includes differences in rater stringency as measurement error.

Correlational analysis was used for validity calculations. Both observed and "true" (or disattenuated) correlations were calculated, the latter based on the generalizability coefficients found with the actual sample size used.

Results

In table 3.3, descriptive statistics of all variables are presented. The cases were judged as very realistic by the Ss (mean fidelity score was 4.2, on a scale from 1=very unrealistic to 5=very realistic), supporting the fidelity of the role-plays.

Table 3.3. Descriptive statistics of all variables across the four cases

	Mean	SD	Scale min	Scale max
Process skills	43.9	5.6	12	65
Content skills	39.4	3.6	14	75
Diagnostic accuracy	7.1	1.3	0	12
Rel. information (%)	11.9	1.4	0	100
Patient satisfaction	65.8	6.0	18	90
DSM-III knowledge	6.0	2.1	0	12

Inter-rater reliability was good to excellent; the intraclass correlation for process skills was .79, for content skills .72, and for diagnostic accuracy .84.

Table 3.4 contains the variance components of process and content skills.

Table 3.4. Absolute and proportional variance components of process and content skills, diagnostic accuracy, relevant information, and patient satisfaction (standard error in brackets)

Source*	Process skills		Content skills		Diagnostic accuracy		Relevant information		Patient satisfaction	
P	19.8 (8.3)	34%	7.1 (3.9)	15%	1.6 (0.7)	27%	.03 (.07)	3%	23.3 (11.5)	26%
C	4.6 (4.1)	8%	13.6 (9.8)	28%	1.8 (1.2)	31%	.09 (.08)	10%	13.6 (10.3)	15%
R:C	0.9 (0.9)	2%	1.9 (1.4)	4%	0.1 (0.4)	2%				
PC	21.6 (5.1)	37%	14.5 (3.9)	30%	1.5 (0.4)	25%	.74 (.14)	86%	52.3 (9.6)	59%
PR:C	11.5 (1.8)	20%	11.8 (1.9)	24%	0.9 (0.2)	15%				

*P is person, C is case, R:C is rater within case, PC is person by case, PR:C is (residual) error

The person component represents the desirable true score variance. It indicates how much subjects vary in ability. The case component reflects absolute differences between cases, and is an index of case difficulty. The rater-within-case component reflects systematic differences in rater stringency. The person-by-case component represents inconsistencies in the rank-ordering of persons across cases, the so called "case specificity problem". The last component refers to the (residual) measurement error. For both process and content skills, the person-by-case component is the largest source of variance. The "true" person variance is also large for process skills, but is much lower for content skills. Content skills appear quite sensitive to absolute differences between cases (high case variance), whereas case variance accounts for only 8 % of the variance of process skills. The general error term is a substantial source of variance for both process and content skills.

Absolute differences in rater stringency are small. Variance components of the other variables are also given.

From the variance components in Table 3.4, generalizability coefficients were calculated. In Table 3.5 generalizability coefficients are presented as a function of the number of cases and raters. Based on four cases and two raters, the sample size of the present experiment, generalizability is acceptable for process skills. Content skills require a larger number of cases (at least ten) to reach acceptable generalizability of scores, mainly as a result of the lower "true" person variance. Generalizability coefficients of the other variables are also shown. The internal consistency of knowledge of DSM-III-R was .38 (intraclass correlation).

Table 3.5. Reproducibility of process and content skills, diagnostic accuracy, relevant information, and patient satisfaction as a function of the number of cases and the number of raters

	Process skills		Content skills		Diagnostic accuracy		Relevant information	Patient satisfaction
N of cases	N of raters		N of raters		N of raters		N of raters*	N of raters*
	1	2	1	2	1	2	1	1
1	.38	.42	.21	.26	.40	.45	.04	.31
2	.55	.59	.35	.41	.57	.62	.07	.47
4	.71	.74	.52	.58	.73	.77	.13	.64
10	.86	.88	.73	.78	.87	.89	.28	.82

* For a single rating only, since only single ratings were available

Table 3.6 contains the correlations between the different variables. It demonstrates that process skills correlate strongly with patient satisfaction. Content skills correlate substantially with the amount of relevant information given by the patient, and moderately with diagnostic accuracy. Contrary to expectations, content skills do not correlate with knowledge of DSM-III-R. As different variables were rated by different judges (except for process and content skills), the relationship between the different variables is not due to the method of data collection (or halo-effect).

Observed correlations were corrected for unreliability of the different measures. As a result of the low generalizabilty of relevant information, disattenuated correlations of process and content skills with this variable reach unity.

Table 3.6. Observed correlations between process skills, content skills, relevant information, diagnostic accuracy, patient satisfaction, and DSM-III-R knowledge (correlations corrected for attenuation between brackets)

	Content skills	Relevant information	Diagnostic accuracy	Patient satisfaction	DSM-III-knowledge
Process skills	.71***	.37 (1.00)	.07 (.09)	.68*** (.99)	.00 (.00)
Content skills		.57* (1.00)	.45* (.67)	.45* (.74)	.02 (.04)
Rel. information			.40* (1.00)	.43* (1.00)	-.01 (-.03)
Diagnostic accuracy				.40* (.57)	.13 (.24)
Patient satisfaction					.16 (.33)

*** = $p < 0.001$
* = $p < 0.05$

Discussion

Results of this study show that diagnostic interviewing skills can be reliably measured with respect to inter-rater reliability. However, differences in rank-ordering of interviewers across cases of different psychiatric content, generally referred to as *the case specificity problem*, threatened overall reliability of the assessment of diagnostic interviewing performance. This is found in virtually all studies to assess clinical competence (Van der Vleuten & Swanson, 1990). What causes these differences in rankordering across cases? With regard to content skills, differences in knowledge and experience between interviewers concerning specific mental disorders can explain this phenomenon. Interestingly, however, process skills suffered even more from person-by-case interaction. Differences in proficiency to handle the interviewing process with patients with different communication styles, for example to structure a very talkative patient or to stimulate a withdrawn patient, may account for the variability of process skills across cases. Moreover, content skills appeared quite sensitive to absolute differences in cases. This may be the result of variations between cases in terms of differential-diagnostic difficulty.

These findings have important implications for the measurement of diagnostic interviewing. The variability of diagnostic interviewing skills across cases has to be taken into account in studies on psychiatric interviewing ability. For examination purposes, a large number of cases is required to obtain reliable scores of general diagnostic interviewing ability. Examination of trainees may lead to unreliable inferences, if they are based on single cases. Finally, results suggest that it is generally unnecessary to use more than one rater per interview, at least if many cases are used and different raters are used per case.

The construct validity of the DIRSM was supported by the correlational analyses. Process skills, supposed to be necessary to establish a good relationship with

the patient, were strongly related to the quality of the interviewer-patient relationship as experienced by the patient. Content skills, supposed to reflect the information-gathering ability of the interviewer, were substantially connected with the amount of relevant information given by the patient. Content skills also moderately correlated with accuracy of the diagnostic formulation and treatment plan, whereas process skills did not. Clearly, both skill aspects are, in a different way, important to meet the purposes of a diagnostic interview: building up a relationship, collecting the necessary diagnostic information, and making correct diagnostic inferences. However, as a result of the low reproducibility of amount of relevant information, correlations between relevant information and other variables have to be interpreted with caution. After correction for attenuation the amount of relevant information correlated perfectly with content and process skills, suggesting that process skills are equally important for gathering relevant patient information.

Contrary to expectation, knowledge of DSM-III-R did not correlate with content scores. Knowledge of classification of mental disorders is assumed to be a precondition for adequately interviewing a psychiatric patient, with respect to content issues. However, *knowing* facts is something different than *application* of this knowledge in real situations. Research of Glaser and others (Glaser, 1984) suggests that the knowledge of experts includes knowledge about the application of what they know. Novices, on the other hand, may lack this "application knowledge". It can be hypothesized that application knowledge of DSM-III-R will correlate with content skills.

The DIRSM seems appropriate to measure psychiatric interviewing ability. Future research on this test should concentrate on different areas. First, by means of latent structural analysis it has to be investigated whether the two aspects of diagnostic interviewing (process and content) are indeed distinguishable factors. Secondly, ways to obtain reproducible results within a limited amount of testing time need to be explored, for example, by shortening interviewing time per case. Finally, another validity question seems relevant: Can the test discriminate between novices and experts in this field? With respect to the teaching of diagnostic interviewing, more research is needed to evaluate diagnostic interview training. It has been shown that medical students and general practitioners can be effectively trained in psychiatric interviewing (Gask, Goldberg, Lesser, & Millar, 1988; Goldberg, Steele, Smith, & Spivey, 1980; Maguire, Fairbairn, & Fletcher, 1986). The effectiveness of diagnostic interview training programmes for the mental health professions still needs to be established.

Chapter 4

Measuring Diagnostic Interviewing Performance: Reproducibility of Skills

Summary

The reproducibility of diagnostic interviewing performance of undergraduate students and mental health professionals was studied under authentic assessment conditions. Subjects (N=27) interviewed four simulated patients, who portrayed various mental disorders from the domain of affective and psychotic disorders. It was found that by means of four interviews (2 hours testing time), a reproducible estimation of diagnostic interviewing ability of the interviewers can be made. These results clearly contrast with the results of a previous study, in which a large number of cases (10) was necessary to obtain reproducible results of general diagnostic interviewing ability. In the previous study, however, cases were taken from a wider range of mental disorders. Therefore, the restriction of the diagnostic categories assessed in this study appears to have contributed to the present findings. It is concluded that diagnostic interviewing performance within a restricted domain of diagnostic categories can reliably and feasibly be assessed with the proposed procedure.

Introduction

Diagnostic interviewing and accurate diagnosis of mental health disorders are important skills for mental health professionals, such as psychiatrists, clinical psychologists, and mental health scientists. These skills are initially acquired during formal training, and develop further through continued practice. Although training programmes have been developed to teach these skills to students (Bögels, 1994; Lovett, Cox, & Abou-Saleh, 1990; Shea & Mezzich, 1988), studies on the assessment and evaluation of diagnostic interviewing performance in the mental health domain (generally termed psychiatric interviewing) are scarce (McCready & Waring, 1986; Rutter & Cox, 1985; Shea & Mezzich, 1988).

This state of affairs sharply contrasts with the abundance of studies available in the domain of diagnostic interviewing in medicine (see, for an overview, Van der Vleuten & Swanson, 1991). One consistent finding of almost all research on clinical performance, is the low correlation between doctors' performance across cases. This phenomenon has been termed *case specificity of clinical competence* (Elstein, Shulman, & Sprafka, 1978). The assumption is that doctors' knowledge and skills are to a large extent case-related. In other words, dependent of the specific experience and specialisation, some doctors perform better on one case, whereas other doctors perform better on another case. Therefore, performance on one case is a poor predictor of performance on other cases. In order to overcome this case-specifity, performance-based tests in medical schools usually consist of a large number of cases (10-40) to derive a reproducible judgment of skills.

The few studies that have been undertaken with regard to the assessment of psychiatric interviewing performance assessed inter-rater reliability, but failed to account for case-specificity (Mumford, Schlesinger, Cuerdon, & Scully, 1987; Pollock, Shanly, & Byrne, 1985). However, it can be hypothesized that not only inconsistencies between raters, but also inconsistencies of interviewing behaviour across different cases (i.e. psychiatric disorders) are an important source of variability. The assessment of diagnostic interviewing performance was studied by Bögels et al. (1994). Twenty advanced undergraduate students of Mental Health Sciences conducted a diagnostic interview with four different simulated patients, portraying four mental disorders; social phobia, panic disorder, major depression, and delusional disorder. Inconsistencies in performance across cases ("case-specificity") proved to be a far larger source of variance than inconsistencies in judgement across raters. Results suggested that a large number of cases (10) is therefore required to obtain reproducible scores of general diagnostic interviewing ability.

However, an examination consisting of 10 encounters with simulated patients generally exceeds the time and money available for this purpose. Therefore, it is

important to look for ways to shorten such tests, without losing reliability. One possible way to reduce testing time is to restrict the number of diagnostic categories to be assessed during the examination.

In the study of Bögels et al. (1994) only students were included. The conclusions of that study cannot automatically be generalized to psychiatric residents or mental health professionals. Growth in experience may increase case-specificity of performance, as expertise is highly case-related (Schmidt, Norman, & Boshuizen, 1990). On the other hand, professionals may differ more from each other than students, as they more and more develop their own style in interviewing patients, which may be relatively stable across patients.

The purpose of the present study was to investigate the reproducibility of a behavioural test to assess diagnostic interviewing performance of (future) mental health professionals. In contrast to the Bögels et al. (1994) study, which included three categories of disorders (anxiety disorders, affective disorders, psychotic disorders) the cases used were restricted to two diagnostic categories (affective and psychotic disorders). Furthermore, the subjects that were assessed in the present study consisted not only of students, but also of mental health professionals.

Material and methods

Subjects

Twenty-seven subjects participated in the experiment, 15 students of Mental Health Sciences and 12 mental health professionals. The mean age of the students was 22.5 years (range 20-29), 2 men and 13 women. During their study they were extensively trained in psychiatric interviewing and diagnosing mental health disorders. In addition, they had substantial experience in interviewing simulated patients. The professionals were 5 psychiatrists, 6 clinical psychologists, and 1 social worker. Their mean age was 34.8 years (range 25-43), and their mean experience with psychiatric interviewing was 7.1 years (range 2-13). Nine of them were men and 3 women.

Cases

Four standardized case histories were developed, based on real patients. The cases concerned a man with a bipolar disorder, who was currently in a manic episode with psychotic features, a divorced woman with recurrent depressive episodes, a young woman with an Amphetamine-induced delusional disorder, and a man with a schizophrenic disorder. The cases were enacted by four simu-

lated patients, each patient portraying one case. They were extensively trained to simulate the case with high fidelity and constancy over time.

Instruments

The Diagnostic Interviewing Rating Scale in Mental Health Care (DIRSM), developed by Bögels and colleagues (1994) was used to measure the quality of diagnostic interviewing skills. The DIRSM consists of two subscales called "process skills" and "content skills". Process scores reflected ratings on 10 aspects of interpersonal skills, such as asking questions, summarizing, thinking aloud and sharing information. Content scores on the DIRSM reflected ratings on 13 aspects of information gathering technique, including clarification of present complaints, history-taking of the main complaint, psychiatric examination, social and family history-taking. Each aspect is scored by raters on a 5-point scale from 1 (insufficient) to 5 (advanced). In addition, raters provide global judgements of process and content quality of the interviewer, both on a 10-point scale.

Diagnostic accuracy was measured by means of a semi-structured questionnaire, validated in previous research, the Problem Solving Questionnaire, developed by Kraan and Crijnen (1987). This instrument requires the interviewer to respond to 7 open-ended questions about diagnosis, differential diagnosis, explanatory hypotheses and plans for further management.

Procedure

Subjects were instructed to conduct a psychiatric interview with the four simulated patients, each for a period of 30 minutes, in order to arrive at a diagnosis and a management plan. The interviews were videotaped. Immediately after each interview, Ss were asked to complete the Problem Solving Questionnaire concerning the patient they had just seen. They were allowed to use the DSM-III-R (American Psychiatric Association, 1987) to answer the questions about diagnosis and differential diagnosis. The whole test session took 3 hours, and the Ss interviewed the patients in varying order.

Scoring

Four skilled raters evaluated the skills displayed by the interviewers. To that end, they were trained in using the DIRSM to a satisfactory level of inter-rater agreement. Inter-rater reliability, based on the ratings of all four raters of 10 Ss interviewing different simulated patients, was .78 for process skills, and .86 for content skills. The inter-rater reliability is expressed as an intraclass correlation, differences in rater stringency are treated as measurement error. To evaluate the answers of the Ss on the Problem Solving Questionnaire, a scoring key was

developed by 5 psychiatric experts, consisting of accurate, semi-accurate and wrong answers for each question. Based on this scoring key, the completed Problem Solving Questionnaires of the Ss were rated by another rater.

Total scores across items were calculated for process and content skills and diagnostic accuracy separately.

Statistical analysis

Reproducibility was analyzed by using generalizability analysis (Cronbach, Glaser, Nanda, & Rajaratnam, 1972). Through generalizability analysis, the relative magnitude of variance caused by interviewers, cases, and their interactions, can be estimated. Generalizability analysis furthermore enables one to estimate reproducibility coefficients given fixed numbers of cases. Variance components were estimated by using a person-by-case design ("pxi" in generalizability terminology). First, generalizability coefficients were calculated according to a norm-oriented score interpretation. In this perspective, scores are given an interpretation relative to one's standing among a group of examinees. This enables the assessor to make relative decisions with regard to pass or fail of examinees. Next, dependability coefficients were calculated according to a domain-referenced score interpretation. In this perspective, the observed score is interpreted absolutely, without regard to the performance of others (Shavelson, Webb, & Rowley, 1989). The coefficients reflect the total reproducibility of the test, including inter-rater reproducibility, since each case was assessed by a different rater. All coefficients were calculated for a variable number of cases.

Results

In Table 4.1, descriptive statistics are provided. Table 4.2 contains the variance components attributable to the different variables involved. The person component represents the (desirable) true score variance. It indicates to what extent subjects vary in diagnostic interviewing ability. The person variability is quite sizable for all measures, and larger than is usually the case in similar performance-based tests (van der Vleuten & Swanson, 1991). The case component reflects absolute differences between cases, and is an index of case difficulty. This component is small, so there were no large differences in case difficulty. The person-by-case component represents inconsistencies in the rank-ordering of persons across cases, the forementioned "case-specificity" phenomenon. It also includes inconsistencies across raters, as raters were nested within cases. The person-by-case component is large, but smaller than generally found in authentic assessment. As the inter-rater reliability estimates were rather high, it is assumed that the factor represents mainly differences between persons across cases.

Table 4.1. Descriptive statistics of all variables across the four cases for all Ss (N=27) and separately for students (N=15) and professionals (N=12)

Group	Variable	Mean	SD	Scale min	Scale max
Total	Process skills	48.3	5.2	11	60
	Content skills	42.0	6.8	14	75
	Diagnostic accuracy	7.0	2.2	0	14
Students	Process skills	45.6	4.2		
	Content skills	37.5	4.9		
	Diagnostic accuracy	5.7	1.6		
Profes-	Process skills	51.6	4.4		
sionals	Content skills	47.6	3.8		
	Diagnostic accuracy	8.7	1.7		

Table 4.2. Absolute and proportional variance components (standard errors in brackets) of process and content skills and diagnostic accuracy, for all Ss (N=27) and separately for students (N=15) and professionals (N=12)

Group	Variance components	*Process skills*	*Content skills*	Diagnostic accuracy
Total	Person	21.5 (7.3) 48 %	40.0 (12.3) 58 %	43.5 (13.2) 63 %
	Case	0.7 (0.3) 2 %	5.5 (4.0) 8 %	2.5 (2.2) 4 %
	Person-by-case*	22.4 (0.9) 50 %	22.9 (3.6) 33 %	23.3 (3.7) 34 %
Students	Person	12.2 (6.3) 31 %	17.8 (8.7) 35 %	18.5 (8.8) 39 %
	Case	6.2 (1.2) 16 %	7.2 (2.8) 14 %	5.0 (4.3) 10 %
	Person-by-case*	21.5 (1.1) 54 %	26.4 (2.8) 51 %	25.0 (5.3) 52 %
Profes-	Person	15.4 (7.8) 47 %	11.3 (5.8) 34 %	25.1 (11.8) 48 %
sionals	Case	0.0 (0.8) 0 %	8.3 (6.9) 25 %	7.4 (1.6) 14 %
	Person-by-case*	17.6 (4.2) 53 %	13.5 (3.2) 41 %	19.9 (4.8) 38 %

*The person-by-case component includes also differences between raters, since each case was rated by a different (single) rater.

From the variance components in Table 4.2, generalizability coefficients were calculated, adopting a norm- and domain-referenced score interpretation. In Table 4.3 reproducibility coefficients are presented as a function of the number of cases. Coefficients were calculated for the group as a whole and separately for the students and professionals. Based on four cases, the sample size of the present experiment, reproducibility is satisfactory for all variables. Reproducibility for the separate groups is somewhat lower than for the group as a whole, and reproducibility is higher for professionals than for students. Content skills and diagnostic accuracy tend to reach higher reproducibility than process skills.

Table 4.3. Reproducibility of process skills, content skills, and diagnostic skills as a function of the number of cases, from a norm- and domain-oriented perspective, for all subjects (N=27) and separately for students (N=15) and professionals (N=12)

		Process skills		Content skills		Diagnostic accuracy	
Group	N of cases	G*	D**	G	D	G	D
Total	1	.49	.48	.64	.59	.65	.63
	2	.66	.65	.78	.74	.79	.77
	4	.79	.78	.87	.85	.88	.87
	10	.91	.90	.95	.93	.95	.94
Students	1	.36	.31	.40	.35	.42	.38
	2	.53	.47	.57	.51	.60	.55
	4	.69	.64	.73	.68	.75	.71
	10	.85	.81	.87	.84	.88	.86
Profes-sionals	1	.47	.47	.46	.34	.56	.55
	2	.64	.64	.63	.51	.72	.71
	4	.78	.78	.77	.68	.83	.83
	10	.90	.90	.89	.84	.93	.92

* Generalizability coefficient (norm-oriented)
** Dependability coefficient (domain-oriented)

Discussion

Results of this study show that diagnostic interviewing performance can be accurately measured using approximately 4 cases (2 hours testing time) and a different rater per case. These results are in clear contrast to those of a previous study (Bögels et al., 1994), in which it was found that content skills require at least 10 cases to reach minimal reproducible levels. The fact that the present group consisted of a mixture of students and professionals did not account for the results, since reproducibility was still high after separation of the group into students and professionals. Therefore, the restriction of diagnostic categories (only affective and psychotic disorders were included in the present study, whereas affective, psychotic, and anxiety disorders were included in the prior study) appears to have increased reproducibility of scores. A disadvantage of restriction of diagnostic categories is, of course, that the observed scores permit us to generalize to a smaller universe of diagnostic categories, namely the universe of psychotic and affective disorders.

As a result of the low absolute differences between case-difficulty (low case variance), reproducibility of scores from a norm-oriented perspective was hardly better than reproducibility from a domain-oriented perspective. For examination purposes, it is important to reflect on the advantages and disadvantages of both

types of judgements. Since there are no clear criteria yet with regard to adequate or inadequate psychiatric interviewing performance, the best choice for the moment may be a norm-oriented judgement, in which scores are given a meaning relative to an examinee's standing in the group. In medical schools, however, substantial work has been done in developing standards for adequate and inadequate performance (Norcini, Stillman, Regan, Sutnick, Williams, & Friedman, 1992). If such standards can validly be established, and differences in case-difficulty are controlled across exams, a domain-based judgement of clinical performance has important advantages. It provides examinees clarity with regard to the level of performance which is required in order to pass the test. Furthermore, domain-based judgement makes examinees independent of the (accidental) level of skills of the group of which they are part.

In conclusion, the results of the present study suggest that by means of authentic assessment various aspects of diagnostic interviewing proficiency can be measured with sufficient reliability and feasibility. Objective assessment of diagnostic interviewing performance seems indeed possible.

Chapter 5

Authentic Assessment of Interviewing and Counselling Skills: Effect of Testing Time per Station on Reproducibility and Validity[1]

Summary

The purpose of this study was to evaluate the effects of station length on the reliability and validity in a test for interviewing and counselling skills. Fifty Health Sciences students were randomly assigned to one of three conditions: four 5-minute interviews, four 10-minute interviews, and four 20-minute interviews. Generalizability analyses showed that after standardizing for time the highest generalizability was obtained with 5- and 10-minute stations. The lowest generalizability was found with 20-minute stations. Differences between persons, or true score variance, tended to decrease in 20-minute stations. Inter-rater reliabilities were rather low, and were not influenced by station length. Validity analyses revealed that students tend to do "more of the same" in longer stations. The quality of the problem-analyses was higher in 20- than in 10-minute interviews, but no differences were found between 5- and 10-minute interviews. We conclude that 10-minute stations are the optimal choice for assessing interviewing and basic counselling skills.

[1] Bögels, S. M., Van Mourik, T. G. C., & Van der Vleuten, C. P. M. (1994) Authentic assessment of interviewing and counselling skills: Effect of station time on reproducibility and validity. *Teaching and Learning in Medicine* (in press).

Introduction

One of the main problems of authentic assessment is the low correlation between examinees' performance on different cases. As a result, tests usually consist of many cases to obtain reproducible results. Authentic assessment is therefore unfeasible and prohibitly expensive in many evaluation settings (Shatzer, 1991). With regard to the assessment of interviewing skills, at least 6-12 cases (3-4 hours testing time) are needed to obtain reproducible results; that is a generalizability greater than .80. (Bögels et al., 1994; Stillman, Regan, & Swanson, 1987; Stillman et al., 1992; Van Thiel, Kraan, & Van der Vleuten, 1992).

It is therefore important to look for ways to reduce testing time of authentic assessment, without compromising reliability and validity. One alternative is to reduce testing time per single case. Testing time per case, however, is primarily a validity issue. The time, necessary for an encounter with a simulated patient, is determined by 1) the task required, 2) the complexity of the case, and 3) interviewers' level of expertise. Nevertheless, tasks and instructions may be varied in such a way that they require more or less time per station. To improve efficiency, it is important to investigate the effects of differences in testing time per case on the generalizability of the results. Post-hoc comparisons using shorter and longer stations reveal that longer stations tend to yield more measurement information, but the fact that more short stations can be completed in a fixed length of time partly compensates for this effect (Van der Vleuten & Swanson, 1990).

Shatzer, Wardrop and colleagues conducted two studies to investigate this issue directly. In a first study (Shatzer et al., 1992), medical students conducted simulated patient interviews in 20 minutes. Scorings of raters after 5, 10 and 20 minutes were obtained. It was found that 10-minute scorings produced higher generalizability than 5-minute or 20-minute scorings. By contrast, in a second study (Shatzer, Wardrop, Williams, & Hatch, 1994) in which actual 10-minute stations were used, with a scoring after 5 minutes, 5-minute scoring resulted in higher generalizability than 10-minute scoring. In both studies, performance was judged by only one rater. It remains unclear whether differences in reproducibility are the result of station length influencing inter-rater reliability rather than station length influencing performance. It can be hypothesized, for example, that differences between raters become larger in case of longer stations, as raters may be unable to remain concentrated for the full 20 minutes. A second limitation of both Shatzer et al. studies is that the shorter stations were artificially created, as the actual length of the stations was 20 minutes in the Shatzer et al. (1992) study, and 10 minutes in the Shatzer et al. (1994) study. Students, as well as simulated patients would probably approach actual 5-minute stations differently from actual 10- or 20-minute stations.

Station length may also influence the validity of the examination. Shatzer et al. (1994) found that 88 % of the questions students asked, were asked during the first 5 minutes. This finding suggests that longer stations yield more redundant information. Other issues concerning the validity effects of station length need further clarification. For example, although students may ask most questions at the beginning, what happens to the quality of the information collected and the quality of data interpretation in shorter stations? Patient information, gathered in shorter stations, may be incomplete, resulting in faulty or superficial analyses of the case. Furthermore, is the quality of the relationship between interviewer and patient influenced by the length of the station? Simulated patients may feel less understood in shorter stations as a result of the speed of the interviewer and the short time for establishing a working alliance.

The purpose of the present study was to investigate these issues with regard to the assessment of interviewing and counselling performance. First, we wished to explore the effects of different length interviews on the generalizability of the test results. The effects of station length on different sources of measurement variance (examinees, raters, and cases) were analyzed, using *actual* 5, 10, and 20-minute stations. Secondly, the consequences of decreasing time per interview with regard to the validity of the test were studied. It was investigated how station length influences students' interview behaviour, the quality of students' problem analyses, and simulated patients' satisfaction with students' communication skills.

Methods

Subjects and conditions

Fifty (of 230) first-year students of Health Sciences[2] voluntarily participated in the experiment. They were randomly assigned to one of three conditions: four interviews of 5 minutes, four interviews of 10 minutes, or four interviews of 20 minutes. Ss were matched with regard to sex and age. In the 5-minute condition, 17 Ss participated, 3 men and 14 women, mean age 20.2 (range 18-33 year). In the 10-minute condition, 17 Ss participated, mean age 19.4 (range 18-25), 3 men and 14 women. In the 20-minute condition, 16 Ss participated, 2 men and 14

[2] The Maastricht Faculty of Health Sciences brings together a variety of studies in the health field: Mental Health Sciences, Health Promotion, Nurse Sciences, Biological Health Sciences, Movement Sciences, Theory of Health Sciences, Management of Health Care, and Health Education. The medical curriculum is not part of Health Sciences, but takes place at the Medical Faculty.

women, mean age 19.4 (range 18-22). Ss did not differ significantly with regard to age (F (2,47)=.48, p=.621).

Prior to the experiment, Ss attended a training course in basic interviewing skills, mainly based on the microcounselling model of Ivey (1971). The training consisted of six 3-hr sessions, was organized in 23 groups of 10 students, and was guided by advanced students of Health Sciences.

Cases

Four cases were developed by the authors, based on real patients. The first case concerned a 27-year old man, complaining about pain, who felt extremely responsible in his work as an electrician. The second was a 51-year old divorced woman with sleeping problems and tantrums. The third was a 33-year old woman working as a psychiatric nurse, who had panic attacks and nightmares after being attacked by a patient at work. The fourth case was a 39-year old man employed as a truck driver, with hypertension, who worried about his daughter and about losing his job. The cases were enacted by eight simulated patients, two patients per case. Each pair was extensively trained to portray the case consistently over time.

Instruments

A rating form was used to assess the skills of the Ss. Furthermore, three instruments were used to assess in what way station length influences the validity of the test; a patient satisfaction questionnaire, a measure to establish the quality of the problem analysis, and ratings of case fidelity and case difficulty.

The Interviewing and Counselling Skills Rating Scale (ICSR) is a rating form to measure the quality of interviewing and basic counselling skills (Bögels, 1992; Bögels et al., 1994). The scale consists of 10 items, each item representing a skill to be used during a problem-based interview. The skills are: 1) attending, 2) asking questions, 3) concreteness, 4) reflection of feelings, 5) paraphrasing of content, 6) summarizing, 7) giving information, 8) asking for feedback, 9) thinking aloud, and 10) structuring. For each skill detailed criteria are described, dealing with *how* and *when* to use the skill. Table 5.1 provides an example of the scoring of an item.

At the end of the interview, each skill is rated on a 5-point scale ranging from 1 (unacceptable) to 5 (advanced). Finally, raters provide a global judgement of the quality of the interviewing performance, on a 10-point scale. In prior research (Bögels et al., 1994) high inter-rater reliabilities were found using the ICSR (intraclass correlation of .79). Scores on the ICSR were furthermore related to patient satisfaction with the communication (r=.68) and to the amount of relevant information given by the patient (r=.37).

Table 5.1. Example of a (scored) item of the Interviewing and Counselling Skills Rating-scale (ICSR). Every time the interviewer uses the skill "giving information" the observer judges to what extent the intervention matches with the criteria of quality and timing and subsequently draws a line somewhere between 1 (unacceptable) and 5 (advanced). At the end of the interview, the observer gives a final rating for overall use of the skill "giving of information" by circling one of the five numbers. The lines serve as an aid to memory on which observers base their final rating.

7. Giving information	1	2	(3)	4	5
Criteria for quality:					
- clear and concrete					
- in small units	///		/		///
- tuned in to the patient					
Criteria for timing:					
- when indicated: introduction, transitions, end, questions of patient					
- well dosed					

In order to assess the extent to which the Ss had established rapport with the patient, a patient satisfaction questionnaire was used, validated in previous research (Bögels, 1992; Bögels & Kreutzkamp, 1990). The questionnaire consists of 9 statements, which are scored by patients on a 5-point scale from 1 (totally disagree) to 5 (totally agree). An example of a statement is: "The interviewer encouraged me to talk about what is bothering me".

To assess the comprehension of the cases by the Ss, a problem-analysis quality measure was used (Robinson & Halliday, 1988). Ss were instructed to describe the aspects of the patient's problem, the relative importance of the aspects, and the factors that caused and/or maintained the problems. A model case report was written for each case. Problem analyses of the Ss were judged on five categories: 1) accuracy of the description of the problem aspects, 2) completeness of the description, 3) clarity of the description, 4) correspondence of relative importance of aspects with the model case report and soundness of arguments for relative importance, 5) correspondence of causes and maintaining factors with the model case report and soundness of arguments for causes and maintaining factors. Each category was scored on a 5-point scale ranging from 1 (insufficient) to 5 (advanced).

In order to check for possible differences in the quality of the simulations between conditions, Ss were asked to indicate how realistically the patient portrayed the role on a 5-point scale (1=very unrealistic/low fidelity to 5=very realistic/high fidelity). To control for possible differences in perceived case-

difficulty between conditions, Ss indicated how difficult they found each case (1=not difficult to 5=very difficult).

Procedure

Ss were instructed to conduct interviews to clarify the problem with each of the four simulated patients. They were informed about the amount of time they had for each interview (5, 10, or 20 minutes, depending on the condition). They were furthermore encouraged to apply all interviewing skills acquired during their prior training. The interviews were videotaped. During the experiment, the authors observed the roleplays in a central registration room, in order to check whether the simulated patients (two per case) portrayed the case in conformity with instructions. Immediately after each interview Ss wrote their problem analyses. In all conditions, they received 10 minutes for this task. They also answered the questions concerning fidelity and difficulty. Simultaneously, the simulated patients completed the patient satisfaction questionnaire. Ss interviewed the patients in random order.

Scoring

Raters for the ICSR were eight advanced students of Health Sciences, familiar with teaching interviewing skills to students, including evaluating students' interviewing performance by means of the ICSR. After a short retraining in using the ICSR, two raters independently rated the videotaped interviews. Raters scored a different case for each time condition.

The problem analyses of the Ss were rated by two other raters (a Health Science student and a Health Scientist). They received a short training in rating the written problem analyses by means of the categories, described earlier. Each problem analysis was scored by only one rater. They were blind to condition. The inter-rater reliability, obtained from a sample of 16 problem analyses from different Ss with different cases, that were rated by both raters, was .64 (intraclass correlation).

Total scores across items for all variables were used for further analyses.

Statistical analysis

Generalizability theory was used for reliability analysis (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). To estimate inter-rater reliability, scores of all rater pairs across all cases were subjected to a random rater-nested-within-persons design ("i:p" design in generalizability theory terminology). Intraclass correlations were calculated (for a single rating). This coefficient provides an index of inter-rater reliability that includes differences in rater stringency as measurement error. A random rater-nested-within-cases-by-persons ("j:ixp")

design was used to estimate overall reliability of interviewing performance. Variance components were estimated and generalizability coefficients were calculated for a variable number of cases and raters (according to a norm-referenced perspective, scores are given relative meaning to each other). These generalizability coefficients can be taken as the expected correlation between this test and a hypothetical other test consisting of a random sample of other raters, cases, and simulated patients.

Results

Reliability

Inter-rater reliability with regard to the quality of the interviewing and counselling skills was rather low, the intraclass correlation was .39 for 5-minute stations, .35 for 10-minute stations, and .35 for 20-minute stations.

Table 5.2 contains the variance components of interviewing performance for the three conditions.

Table 5.2. Estimated absolute and proportional variance components (standard error between brackets) of the quality of skills, problem-analysis quality and patient satisfaction at different station lengths.

Variable	Source*	5 min		10 min		20 min	
Quality of skills	P	6.8	(3.1) 24%	13.1	(5.4) 31%	2.7	(2.7) 7%
	C	1.7	(3.3) 6%	0.0	(4.1) 0%	0.0	(3.0) 0%
	R:C	4.7	(3.1) 16%	10.4	(6.5) 24%	7.5	(4.8) 21%
	PC	3.5	(2.2) 12%	4.1	(2.7) 10%	11.5	(4.1) 32%
	PR:C	12.0	(2.1) 42%	15.3	(2.7) 36%	14.5	(2.6) 51%
Problem-analysis**	P	3.6	(1.8) 30%	5.2	(2.3) 46%	0.6	(1.1) 6%
	C	1.4	(1.1) 12%	0.0	(0.1) 0%	0.2	(0.5) 2%
	PC	7.0	(1.4) 58%	6.1	(1.2) 54%	9.0	(1.9) 92%
Patient-satis-faction**	P	7.3	(5.2) 15%	29.1	(12.7) 44%	8.7	(5.5) 20%
	C	13.3	(9.5) 26%	2.4	(2.8) 4%	9.1	(6.8) 20%
	PC	29.8	(6.0) 59%	34.3	(6.8) 52%	26.8	(5.5) 60%

*P=person, C=case, R:C=rater within case, PC=person by case, PR:C=(residual) error.
**For the variables problem-analysis quality and patient satisfaction only P, C, and PC are given, since these variables were scored by a single rater.

The person component represents the (desirable) true score variance, it indicates how much persons vary in ability. The person variance is highest in 10-minute stations, and quite low in 20-minute stations. The case component is an index of case difficulty; the variability between average scores on each case. The

pure case variance is negligible in all conditions. The rater-within-case component reflects differences in rater stringency (within cases), and is a large source of variance in all conditions. The person-by-case component reflects inconsistencies in the rank-ordering of persons across cases; the so-called "case-specificity issue". It is rather small, but is largest in 20-minute stations. The last component refers to the (residual) measurement error. The error term is large in all conditions. The variance components of the quality of problem-analysis and patient satisfaction are also given.

From the variance components, generalizability coefficients were calculated as a function of the number of cases and raters, and aggregated to 1 hour testing time (three interviews of 20 minutes, six interviews of 10 minutes, and 12 interviews of 5 minutes). With the actual design, two raters and four cases, satisfactory overall reliability (generalizability coefficient greater than .80) for quality of skills is accomplished only in the 10-minute stations (see Table 5.3). After standardization to 1 hour of testing time, generalizability of 5-minute stations is somewhat better than generalizability of 10-minute stations, the difference is, however, small. Generalizability coefficients of the problem-analysis quality and patient satisfaction are also given. For both variables, 10-minute stations produce the most favourable generalizabilities.

Table 5.3. Actual and one-hour testing time generalizability coefficients of the quality of skills, problem-analysis quality and patient satisfaction for the three station lengths.

Variable	N of cases	5 min		10 min		20 min	
		1 rater	2 raters	1 rater	2 raters	1 rater	2 raters
Quality of skills	4 cases	.64	.74	.73	.82	.30	.37
	1 hour	.84	.90	.80	.87	.24	.31
Problem-analysis*	4 cases	.67		.77		.21	
	1 hour	.86		.84		.17	
Patient-satisfact.*	4 cases	.50		.77		.56	
	1 hour	.75		.84		.49	

*For the variables problem-analysis quality and patient satisfaction generalizability coefficients are only provided for one rater, since these variables were scored by a single rater.

Validity

In Table 5.4 mean scores per condition on all variables are given, averaged across cases and -where appropriate- across judges.

Table 5.4. Mean total scores, standard deviations (between brackets), and scale minimum and maximum, averaged across cases and -where appropriate- across raters of all variables at the different station lengths.

	5 min	10 min	20 min	Scale min.	Scale max.
N of interventions	18.1 (2.6)	27.7 (4.8)	41.7 (4.5)	0	∞
Quality of skills	36.9 (3.3)	37.5 (4.3)	37.3 (3.1)	10	55
Patient satisfaction	42.7 (3.9)	42.7 (6.1)	43.1 (3.9)	10	55
Problem analysis	12.4 (2.3)	12.5 (2.6)	14.9 (1.7)	5	25
Case difficulty	2.6 (0.7)	2.8 (0.7)	2.8 (0.6)	1	5
Case fidelity	4.3 (0.5)	4.3 (0.4)	4.2 (0.8)	1	5

As was expected, the number of interventions increases when more time is available for the interview (F (2,47)=139.0, p=.000). However, as is shown in Figure 5.1, students generally do "more of the same" if they have more time; they mainly ask more concretising questions, the already most frequently used skill in the 5-minute condition. One would expect that higher-order skills, like "reflection of feelings" or "thinking aloud", would be used more frequently in longer interviews. On the contrary, these skills are hardly used more often when more time is available.

No mean differences were found for quality of skills (F (2,47)=.1, p=.875) or patient satisfaction with the communication (F (2,47)=.0, p=.953) across conditions. In accordance with expectations, a significant difference was found in problem analysis quality between conditions (F (2,47)=6.6, p=.003) in favour of the longer stations, but this difference was only found between 10- and 20-minute stations, and not between 5- and 10-minute stations. No significant differences were found in case fidelity (F (2,47)=.1, p=.925) and case difficulty (F (2,47)=.9, p=.418) between conditions. Therefore, station length did not influence the fidelity of the role-plays and the complexity of the patient-case as experienced by the Ss.

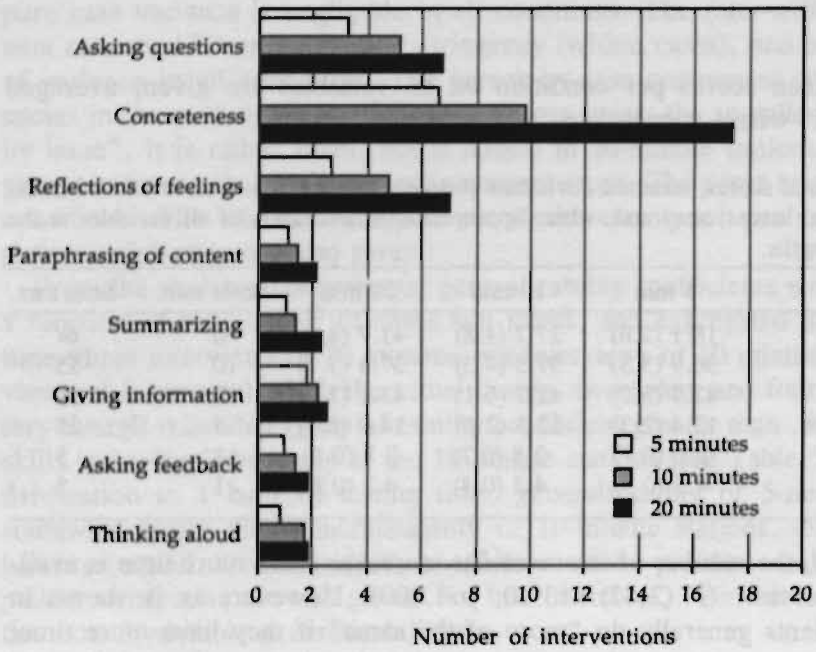

Figure 5.1. Number of times each skill is used at the three station lengths (averaged across cases and across raters). The number of interventions with regard to "attending" and "structuring" were not scored, as "attending" happens constantly and "structuring" is a composite skills in the sense that the other skills all contribute to it.

Discussion

The first purpose of this study was to investigate the effect of different length interviews on the reproducibility of results. It can be concluded that station length does influence generalizability. Generalizability coefficients of 5- and 10-minute stations appeared highest. The desirable true person variance was highest in 10-minute stations, and *lowest* in 20-minute stations. Results need to be interpretated cautiously, since the small numbers of Ss, combined with the large standard errors around the variance components, limit the power of the results. However, our results correspond with those of a recent, well-designed study of Shatzer, Darosa, Colliver, and Barkmeier (1993) in which generalizability was studied directly. They also found that shorter stations produced more variability among students' performance than longer stations. If students are offered more time, differences tend to disappear. Therefore, in order to value efficiency in performance decisions, it may be important to restrict time.

The second purpose of the study was to investigate the effects of station length on the validity of the test. Analysis of the interview patterns revealed that

beginning (but trained) students tend to do "more of the same" when they have more time, at least with regard to interviewing and basic counselling skills. They use, for example, the already most frequently applied skill "concreteness" more often when more time is available, but hardly apply the sparingly used skills like "reflection of feelings" or "thinking aloud" more often when they have more time. In conclusion, the validity analyses gave further evidence to the notion that longer stations do not necessarily yield more measurement information but seem to provide redundant information with regard to assessing interviewing performance. Possible negative side effects of shorter stations, such as problems in establishing rapport or low fidelity simulations were not found. It should, however, be noted that the focus of our assessment was interpersonal skills, and that our results cannot be generalized to history-taking skills in general. In establishing history-taking performance longer time may be necessary, dependent on the time needed to conduct the whole task.

As was expected, longer stations produced higher-quality problem analyses (comparing the 10- and 20-minute stations). When students are given more time to interview patients, they achieve a better understanding of their problems, probably because they have obtained more information about the patient. The lack of difference in problem analyses quality for the 5- and 10-minute stations is puzzling in this respect. A tentative explanation may be that simulated patients disclose their problems more easily and quickly to the interviewers in the 5-minute condition than in the 10-minute condition, resulting in a equal quality problem analysis after 5 or 10 minutes interviewing time. Our findings are in line with those of Shatzer et al. (1993) who found no performance differences in the written responses about important findings and diagnostic considerations between 5- and 10-minute stations. Furthermore, although 20-minute stations produced higher-quality problem-analyses, generalizability for this measure was low in the 20-minute stations, compared to the 5- and 10-minute stations, mostly attributable to less variability among students' performances. Therefore, if the purpose of an authentic assessment is to estimate problem analysis quality (or comparable outcome measures such as the amount of relevant information obtained, diagnosis, management plan, etc.) shorter stations may produce more reliable scores.

Station length did not influence inter-rater agreement. A general complication of this study was the rather low inter-rater reliability of the skills-ratings, regardless of condition. Inter-rater reliability can be influenced by at least four factors; 1) the scale used, 2) the type of skills rated, 3) the raters used, and 4) the training of raters. First, the instrument used in this study to measure interviewing skills was a global rating form. Global ratings may give more room for subjective interpretation and, as a result, produce lower inter-rater reliability (Van Luyk & Van der Vleuten, 1992; Streiner, 1985). However, prior research with this scale yielded fairly high reliabilities (Bögels et al., 1994; Bögels, 1994). Recent research by Van Thiel, Van der Vleuten, and Kraan (1992) also showed that high inter-rater reliabilities can be achieved with similar global rating forms.

Secondly, basic interviewing and counselling skills may be more difficult to rate than content-related skills such as history-taking. This explanation is not supported by different studies in which inter-rater reliability of "process skills" was as high or higher than that of "content skills" (Bögels et al., 1993; Van Thiel, Kraan, & Van der Vleuten, 1991). However, in the studies mentioned, process and content skills were both rated by the same raters. It might well be more difficult to agree upon the quality of process skills separate from content. Thirdly, our use of student raters may have caused a problem. Although we selected advanced students with experience in training and assessing students' process skills, their base-line is not as well established as that of the teaching staff, who have observed far more students. As a consequence, student-raters may score with greater variation. Fourthly, the rater training was probably insufficient. Our results are however comparable to, for example, the results of Stillman et al. (1991), who reported an inter-rater reliability for interview scores of only .26. In conclusion, results of this study indicated that absolute differences between raters are an important source of error and need more attention during rater-training and in research on authentic assessment.

Contrary to expectations, case-specificity of performance did not appear to be a great problem, except for the 20-minute stations. A first explanation for this finding can be found in the type of skills measured. Basic interviewing skills may not be very case-dependent, as these skills are less related to case-content. However, differences in proficiency to handle the interviewing process with patients with different communication styles, for example to structure a talkative patient or to stimulate a silent patient, become more pronounced when more time is available for the interview. A second explanation for the low case-specificity found in the present study, compared to most studies in the medical domain, concerns the type of instrument used to assess interviewing performance. We used a station-independent rating form, whereas in studies on clinical skills of medical students, generally, checklists are used which are tailored to station content. It has been demonstrated that such checklists produce larger variability across stations (Van Luyk & Van der Vleuten, 1992).

When standardized to 1 hour of testing time, 5-minute stations produced somewhat higher generalizability coefficients than 10-minute stations, but this is mainly the result of a larger rater-sample. An alternative strategy, which is suggested by the present results, is to increase the number of raters within stations. It should be realized that in comparing reliability findings by standardizing for time, that each time condition requires quite different resources. Tests using 5-minute stations require considerably more raters, patients, training time, room facilities etc, than longer stations. Moreover, beginning students may feel hurried when they have so little time. Taken together, it seems therefore that 10-minute stations provide an adequate compromise, combining optimal reliability information with manageable resource-requirements.

Chapter 6

Effects of a Training in Interviewing and Counselling Skills[1]

Summary

During their study in the Mental Health Sciences students are systematically trained in interviewing skills, diagnostic skills and basic therapeutic skills. The purposes of the present study were to evaluate effects on students of a short training in interviewing skills and to compare the level of proficiency in applying these skills as acquired by students with the level of proficiency displayed by professional helpers who were not trained. By means of a pre-pre-post design the effects of training on skills, self-efficacy, and patient satisfaction were measured, using simulated patient encounters. Results showed a substantial improvement in interviewing skills and self-efficacy as a result of training. Furthermore, students attained levels comparable to those of professional helpers on skills-rating. In addition, no differences were found in simulated patient satisfaction between students and professionals. However, professionals judged themselves as to be interpersonally more effective than students.

[1] Based on: Bögels, S. M. & Kreutzkamp, R. (1990). Effecten van een training in basale gespreksvaardigheden. Tijdschrift voor Onderwijsresearch, 4, 200-214.

Introduction

A central aim of education in the health professions, such as medical education, psychology, nursing education, and health sciences, is to achieve an effective communication between a helper and a patient. The helper needs to acquire various interpersonal and communicative skills, in order to interact effectively with patients. Skills training programmes have been developed to teach these skills to students.

At the faculty of Health Sciences of the University of Limburg, an academic programme was established that provides undergraduate education in what is called Mental Health Sciences. A training curriculum has been developed that runs parallel to theoretical education. Students are systematically trained in basic interviewing and counselling skills, observation skills, behavioural assessment skills, diagnostic interviewing skills, and psychotherapeutic skills. In the training courses, innovative teaching methods are applied, such as simulated patient encounters and staff-independent exercise. Empirical evaluation of such a training curriculum is important for several reasons. First, the question needs to be answered whether students acquire the skills that the training aims at. Demonstrable effects are a precondition for the continuation of intensive training programmes like the one discussed here. Secondly, by assessing growth in skills students can be provided with precise feedback with regard to their skill level compared to other students and compared to pre-training performance. Systematic feedback is, from an educational point of view, essential for the acquisition of skills (Patrick, 1992; Riccardo & Kurtz, 1983); one does not learn from experience if it is not guided by information concerning the correctness of the performance.

Since the work of Ivey (1971) on micro-training, a large number of studies has been published on effects of interpersonal skills training for students and professionals. Marshall, Sharping, and Bell (1979) found in a meta-analysis of 122 studies on effects of skills-training for social workers, psychologists, psychiatric residents, and counsellors, that in only half of the studies significant effects were reported. In contrast, Carrol and Monroe (1980) concluded in a meta-analysis of 33 studies on interpersonal skills training for medical students, doctors, and nurses, that 80 percent of the training courses were effective. However, in most of the studies which were analyzed by Carrol and Monroe (1979) pretest-posttest designs without a control group were used. Such pre-experimental designs are sensitive to confounding factors, for example, the possible learning effect of testing itself.

In most of the evaluation studies, only statistically significant effects were used as an indicator of effectivity. But, statistical significance does not give much information about the strength and relevance of the observed differences for educational practice. One way to gain insight in the meaning of the skills

level that has been achieved as a result of training, is to compare the performance of trained students with that of a suitable reference group (i.e. their later professional group). Stone (1982) calls this "the social comparison method". Based on this comparison the question, what end-terms the student needs to achieve, can also be answered.

Although much research has been done to evaluate the effects of training courses and training procedures, the validity question, namely whether the trained skills are relevant to accomplish effective helper-patient communication, generally remained unanswered (Ford, 1979; Kurtz, Marshall, & Banspach, 1985; Stone, 1982). We know, for example, very little about the relation between helpers' skills and patient variables, like patient satisfaction with the helpers' communication.

Potentially, training may produce many changes; changes in trainees' behaviour, changes in trainees' cognitions, changes in the experience of the patient. The performance of the trainee in terms of observable behavioural changes, however, was the only dependent variable in most evaluation research. Although behavioural change is an important outcome measure, relying solely on behavioural changes involves a risk. The trainee may, as a consequence of knowing that only behavioural entities (and not the outcome of behaviour) is judged, concentrate on mechanical behaviour units ("tricks") without knowing what function this behaviour has. In this line, Van Luyk, Van der Vleuten, and Van Schelven (1990) demonstrated that as a result of the fact that medical students memorized behavioural checklists used in previous skills-tests, a deeper understanding of the skills seemed to be missing.

One of the potential effects of training may be an increase in the trainee's confidence in the effectivity of his interpersonal behaviour. Bandura (1977) assumed that the degree to which people are convinced of the effectivity of their behaviour (called "self-efficacy"), may be an important determinant of their behaviour, since self-efficacy can increase the intensity and persistence of behaviour. People's beliefs in their capabilities to fulfil task demands within a wide range of psychological domains, assessed by self-efficacy scales, have been shown to enhance or impair performance. Furthermore, it has been suggested that skills-training can produce more generalized and lasting effects if it raises self-beliefs of efficacy as well as imparts skills (Bandura, 1989). In sum, in order to validly assess training effects, it is important to consider a variety of aspects; the overt trainee behaviour, the cognitions of the trainee, and the experiences of the patient.

Taking into account the above mentioned considerations, a study was conducted to evaluate effects on students of Mental Health Sciences who followed a training in basic interviewing and counselling skills. The questions we aimed to answer were:

1. What are the effects of a training in basic interviewing and counselling skills on the trainees' behaviour, trainees' self-efficacy, and the satisfaction of the simulated patient with the communication?
2. Are the trained interviewing skills related to patient satisfaction and self-efficacy?
3. How skilled are trained students compared to professionals?

Method

Subjects

Of the 36 second-year students of Mental Health Sciences, 16 students (5 men, 11 women) voluntarily participated in the study. Their mean age was 22.4 years (range 19-45 years). One subject had prior experience in mental health care. Twelve professionals (7 men, 5 women), working in various mental health institutes, took part in the study. Their mean age was 29.5 years (range 23-37 years). Four were psychiatric residents, 6 were psychiatric nurses, and 2 were psychologists. Their mean working experience in mental health care was 4.5 years. Their work consisted of initial interviews, counselling, and psychotherapy.

Training

The training was organized for all (36) second-year students in four groups of 9 students each. The experimental Ss were randomly divided over the four groups. Trainers were faculty members, experienced in problem-based interviewing and training novices in these skills. The training consisted of five 3-hr sessions, and aimed at learning to conduct an initial interview in order to clarify the patient's problem(s). The complex skill of problem clarification was divided into the following (sub)skills (see Lang & Van der Molen, 1984): attending, open-ended questions, paraphrasing of content, reflection of feelings, summarizing, concreteness, and regulating skills. The training method was mainly based on the micro-counselling approach of Ivey (1971). The course started with a video-demonstration of the integrated and functional use of all subskills, in order to inform students about what skills they were expected to acquire during the training course. Thereafter, one or two skills were introduced in each session. The training format consisted of the following components. First, the skill under study was demonstrated by means of a video-example. Secondly, an introductory exercise was carried out in order to achieve a better understanding of the skill. Thirdly, the skill was practised in 5-10 minute role-plays. Role-playing was followed by immediate feedback by the trainer and students, focused on the

single skills being taught. Videotaped feedback was also provided. Finally, the integrated and functional use of all skills in a problem-based interview was practised (for a description of the training method see Lang & Van der Molen, 1992).

Design

A "pre-pre-post" or "time-series" design (Campbell & Cook, 1979) was used to measure changes in performance as a result of training. Education is organized in 6-weeks units, in which a theme is studied. The theme of the first unit of Mental Health Sciences (in which the training took place) was called "Introduction in mental health care and mental health problems". At the start of this unit students were tested for the first time (pretest 1). Three weeks later, just before the start of the training, the test was repeated (pretest 2), in order to control for the combined effects of theoretical education and testing itself. After the 3-week period, in which the training took place, Ss were tested for the third time (posttest). Since the theoretical education in the first 3 weeks did not differ essentially from the theoretical education in the second 3 weeks -in which students received training-, a greater change in performance in the expected direction between pretest 2 and posttest than between pretest 1 and pretest 2, could therefore be attributed to the effect of training.

Professionals were tested once, after the posttest of the students. Simulated patients were blind to whether the Ss were students or professionals.

Simulated patients and cases

Four simulated patients, all women, who regularly played patient roles for educational purposes, were instructed to simulate a case. Three cases were developed, miss A, B, and C. The cases were identical with regard to the reason for consulting the general practitioner (longlasting and frequent headaches), and had the same structure with regard to underlying problems that caused and/or maintained the headaches. However, the specific problems at work, in the marriage, and in the family, differed from one case to the other. In constructing the cases, special attention was paid to building in the same amount of interactional difficulty in each case, in order to make them comparable.

One simulated patient role-played miss A, one role-played miss B, and two role-played miss C. At pretest 1, half of the students interviewed miss A and half miss B. At pretest 2, the students who had interviewed miss A at pretest 1 now interviewed miss B, and vice versa. At posttest all students interviewed miss C. The professional also interviewed miss C.

Instruments

Interviewing skills

The interviewing and counselling skills were judged by means of a rating form. The rating form consisted of 10 items. Seven items referred to skills important in a problem-clarifying conversation, namely: attending, asking questions, paraphrasing of content, reflection of feelings, summarizing, concreteness, and regulating skills. Item 8 concerned the attitude of the interviewer. Item 9 reflected the regulation of the speed of the conversation. The last item assessed to what extent the problem was clarified by the interviewer. A combined judgement of the quality and functional use of each skill was given by raters on a 5-point scale from 1 (insufficient) to 5 (advanced). In addition, raters provided global judgements of the overall performance of the interviewer on a 10-point scale.

Patient satisfaction

In order to assess to what extent Ss had established a working alliance with the patient, a short patient satisfaction questionnaire was developed, inspired by the Patient Satisfaction with Communication Checklist (Crijnen & Kraan, 1987), originally developed to measure patient satisfaction related to physician's communication. Simulated patients indicated how strongly they agreed or disagreed with 7 statements, on a 6-point scale from 1 (totally disagree) to 6 (totally agree).

The statements were: 1) The interviewer showed that he understood me and my problems; 2) The interviewer gave me the time and scope to talk about what's bothering me; 3) The interviewer went further into the matters I raised during the conversation; 4) The interviewer paid attention to the different aspects of my problem; 5) The interviewer got to the root of the problem; 6) I was put at ease by the interviewer; 7) I would return to this interviewer.

Self-efficacy

Self-efficacy was measured by asking the Ss just before the interview to indicate on a 6-point scale how effectively they thought they would perform with regard to the most important interviewing and counselling skills to be used in a problem-based interview. A description was given of the following aspects: 1) attending, 2) listening, 3) asking further, 4) handling feelings, 5) summarizing, and 6) regulating skills. Finally, Ss were asked give an overall rating of their predicted capability to successfully conduct a problem-based interview. The 7 items were scored on a 6-point scale from 1 (not at all effective) to 6 (extremely effective).

Tension

Experiencing too much tension may affect interview performance during a simulated patient encounter. Especially since it was, at the first pretest, the first time that Ss were confronted with a simulated patient, tension could be expected to

distract them. To control for this factor Ss were asked to rate, after the interview, how tense they were during the interview (1=not at all tense, 6=extremely tense).

Procedure

After entering the behavioural laboratory, where the tests took place, Ss were asked to fill in the self-efficacy questionnaire. Next, Ss received some information concerning the patient they had to interview, in the form of a letter from the referring general practitioner. Ss were instructed to have a conversation with the patient lasting 20 minutes, in order to clarify the problem(s). The conversations were videotaped by means of two camera's. One camera videotaped the interviewer and patient together, the other made a 15 second close-up of the interviewer every five minutes, in order for the raters to be able to observe the nonverbal behaviour of the interviewer in more detail. After the conversation Ss were asked to indicate how tense they had been. Simulated patients completed the patient satisfaction questionnaire directly after each conversation.

Scoring

Of each videotape the 5th- to 15th minute were selected for further analysis. This was done because there was a great variety in the way Ss started the conversation and ended the conversation. The 5th to 15th minute gave more comparable material.

Four psychologists, two men and two women, were selected to rate the videotapes. They worked or had worked in mental health care and had experience in training students in interviewing and counselling skills. Therefore, they could be considered as expert judges. They were told that it concerned a study on the effects of skills-training and that students as well as professionals participated in the study. They were extensively trained in using the rating form. A total of 60 video-fragments (3 per student and 1 per professional) were randomly scored by the judges, each fragment was scored by all four judges. The rating was done in plenary 4-hrs sessions. Judges were blind with regard to whether the video-fragment concerned a pretest or a posttest, a student or a professional. To prevent judges from identifying professionals a priori as a result of age-characteristics, only relatively young or young-looking professionals took part in the study.

Results

Inter-rater reliability

Inter-rater reliability per item was calculated by means of a weighted Cohen's kappa. The weighted kappa is a reliability coefficient for ordinal and interval scales, which includes absolute differences between raters as error. A kappa of .40 is considered as moderate, a kappa of .60 as high/substantial (Pritchard, 1981). In Table 6.1 kappa's per item are presented. Inter-rater reliability per item was generally moderate to high. To calculate inter-rater reliability of the total scores (the 10 items plus the global overall rating) an intraclass correlation was used, which includes absolute differences between raters as measurement error. The inter-rater reliability of total scores was satisfactory; at pretest 1 .72, at pretest 2 .62, at posttest .71 and at the test of the professional .78. Since the total scores on the 10 items were strongly related to the global overall score (Pearson's correlations varying from .94 to .98), these scores were added. The skills score to be used for further analyses consists of the total score of the 10 items plus overall rating across the four raters.

The internal consistency of the patient satisfaction questionnaire was very high (Cronbach's α = .95), the internal consistency of the self-efficacy scale was satisfactory (α = .75).

Table 6.1. Inter-rater reliability per item across the four raters (weighted Cohen's kappa) for the students (N=16) at three measurements and for the professionals (N=12)

	Pretest 1	Pretest 2	Posttest	Test professionals
1. Attending	.49	.45	.44	.59
2. Asking questions	.49	.21	.35	.31
3. Paraphrasing of content	.42	.46	.43	.61
4. Reflection of feelings	.60	.52	.67	.56
5. Summarizing	.45	.32	.51	.65
6. Concreteness	.51	.33	.53	.67
7. Regulating skills	.41	.55	.53	.50
8. Timing	.42	.25	.61	.59
9. Attitude	.40	.51	.44	.65
10. Problem clarification	.59	.31	.45	.66

Effects of training

The effects of the training are presented in Table 6.2 and Figures 1, 2, and 3. Multivariate and univariate analyses of variance over the difference-scores of pretest 2 minus pretest 1 and posttest minus pretest 2 were conducted. No overall differences were found between pretest 1 and pretest 2 (F Hot.(3,13)=2.49, p=.106). Therefore, it can be assumed that differences between pretest 2 and

posttest are the result of training and not the result of testing, time, or other educational activities. A significant overall change was found between pretest 2 and posttest (F Hot.(3,13)=24.49, p=.000) in favour of the posttest. Univariate tests showed that the change was caused especially by the skills-rating (F (1,13) =27.29, p=.000) and to a lesser extent by self-efficacy (F (1,13)=-10.81, p=.005). No changes were found on patient satisfaction (F (1,13)=.46, p=.51). The growth in skills was almost two standard-deviations, which is quite large. Contrary to expectations, students indicated that they experienced the same amount of nervousness at the three tests, no differences were found between pretest 1 and 2 (F (1,15)=.24, p=.63) and between pretest 2 and posttest (F (1,15)=.81, p=.38).

Table 6.2. Means and standard deviations (between brackets) of all variables for the students at three measurements and for the professionals

	Students						Professionals	
	Pretest 1		Pretest 2		Posttest			
Skills-rating*	27.8	(6.7)	30.5	(6.0)	41.2	(7.0)	39.2	(10.3)
Patient Satisfaction	29.4	(9.4)	27.9	(7.3)	29.3	(5.2)	32.8	(6.8)
Self-efficacy	24.7	(3.0)	24.8	(4.2)	27.2	(2.4)	31.1	(1.5)
Tension	2.9	(1.0)	2.7	(0.7)	2.6	(0.6)	2.5	(0.8)

* averaged across raters

Validity of the skills-rating

In Table 6.3 the correlations of the skills-rating with simulated patient satisfaction and self-efficacy are given. A substantial relation was found between the skills, as rated by the judges, and the satisfaction of the patients with the communication of the interviewer. A low association was found between students' skills and their self-efficacy, and, moreover, a negative association was found between the professionals' skills and their self-efficacy.

Table 6.3. Pearson correlations of the skills-rating with simulated patient satisfaction and self-efficacy, analyzed for the three tests of the students (N=16) and the test of the professionals (N=12)

			Patient Satisfaction	Self-efficacy
Skills-rating	Students	Pretest 1	.48*	.19
		Pretest 2	.54*	.19
		Posttest	.81***	-.07
	Professionals		.54*	-.54*

* = $p < 0.05$
***= $p < 0.001$

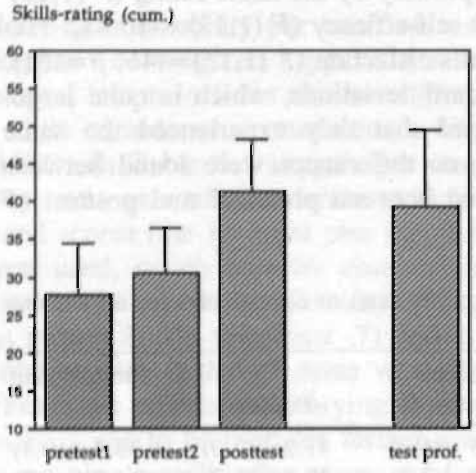

Figure 6.1. *Skills-rating of students* (at 3 measurements) and professionals*

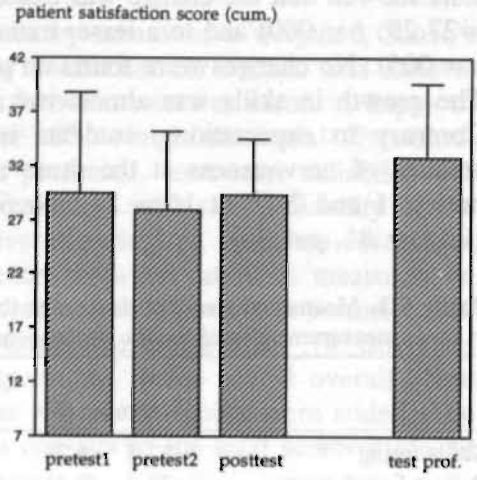

Figure 6.2. *Patients' satisfaction with students* (at 3 measurements) and professionals*

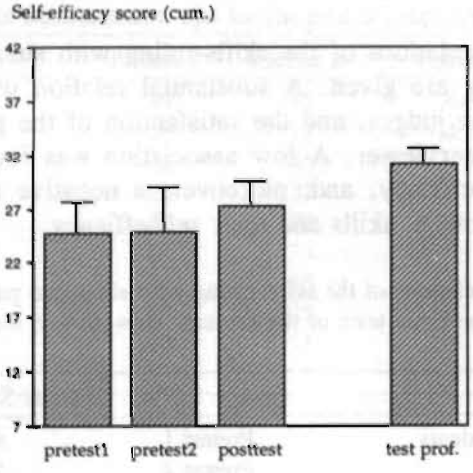

Figure 6.3. Self-efficacy of students (at 3 measurements) and professionals*

* Interview training took place between pretest 2 and posttest. No training was given between pretest 1 and 2, in order to control for the possible effects of testing, time, or theoretical education.

Validity of the training results

Trained students were compared with professional helpers (see Table 6.2 and Figures 6.1, 6.2, and 6.3). Multivariate analysis showed a significant overall difference (F Hot. (3,24)=8.40, p=.001) in favour of the professionals. Univariate analyses, however, revealed that the difference was caused solely by higher self-efficacy scores of the professionals (F (1,26)=25.21, p=.000) and not by higher skills-ratings (F (1,26)=.36, p=.55) or higher patient satisfaction (F (1,26)=2.4, p=.13). So, the professionals can be distinguished from the students only by their higher expectations of their performance.

Discussion

The first aim of this study was to determine the effects of a short systematic training in interviewing and counselling skills. The training was, at least in the short term, effective in establishing behavioural change in the desired direction. Also, a cognitive change was observed; trained students had higher expectations of their interpersonal effectiveness than untrained students. Contrary to expectations, the satisfaction of the simulated patients did not change as a result of training. A possible explanation can be found in a methodological shortcoming of this study. At posttest, students interviewed other simulated patients than at the two pretests. As a result, the simulated patients at posttest had a different baseline from that of the simulated patients at the pretests (namely, more skilled students). The simulated patients from the posttest, therefore, may have been more stringent in their judgement. The finding that patient satisfaction did correlate with skills-rating, supports this hypothesis. For, since higher quality skills coincided with more satisfied patients, it is reasonable to expect that substantial growth in skills would also result in higher satisfaction scores. No definite conclusions can be drawn yet with regard to the effects of skills-training on patient satisfaction.

The second question of this study concerned the validity of the skills being taught. The finding that judges' ratings of the skills of both students and professionals were strongly related to simulated patients' satisfaction with the communication, supports the validity of the skills.

Contrary to expectations, no relationship was found between students' efficacy expectations and observed skills, and even a negative relation between helpers' self-efficacy and observed skills. Although the small sample size of the professionals requires prudence with regard to conclusions concerning the negative relation, the low association between students' self-efficacy and skills requires an explanation. The low correlation is probably not the result of lack of knowledge about one's skills, since the self-efficacy of trained students, who

have had frequent interview experiences and received continuous feedback with regard to their skill-level, is still unrelated to observed skills. Possibly, a certain amount of insecureness (or an underestimation of one's self-efficacy) is related to empathic understanding, which is important in skilled performance. Or, to put it in another way, too high self-efficacy may result in a arrogant attitude which interferes with empathic understanding. It is also possible that interviewers who are insensitive to patients' communication are similarly insensitive to their own abilities. The relationship, claimed by Bandura (1977, 1989), between belief in one's efficacy and effective behaviour, could not be replicated in this study. Our results however correspond with those of Marteau, Humphry, Matoon, Kidd, Lloyd, and Horder (1991), who found no or a negative association between students' confidence at communicating effectively and their level of skill. In line with this, Kraan, Crijnen, Van der Vleuten, and Imbos (1992) found no or a negative correlation between students' self-ratings of interpersonal and communication skills after an interview with a simulated patient and judges' ratings of the same skills. The low association between self-efficacy (or self-rating of skills) and observed skills needs further research, especially since the inability to judge communication skills accurately may serve as a barrier to improving skills, as those with poorer skills may not have the motivation to learn from training.

The third question of the study concerned the validation of the training results: How does students' performance relate to professionals' performance? Trained students performed as well as professionals, and no significant difference in simulated patient satisfaction was found between trained students and professionals. Professionals, however, believed more in their own effectivity. These results support the meaning of the size of the training effect; training brought students to the same skill-level as professionals. However, the results are somewhat unexpected. No matter how effective the training, on the basis of the professionals' more advanced knowledge and experience, one still would expect professionals to perform better.

Several remarks can be made with regard to these unexpected findings. First, in the present study only basic interviewing and counselling skills were measured. Advanced counselling skills, like confrontation, were not part of the rating. Professionals are expected to perform better on such skills. Secondly, we did not assess interviewing performance in more difficult situations, such as interviewing silent, unwilling, or suspicious patients. The simulated patients in this study portrayed psychosocial problems and were not extremely difficult to interact with. Since expertise is characterized by a flexible or adaptive use of skills in a range of situations with changing task requirements (Patrick, 1992), one could hypothesize that in more difficult situations professionals would perform better than students. Thirdly, in the present study only short-term effects are assessed. It is unclear to what extent these results are maintained. Possibly, at long term students' skills would decrease compared to professionals' skills, since professionals practise their skills constantly. Fourthly, the rating scale used

to assess interviewing skills was based on the final terms of the training course. Although students were not familiar with the rating scale, they may have known better than the professionals what skills they were expected to demonstrate.

Several questions with regard to the effects of interpersonal skills training deserve attention. A first question concerns the long-term effects of training. In the present study, educational demands made it impossible to assess whether the training effects are maintained in the long run; directly after this training students followed other training courses, such as diagnostic interviewing, which could be expected to also affect basic interviewing skills. It is important that future research on skills-training effects addresses this question, since there is evidence that interpersonal skills-training suffers from loss of effect over time (Baker & Daniels, 1989; Sanson-Fisher & Maguire, 1980). Another question of interest concerns the effect of interviewing skills training on other outcome variables; does training result in better analyses of the patient's problem? Finally, there is hardly any research yet on the effects of training in higher-order skills, like confrontation, interpretation, or diagnostic interviewing. There is evidence that such complex skills need more training to develop (Patrick, 1992), and perhaps training methods that rely more on cognitive conceptualizations of the problem than the micro-counselling method (Robinson & Halliday, 1987). Evaluation research on the possible effects of training in higher-order skills could contribute to the present knowledge on how to teach them to novices.

Chapter 7

Training Diagnostic Interviewing to Students of Mental Health Sciences: Immediate and Long-term Effects[1]

Summary

This study evaluates the immediate and long-term effects of systematic training in psychiatric interviewing. Students of Mental Health Sciences (n=14) were first trained in diagnostic interviewing with regard to affective disorders. Six months later, they received training focusing on psychotic disorders. By means of a quasi-experimental design effects of instruction were examined. Four dependent variables were measured: process interviewing skills, content-related interviewing skills, recall of relevant patient information, and diagnostic accuracy, using simulated patient encounters. After the first training, students' performance substantially improved with regard to content-related interviewing skills and recall of relevant patient information. Performance however decreased after the 6 months during which students received no training. The second training did not improve performance. It is concluded that diagnostic interviewing can be effectively trained in a short period with regard to affective disorders, but teaching students how to conduct a diagnostic interview with psychotic patients may need more practice.

[1] Submitted for publication by Bögels, S. M.

Introduction

Diagnostic interviewing is an important professional skill for mental health professionals such as clinical psychologists, psychiatrists and counsellors. The diagnostic interview constitutes the main tool for assessment and, hence, for treatment of mental disorders. Although training programmes are being developed to teach these skills to future mental health professionals, controlled studies on the effects of these training programmes are lacking. Training effect studies on three other types of skills, related to diagnostic interviewing in mental health care, may however provide a relevant background for the present study.

First, a good number of empirical inquiries have been conducted into the effectiveness of counselling skills training. Baker and Daniels (1989) conducted a meta-analysis on 81 experimental studies of the effects of microcounselling. They concluded that microcounselling is an effective instructional approach, although selective reporting has led to the impression that the effects are greater than they really are, and training suffers from decrease in effects over time. Studies on the effects of training in higher-order skills are scarce.

Secondly, substantial research has been devoted to effects of training in medical interviewing, providing some evidence of short-term benefits of training (Carrol & Monroe, 1980; Evans, Stanley, Burrows, & Sweet, 1989). It is, however, unclear what happens with the acquired skills in the long term. Several investigators noted that after some initial improvement in interviewing skills resulting from training programmes, a decrease in skills occurred (Engler, Saltman, Walker, & Wolf, 1981; Kraan et al., 1990; Sanson-Fisher & Maguire, 1980). Studies on changes in medical students' interaction style have shown a shift in approach that offers an explanation for this decrease in communication skills (Evans, Stanley, Mestrovic, & Rose, 1991; Helfer, 1970; Scott, Donnelly, & Hess, 1975). In their early clinical days, students display an interest in talking with patients and demonstrate skills in facilitating communication, but at the same time lack skills in medical history-taking and deriving a diagnosis (Preven, Krajic Kachur, Kupfer, & Waters, 1986). However, as students' medical knowledge and skills develop during their clinical training, they tend to focus on symptoms and functional inquiry and pay less attention to maintaining an effective communication with the patient.

Thirdly, some studies have evaluated effects of interview training on the psychiatric skills of doctors. Psychiatric interview training was effective in improving family doctors' recognition of psychiatric disturbances (Goldberg, Steele, Smith, & Spivey, 1980) and their interviewing skills (Gask, Goldberg, Lesser, & Millar, 1988; Gask, McGrath, Goldberg, & Millar, 1987). The effects of a psychiatric clerkship for medical students were evaluated by Mumford, Schlesinger, Guerdon, and Scully (1987). They distinguished two types of interviewing skills; process skills, reflecting the interpersonal qualities of the inter-

viewer, and content skills, referring to the interviewer's information-gathering ability. A modest improvement was found on process scores, no improvement was found on content scores.

An important finding, consistent in almost all research on medical competence, is the low correlation between a doctor's performance on different cases. This has been termed the *"content-specificity of clinical competence"* (Elstein, Shulman, & Sprafka, 1978). Research on the assessment of diagnostic interviewing performance of students in Mental Health Sciences also revealed that performance was highly content-specific (Bögels et al., 1994), suggesting that skills, acquired on one type of mental disorder, may not automatically generalize to other types of disorders.

To summarize, some positive effects can be achieved by training in various interpersonal skills, such as counselling, medical interviewing, and psychiatric interviewing for doctors, but acquired skills tend to deteriorate over time. The natural course of interviewing behaviour as observed in medical students (shifting from interpersonal interest to diagnostic issues) may contribute to this deterioration. Furthermore, skills that are acquired on one type of case, may not generalize to other types of cases. Research on the effects of diagnostic interview training for (future) mental health professionals is lacking.

The purpose of this study was to evaluate, in a controlled way, effects on students of Mental Health Sciences of training in diagnostic interviewing (mostly termed "psychiatric interviewing"). We intended to answer the following questions:

1. What do advanced undergraduate students learn from training in diagnostic interviewing?
2. Is improvement by training "content specific", or do skills generalize to types of disorders not dealt with during a course?
3. Are the effects of training maintained over time?

Methods

Training

A training programme in diagnostic interviewing has been developed for students of Mental Health Sciences. The training courses take place in the context of a problem-based academic curriculum (Barrows, 1980). Theoretical instruction is organized in educational units of 6 weeks, in which integrated, multidisciplinary teaching takes place by means of a specific theme. Skill-training is linked to the theme of the unit, in order to integrate theory and practice.

The first training students attend (at the beginning of the second year) is a short training in process interviewing skills, based on the microcounselling

approach of Ivey (1971). A study, conducted to evaluate the effects of this training, revealed that the training was highly effective in improving interview performance with respect to process skills (Bögels & Kreutzkamp, 1990). Next, students are trained in behavioural observation and behavioural assessment interviewing.

Halfway through the second year, a first training course in diagnostic interviewing is scheduled, during an educational unit in which "mood" is the central theme. The training focuses on affective disorders, and aims at learning how to conduct a diagnostic interview and to produce an accurate diagnosis, formulate explanatory hypotheses, and make a plan for further management, based on data gathered during the interview. Following the didactical principles of microcounselling (Ivey, 1971), the complex skill of diagnostic interviewing is divided into specific skills, that are trained separately by means of video-modelling, role-playing, and feedback. Role-playing takes place in triads, with one of the students playing a patient, another being the helper, and the third being the observer. Trainees impersonating a patient are provided with written scripts of cases with affective disorders. Diagnostic accuracy is also trained by showing videotaped vignettes of patients with mental disorders. Students formulate and motivate their diagnostic hypotheses, using the DSM-III-R (American Psychiatric Association, 1987). The training consists of six 3-hr group meetings.

An advanced training in diagnostic interviewing is scheduled at the beginning of the third academic year, during an educational unit about psychoses. It aims at learning to conduct a diagnostic interview with patients having psychotic disorders. High-fidelity role-playing is organized with the help of simulated patients, who are psychiatric nurses, specifically trained to portray psychotic cases. A simulated patient is introduced in each session and four students are given the opportunity to conduct part of the interview. Other students are required to observe and to provide feedback afterwards. The size of the training is six 3-hr group meetings.

The two training courses in diagnostic interviewing were the subject of the present study.

Subjects

Training was organized in groups of 8-10 students, each group was guided by a faculty member. From the entire group of 45 second-year students of Mental Health Sciences, a sample of 16 students participated in the study. They were randomly divided over the training groups. The results of 2 Ss were deleted because of missing data. Fourteen Ss remained, 2 men and 12 women, mean age at the beginning of this study 21.6 year (range 19 to 28 years). One of them had prior experience working in mental health care. The Ss were representative of the entire sample with regard to sex, age, and prior working experience (6 of

them are men, their mean age is 22.3 with a range form 19 to 40, 3 of them had already worked in a mental health setting).

Design

A quasi-experimental design was used to measure progress in diagnostic interviewing as a result of training (see Figure 7.1). Three weeks before the first training in diagnostic interviewing Ss were tested for the first time (Test 1). After 3 weeks of theoretical study of affective disorders, but without training in diagnostic interviewing of patients with affective disorders, this test was repeated (Test 2), in order to control for the effects of theoretical education without training and for the effect of testing itself (Cook & Campbell, 1979). In this way Ss were their own control. After the 3 weeks in which students were trained in interviewing patients with affective disorders, they were tested again (Test 3). Six months later, Ss were tested for the fourth time (Test 4). During the 6 months prior to Test 4 Ss received neither training in diagnostic interviewing nor theoretical education with regard to affective or psychotic disorders. Therefore, at Test 4 we would be able to assess to what extent diagnostic interviewing skills were maintained. During the 6-week period between Test 4 and the final test (Test 5), students followed theoretical education with regard to psychotic disorders as well as training in interviewing psychotic patients.

Cases

Twenty standardized case histories were developed, derived from real patients. Four cases were used at each test, two affective disorders, and two psychotic disorders. The cases were enacted by 20 simulated patients, each patient portraying one case. Simulated patients were selected from our simulated patient file, which contains psychiatric nurses, students from the academy of dramatic arts, and other people with considerable experience in role-playing psychiatric cases for educational purposes.

Instruments

Changes in interviewing behaviour were measured by means of the Diagnostic Interviewing Rating Scale in Mental Health Care (DIRSM, Bögels et al., 1994), a rating form to assess the quality of psychiatric interviewing performance. The DIRSM consists of two scales called "process skills" and "content skills". Process scores reflect ratings on 10 aspects of interpersonal and communication skills, such as asking questions, reflection of feelings, summarizing, thinking aloud and giving information. Content scores reflect ratings on 13 aspects of information-gathering ability, including clarification of the presenting problem, history of the main problem, psychiatric examination, personal, social and fam-

ily history. Each aspect is scored on a 5-point scale from 1 (insufficient) to 5 (advanced). Finally, raters are asked to provide global judgements of process and content quality of the interviewer, both on a 10-point scale.

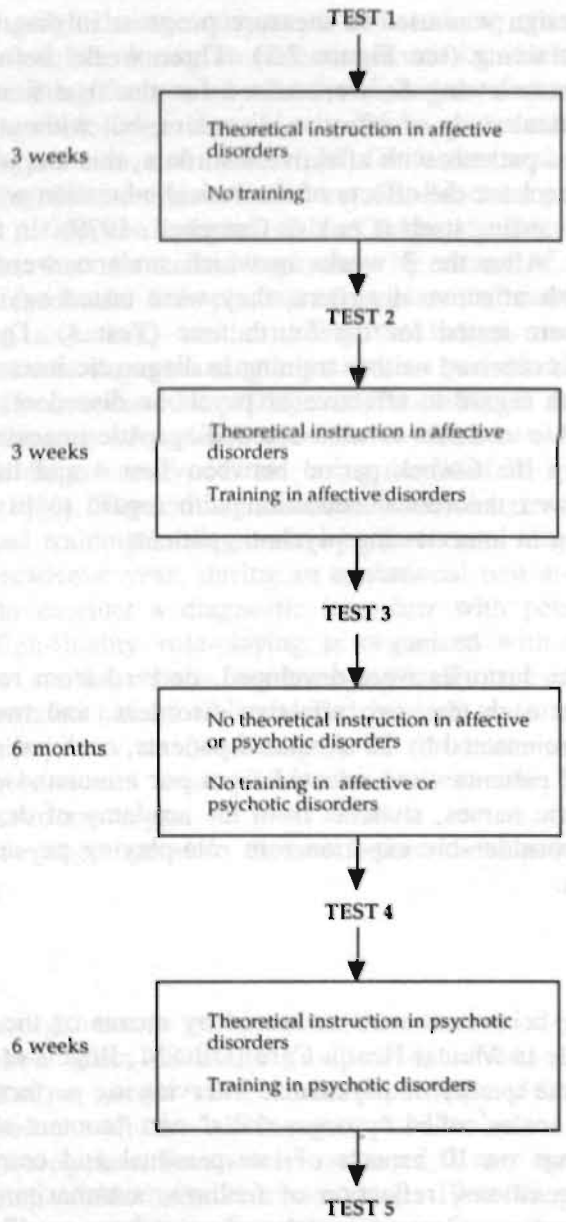

Figure 7.1. Schematical design of the study

In order to measure how much information, relevant for an adequate comprehension of the patient-case, was obtained, Ss were instructed to "recall everything they remembered about the patient they had just interviewed". This is a well known procedure to chart the mental representation of a case by the diagnostician (Schmidt & Boshuizen, 1993). These recalls were audiotaped. For each case a scoring list was developed to measure the amount of relevant patient information recalled. Each list contained 30-40 items of relevant patient information concerning diagnosis, differential diagnosis, etiologic factors, predisposing factors, and help request/previous treatments. Each item was scored on a 2-point scale; 0=not mentioned, 1=mentioned by the interviewer.

Diagnostic accuracy was measured by means of the Problem Solving Questionnaire (Kraan & Crijnen, 1987). The interviewer responded to 7 open-ended questions concerning diagnosis, explanatory hypotheses and plans for treatment. In order to rate the answers of the Ss on the Problem Solving Questionnaires, 5 psychiatric experts developed a scoring key for each case, consisting of accurate (2 points), semi-accurate (1 point) and wrong (0 points) answers.

Research on the measurement of diagnostic interviewing (Bögels et al., 1994) revealed that the dependent measures are sensitive to absolute differences in case difficulty, i.e. test difficulty. Although the cases used in the present study at each measurement point were randomly selected, the possibility that results must be attributed to the specific sample of cases rather than the effect of training cannot be excluded *a priori*. Differences in case difficulty were therefore measured in two ways. First, Ss were asked to rate on a 5-point scale their impression of the difficulty of each simulated patient interview (from 1=not difficult to 5=very difficult). Next, the Amsterdam Clinical Challenge Scale (Gercama, 1992) was used to establish the difficulty of the cases in an objective way. This scale, designed specifically to assess the difficulty of family practice cases, was adapted for use in the mental health context. The modified version measures the complexity of the case with regard to the following aspects: problem, psychiatric history, actual context, personal history, communication, and management. All items are scored on a 5-point scale, from 1 (easy) to 5 (difficult).

Procedure

During each test, Ss were instructed to conduct a diagnostic interview with the four simulated patients, each for a period of 30 minutes. The interviews were videotaped. Immediately after each interview Ss were asked to recall (on audiotape) all information remembered about the patient just interviewed. Between the interview and the recall Ss were given a short counting task, in order to control for recency effects (Peterson & Peterson, 1959). Finally, Ss completed the PSQ concerning the patient they had just seen. They were allowed to use the DSM-III-R to answer the questions about diagnosis.

Scoring

Four students of Mental Health Sciences who had nearly completed their study, rated the interview performance of the Ss by means of the DIRSM. Raters were extensively trained, to a satisfactory inter-rater agreement. The inter-rater reliability, based on the ratings of all four raters of 10 Ss with 10 different simulated patients, was .78 for process skills, and .86 for content skills (measured by an intraclass correlation, including absolute differences between raters as measurement error). To ensure that observers did not drift away from the original standards, a short retraining was provided twice during the 2-month period in which they rated all videotapes. The videotaped interviews were scored by only one rater. Raters were nested within patients; rater 1 scored the interviews of all Ss with case A, rater 2 scored the interviews of all Ss with case B, and so on. All raters judged the interviews with one case at each test occasion. Videotapes were rated in a select order, and raters were blinded with regard to the moment of measurement. Audiotaped recall of relevant patient information of all Ss at all tests was rated by one of the raters. The completed Problem Solving Questionnaires of the Ss were scored by another rater, based on the scoring key. Two clinical psychologists were asked to rate the complexity of each case by means of the Amsterdam Clinical Challenge Scale. Of each patient-case a videotape was selected from an interviewer who had received a psychiatric interviewing score comparable to the total group mean. In this way, raters were the least influenced by the performance of the interviewer. Both raters rated all 20 cases. The inter-rater reliability was .87 (intraclass correlation).

Statistical Analyses

Total sumscores and subscale sumscores across the four cases (each subject interviewed four patients at each test occasion), were calculated for all variables. In addition, scores were differentiated to affective and psychotic disorders. Because recall had different maximum scores in different cases, a proportional score was calculated for this variable.

To minimize the risk of chance findings, only total scores of the dependent variables (process and content skills of the DIRSM, diagnostic accuracy and recall), averaged across the four cases, were used to test the effect of training. Results with regard to subscales of the dependent variables, as well as results with regard to the type of disorder (affective or psychotic) were used for descriptive purposes. Repeated measures analyses of variance were carried out to test the overall effect of training. Differential training effects were tested with MANOVAs of difference-scores between two tests.

Results

Case Difficulty

In Table 7.1 mean scores of all variables are presented.

Table 7.1. Means and standard deviations between brackets of all variables (total scores and subscale scores), averaged across cases, at the five moments of measurement ($N=14$)

Variable	1	2	3	4	5
Process skills (total)	46.5	45.1	45.4	44.8	45.9
	(4.1)	(3.2)	(4.1)	(4.8)	(4.2)
Content skills (total)	34.8	32.3	39.7	36.1	37.6
	(4.5)	(4.6)	(4.6)	(4.4)	(5.1)
Content skills (diagnosis)	19.2	17.9	21.6	19.4	21.3
	(2.6)	(3.0)	(2.5)	(2.3)	(2.6)
Content skills (etiology)	10.7	10.4	12.6	11.9	11.3
	(1.3)	(1.1)	(1.9)	(1.8)	(1.6)
Diagnostic accuracy (total)	5.0	5.3	5.9	6.0	5.8
	(1.8)	(1.6)	(1.9)	(1.6)	(1.6)
Diagnostic accuracy (diagnosis)	1.3	1.5	1.8	1.7	1.6
	(0.8)	(0.5)	(0.6)	(0.6)	(0.4)
Diagnostic accuracy (etiology)	2.8	2.7	2.8	2.7	2.3
	(0.7)	(0.8)	(0.9)	(0.6)	(0.9)
Diagnostic accuracy (treatment)	0.9	1.1	1.3	1.6	1.9
	(0.7)	(0.7)	(0.8)	(0.9)	(0.8)
Recall (%) (total)	27	23	35	25	27
	(4.8)	(6.4)	(6.1)	(5.3)	(6.1)
Recall (%) (diagnosis)	26	27	36	34	36
	(8.7)	(8.9)	(10.1)	(10.3)	(8.4)
Recall (%) (d.d.)	0	5	14	4	12
	(0.0)	(4.1)	(13.1)	(5.6)	(9.3)
Recall (%) (etiology)	43	28	45	38	29
	(7.5)	(7.6)	(8.6)	(10.1)	(9.9)
Recall (%) (predisposition)	9	12	13	14	8
	(10.0)	(8.0)	(14.4)	(10.2)	(7.3)
Recall (%) (treatment)	26	21	41	9	23
	(10.0)	(18.0)	(10.9)	(7.6)	(10.6)
Case difficulty (subjective)	3.0	3.0	3.1	3.0	2.7
	(0.4)	(0.6)	(0.6)	(0.7)	(0.8)
Case difficulty (objective)	17.8	17.9	18.3	19.6	18.5
	(4.5)	(6.4)	(5.5)	(3.2)	(2.7)

Ss did not report any subjective differences in degree of difficulty between cases at different tests (F (4,32)=1.1, p=.389). No differences were found either in objective case difficulty between the five tests (F (4,15)=.1, p=.980). Furthermore, there were no mean differences in case difficulty between affective and

psychotic disorders (F (1,18)=.62, p=.443). The objective case difficulty score correlated, as expected, negatively with performance, namely Pearson's r= -.35 with content skills, -.38 with problem solving, -.32 with recall. No correlation was found between case difficulty and process skills; r= -.03. These results indicate that case difficulty indeed influences performance. In the present study however, differences in performance between tests cannot be attributed to differences in case difficulty, as test occasions were comparable in mean case difficulty.

Effects of Training

Results of the analyses of variance are given in Table 7.2. In Figures 7.2-7.5 the results on each of the outcome variables are presented graphically.

Table 7.2. Repeated measures analyses and multivariate and univariate analyses to test for the effects of training versus no training

Variable	trial (repeated measures)	Test 2-Test 1 (no training)	Test 3-Test 2 (training I)	Test 4-Test 3 (follow-up)	Test 5-Test 4 (training II)
Overall		F(4,10)=2.9	F(4,10)=18.5•••	F(4,10)=11.9••	F(4,10)=1.2
Process skills	F(4,52)=1.1		F(1,13)=0.3	F(1,13)=0.3	
Content skills	F(4,52)=13.1•••		F(1,13)=51.0•••	F(1,13)=11.8••	
Diagnostic accuracy	F(4,52)=2.52		F(1,13)=2.9	F(1,13)=0.0	
Recall	F(4,52)=20.2•••		F(1,13)=49.2•••	F(1,13)=49.2•••	

•• = $p < 0.1$
••• = $p < 0.001$

Repeated measures analyses of variance showed a highly significant overall effect of time on content skills and recall, a nearly significant overall effect of time on diagnostic accuracy (p=0.052), and no effect of time on process skills.

Four MANOVAs were carried out to test for the effects of training versus no training, each for two adjacent measurement points. Ss were tested and retested 3 weeks later to control for the combined effects of theoretical education only and testing. No significant overall changes were observed between the first and second pretest, but there was a trend towards deterioration in performance.

From these data it can be concluded that neither the experience of being tested (conducting four diagnostic interviews) nor theoretical instruction without skills-training in itself resulted in better diagnostic interview performance. Therefore, differences between Test 2 and Test 3 can reasonably be attributed to the effect of training, or to the interaction between training and theoretical instruction. In accordance with expectations, a significant overall improvement was found after the first training. Univariate analyses indicated that improvement was made on content skills and recall. No significant improvement was found with regard to diagnostic accuracy and process skills. At the 6-month follow-up, Ss had deteriorated significantly. This relapse was found with regard to content skills and recall (univariate analyses). The second training did not produce significant progress.

As is shown in Figures 7.2-7.5, the first training, in which the case material consisted mainly of mood disorders, improved performance with regard to affective disorders, but performance hardly generalized to psychotic disorders. The second training, interviewing psychotic patients, did not have any specific effects on the outcome variables, neither with regard to affective, nor with regard to psychotic disorders.

Next, results with regard to subscales are analyzed (see Table 7.1). Content interviewing can be distinguished in the subscales 1) interview behaviour, necessary for gathering diagnostic information, and 2) interview behaviour, necessary for the collection of etiologic and context information. Most progress was made with regard to diagnostic information-gathering ability, less progress was made with regard to etiologic interviewing performance. In fact, etiologic interviewing deteriorated after the training course in interviewing psychotic patients. Diagnostic accuracy consisted of the subscales 1) quality of the (differential) diagnosis, 2) quality of etiologic hypotheses, and 3) quality of treatment plans. Substantial progress was made on quality of treatment plans up to Test 5. Only little progress was made on quality of the (differential) diagnosis, and no progress whatsoever on etiologic hypotheses. Recall showed great fluctuations on subscale level. Comparing pre-training levels (Test 1 and 2) to the final level (Test 5), recall of diagnostic and differential-diagnostic information had increased substantially, whereas recall of etiologic information showed no progress, and, in fact, deteriorated after the training in interviewing psychotic patients.

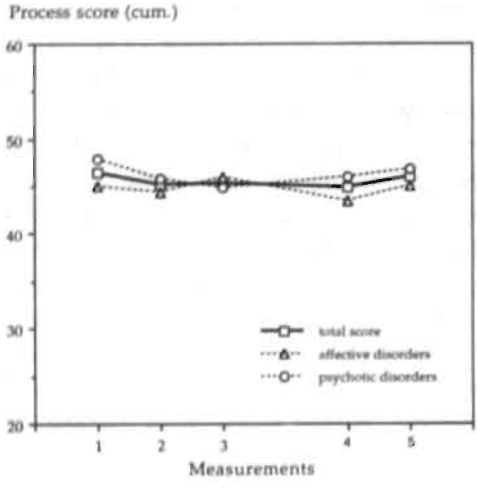

Figure 7.2. Effect of training on process interviewing skills*

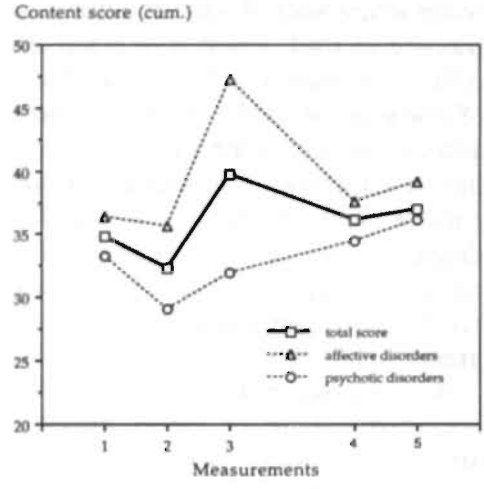

Figure 7.3. Effect of training on content interviewing skills*

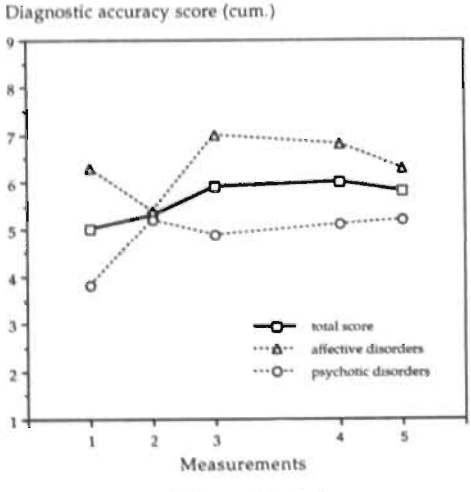

Figure 7.4. Effect of training on diagnostic accuracy*

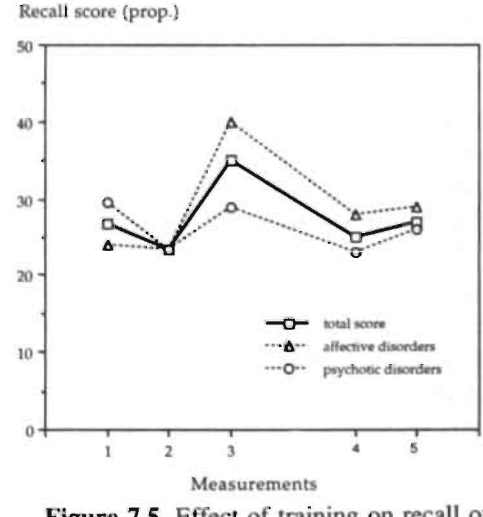

Figure 7.5. Effect of training on recall of relevant patient information*

* Diagnostic interview training with regard to affective disorders took place between measurement 2 and 3, diagnostic interview training with regard to psychotic disorders took place between measurement 4 and 5. No training was given between measurement 1 and 2, to control for the effects of testing, time, or theoretical education. At measurement 4, the retention of skills after 6 months without training is assessed.

Discussion

Effects of Training

The purpose of this study was to evaluate the cumulative effects of two training courses in psychiatric interviewing by means of a quasi-experimental design. The first training in diagnostic interviewing was effective with regard to the acquisition of content skills. Ss did also reproduce more relevant patient information. Some improvement was made with regard to diagnostic accuracy, although this trend was not significant. The effects of training were specific for affective disorders and hardly generalized to psychotic disorders, so the acquisition of diagnostic interviewing skill appears indeed to be "content-specific". The small improvement on diagnostic accuracy after training is puzzling, since producing accurate diagnoses and management plans was one of the objectives of the training. It seems that the training time, devoted to these skills, in addition to theoretical instruction and the reading materials provided, was not enough to increase the knowledge base required for diagnostic reasoning. Research on medical expertise has revealed that accurate diagnosis is highly dependent on domain-specific knowledge and experience and needs much exposure to develop (Schmidt, Norman, & Boshuizen, 1990). Therefore, more instruction and more exposure to a variety of clinical problems may be needed to improve the diagnostic accuracy of novices.

Performance relapsed significantly during the 6 months in which no further training in interviewing skills took place. During this period Ss had no opportunity (at least not organized by university) to apply or retrain their skills. The relapse found in this study is in line with other research showing that training suffers from the loss of effect over time (Baker & Daniels, 1989; Engler et al., 1981).

Contrary to expectation, the training in diagnostic interviewing of psychotic patients (combined with theoretical education on psychotic disorders) did not lead to significant improvement. A tentative explanation for this lack of effect can be found in the great variety of symptoms that psychotic patients may exhibit, and the generally more complex differential diagnosis of psychotic symptoms. Furthermore, psychotic patients often make great demands on the interpersonal qualities of interviewers, as they may, for example, mistrust the interviewer or be difficult to keep in reality. A training of six sessions, therefore, may usually be sufficient to teach the diagnostic interview with affective disorders, but too short to acquire the competencies necessary for the interviewing of psychotic patients. The method of training could be a second tentative explanation for the lack of effect of the advanced training. The introductory training in diagnostic interviewing was developed following the microcounselling method, with frequent video-modelling, behaviour rehearsal and feedback, whereas the

advanced training consisted mainly of plenary sessions with a simulated patient, providing less opportunities for each student to practise the skills. There is evidence that the mastery of skills is a direct result of the amount of standardized modelling (Carrol & Monroe, 1980; Fyffe & Oei, 1979) and practice (O'Toole, 1979). A third explanation can be found in the phenomenon that learning is initially rapid, but more and more practice is required to achieve the same improvement later in training (Patrick, 1992). The students of this study were in their third year of study and had already received substantial training at the time of the advanced training in diagnostic interviewing.

The finding that students' *process* skills did not improve by training in diagnostic interviewing skills is not surprising since its level was already high to begin with. Prior to the training in diagnostic interviewing, they received training in process interviewing skills. Research (Bögels & Kreutzkamp, 1990) showed a great improvement on process skills after that training. In fact, the post-training level students reached in that study, is similar to the pre-training level found in this study. Taking into account the tendency of medical students' process skills to deteriorate during the curriculum as a result of increasing attention to diagnostic issues (Diseker & Michielutte, 1981; Scott, Donnelly, & Hess, 1975), one may even say that the fact that students' process skills in this study did not deteriorate despite increasing attention to diagnostic concerns, might be a result of the continuing training.

Inspection on subscale-level gave a more differentiated insight into what students learned from their training. Progress was made with respect to the gathering and recall of (differential) diagnostic information and treatment-relevant information. The quality of diagnosis and management plan also improved after training. No progress however was made in the long term with respect to the recall of etiologic information and the quality of etiologic formulation. Commonsense ideas, based on everyday insights into human nature, which are not empirically tested, are nearly always an important part of etiologic formulation in psychiatry (Gelder, Gath, & Mayou, 1988). Contrary to the state of art of etiological explanations, there is a firm body of knowledge on diagnostic classification. Knowledge of treatments, effective for specific disorders, is also developing rapidly. Consequently, instruction in diagnosis and treatment planning may have been more clear and may have received more attention during training than the instruction in etiologic formulation. The lack of effect with regard to etiologic issues may therefore be the result of both the nature of instruction and state of the art with respect to etiologic comprehension of mental disorders.

A final issue concerns the research design of the present study. Since educational demands made it difficult to include a control group, a quasi-experimental design was used. By including two pretests we controlled for the competing effects of time, other education than training, and testing (Cook & Campbell, 1979). Furthermore, differences in difficulty of patient cases were controlled. Results indicated that neither exposure to diagnostic interviewing of simulated

patients (without feedback) nor theoretical education without training produced progress in performance. On the contrary, there was a trend towards deterioration of performance. A possible explanation for this unexpected finding is that Ss might be less motivated at the second test than at the first test, because the test situation was not new anymore, and therefore less challenging. Thus, a limitation of the quasi-experimental design applied in this study, is that it cannot be precluded that other factors than training have also influenced the results, for example a loss of motivation for the test over time.

Directions for Educational Improvements and Further Research

The results of this study suggest several avenues for improving instruction in diagnostic interviewing skills. First, the data suggest that training in higher-order skills such as diagnostic interviewing can be developed following learning principles of modelling and practice of separate skills. Training formats consisting of plenary sessions in which a few trainees practise, and the other participants observe, may be less effective than formats in which each student practises frequently. Secondly, interview training with difficult cases such as violent or paranoid patients has to be sufficiently extensive if the purpose is not only introduction but also mastery. Thirdly, since diagnostic skills seem to relapse without training or application possibilities, it is advisable to start a new training that builds on earlier learned skills with a short retraining of the previously acquired skills.

A number of issues with regard to research on training in diagnostic interviewing need further clarification. There is an urgent need for empirical evaluation of training programmes in higher-order skills such as the ones discussed in this article. Without such studies, the value of the training programmes in use remains questionable. Furthermore, it is important to examine the natural course of process- and content-related interviewing skills of trainees in mental health professions, and compare the results with medical students. The "natural" shift in medical students' interviewing behaviour towards diagnostic issues may not be found with students in mental health professions to the same extent, which has important implications for the training needs of these students. Finally, research is needed to compare the effects of training in the context of the curriculum with the effects of clinical clerkships, and the relative costs of the two approaches to skills acquisition.

Chapter 8

Diagnostic Expertise in the Mental Health Professions: Differences Between Novices and Experts[1]

Summary

In this study, the nature of diagnostic expertise in psychiatry is examined. Novices (15 third-year students of Mental Health Sciences, trained in psychiatric interviewing) and experts (12 psychiatrists and clinical psychologists) each conducted diagnostic interviews with four simulated patients who portrayed different mental health disorders. It was found that experts performed better than novices with regard to process interviewing skills, content-related interviewing skills, diagnostic accuracy and recall of relevant patient information. Only on a measure of patient satisfaction, no differences were found between novices and experts. The results support the validity of the measures proposed to assess growth in diagnostic skills.

[1] Submitted for publication by Bögels, S. M., Van Mourik, T. G. C., & Schmidt, H. G.

Introduction

Diagnostic interviewing skill, the ability to arrive at an accurate diagnosis and adequate management planning of patients with mental disorders are important professional competencies for mental health professionals. Gaining expertise in these skills should be one of the main objectives of education in the mental health field (Langsley & Hollender, 1982; McCready & Waring, 1986; Shea & Mezzich, 1988; Sullivan, 1954). In spite of the recognition of the importance of these skills, almost no empirically-based knowledge exists on the nature of diagnostic (interviewing) expertise in psychiatry. This state of affairs contrasts sharply with the extensive body of knowledge available on what constitutes diagnostic expertise of physicians. In this area, research has been conducted since the early 1960's (Rimoldi, 1961), resulting in a fairly detailed understanding of what goes on during the medical interview and how expertise in this field develops. This understanding is based on two lines of investigation: (1) Cognitive psychology research on medical expertise (Elstein, Shulman, & Sprafka, 1978; Norman, Brooks, & Allen, 1989; Patel & Groen, 1986; Schmidt & Boshuizen, 1993) and (2) research using performance-based tests (Rethans, Drop, & Van der Vleuten, 1991b; Stillman et al., 1991; Swanson, 1992). Both research traditions have utilized the novice-expert paradigm. In this paradigm, performance of subjects of presumably different levels of expertise is compared and inferences are drawn with regard to the nature of the competencies underlying differences in performance. In the present study, this (quasi-) experimental approach has been applied to the study of expertise development in psychiatric diagnosis. Before we turn to a discussion of the design and outcomes of this study, we will briefly review findings in the medical expertise domain to provide a context against which our findings can be judged.

Numerous studies have been conducted in the medical domain concentrating on differences in the way subjects of different levels of expertise process and cognitively represent patient information. In these studies, various measures are used, including response times, think-aloud measures, diagnostic accuracy, mental representation and recall of clinical cases (for a review of the important findings in this area see Schmidt, Norman, and Boshuizen, 1990). A consistent finding is that medical expertise is highly case-specific, that is: a clinician's actual experience with a particular problem appears to have the greatest influence on clinical reasoning and problem-solving skills (Elstein, Shulman, & Sprafka, 1978). Experts, furthermore, diagnose by means of rapid pattern recognition. They diagnose a new patient on the basis of similarity to previously encountered patients, and reason about a patient's signs and symptoms, applying knowledge in encapsulated mode (Schmidt & Boshuizen, 1992). Therefore, ordinary testing methods, like general knowledge tests, general reasoning tests, and tests that reward mainly comprehensive history-taking or memory for all

patient information, may put experts at a disadvantage and do not assess their domain-specific, experience-based, encapsulated knowledge that makes them diagnostically more efficient.

Performance-based testing has become a widely used method to assess interviewing as well as diagnostic or problem-solving skills in medical education (Van der Vleuten & Swanson, 1990). A performance-based test usually consists of a number of "stations" at which examinees are required to perform a variety of patient-related clinical tasks, including history-taking, physical examination and counselling, using simulated patients. Examinee performance is scored on checklists and rating forms tailored to the content of each station. Although simulated patient examinations have been received favourably by trainees as well as teachers, the validity has also been questioned as a result of research outcomes. First, Rethans, Sturmans, Drop, & Van der Vleuten (1991b) studied the behaviour of 39 general practitioners who were each visited during normal surgery hours by four simulated patients. It was found that doctors did not meet predetermined consensus standards of care; the mean group scores ranged from 33 to 68%. The same doctors were asked to act to the best of their ability in a controlled test at the medical school (Rethans, Sturmans, Drop, Van der Vleuten, & Hobus, 1991c). Performance during the controlled test only predicted performance during practice when efficiency and consultation time were taken into account. Secondly, several studies have been conducted using performance-based tests, in which no differences were found between medical students and experts (or experts performed even worse than students), no relationship was found between performance and years of experience, or no effects of medical education could be established by means of performance-based tests (see for a review of these studies Norman, 1994). As it is unlikely that experienced doctors are no better or even worse than beginners, the validity of performance-based tests can be questioned. Simulated patients have proven to be valid "stimuli", as they could not be differentiated from real patients by doctors (Norman, Tugwell, & Feightner, 1982; Rethans, Drop, Sturmans, & Van der Vleuten, 1991a; Sanson-Fisher & Poole, 1980). However, the way clinical skills are scored may be the problem. The checklists that are commonly used to rate doctor's performance tend to reward thouroughness (Swanson, 1992). As a result, experts may receive lower scores than students, since they gather less, rather than more, data.

The development of psychiatric expertise has hardly been studied yet. Kolodner (1983) examined the role experience plays in experts' reasoning in the field of psychiatry. She found that psychiatric experts possess two types of domain knowledge: 1) diagnostic knowledge, which is compiled so that uncomplicated cases of particular disorders can be easily identified, and 2) symptom knowledge, that is knowledge of how particular symptoms and their combinations tend to predict particular diagnostic categories. The expert, furthermore, is able to deal with unexpected or deviant cases, as a result of experience-based refinement of knowledge.

Nurcombe and Fitzhenry-Coor (1982) examined clinical reasoning in the psychiatric interview. They showed 12 third year medical students, 7 psychiatric residents, and 8 experienced psychiatrists a single videotaped patient case, and asked them to take notes with regard to the state of their reasoning. Protocols were analyzed with respect to onset of hypotheses, rationales for particular decisions in reasoning, and re-evaluation of previously generated hypotheses in light of new evidence. No linearly increasing relation was found in favour of those with the most experience. This may, however, be the result of the fact that only one patient case was presented, which may yield unreliable data, as expertise is highly content-specific. The indicators used in this study may furthermore be less relevant, since it has been argued that expertise is not characterized by extensive reasoning, but, on the contrary, by taking shortcuts.

In summary, diagnostic expertise does not seem so much a matter of extensive history-taking or superior reasoning skills. Diagnostic expertise is based on domain-specific, compiled diagnostic knowledge. Expertise can be characterized by diagnostic efficiency and is case-specific. If performance-based tests are to be used, these should consist of rating procedures that equally reward qualities such as speed and accuracy, and not only comprehensiveness. The number of cases selected may, furthermore, be essential in studying expertise.

In the present study, differences between novices and experts in the field of psychiatric diagnosis were examined. The study first attempts to obtain insight in what constitutes psychiatric expertise. A deeper understanding of the skills involved in expert psychiatric diagnosis may improve practice and contribute to the development of more efficient teaching methods. Secondly, there is a need for valid procedures to assess diagnostic interviewing skills of mental health professionals in training (Shea & Mezzich, 1988). If one is to teach these skills, it is essential to have valid procedures available that make a fair and informative assessment possible. Recently, a set of such measurements has been developed (see Bögels et al., 1994). It concerns four procedures to measure various dimensions of the psychiatric interview: the ability to conduct the interview in an effective manner, the quality of the mental representation of the information provided by the patient, the accuracy of the resulting diagnosis, and patient's satisfaction with the communication of the interviewer. It was assumed that if these dimensions of the diagnostic interview develop through training and practice, differences on these measures should be demonstrated between experts and novices. Such differences would allow for an evaluation of these measurements' discriminant validity and, hence, their utility in both teaching and assessment.

Method

Subjects

The expert-group consisted of mental health professionals, working in various institutes of mental health care. Only those were included who had at least 2 years experience in diagnostic interviewing of patients with mental disorders and who were familiar with classifying mental health disorders by means of the DSM-III-R (American Psychiatric Association, 1987). The expert-group consisted of 12 Ss; 5 psychiatrists, 6 clinical psychologists, and 1 social worker with a counselling degree. Their mean age was 34.8 years (range 25-43 years), and their mean experience with psychiatric interviewing was 7.1 years (range 2-13 years). The expert-group consisted of 9 men and 3 women.

Novices were 15 students of Mental Health Sciences who were in the third year of a four-year training programme[2]. They were extensively trained during their study in diagnostic interviewing and diagnosing mental disorders by means of the DSM-III-R, but they had not followed clinical clerkships yet. Although they had no clinical experience, they had substantial experience in interviewing simulated patients, which was part of their training. The student-group consisted of 2 men and 13 women. Their mean age was 22.5 years (range 20-29 years). Experts as well as students voluntarily participated in the study.

Cases

Four standardized case histories were developed, derived from real patients that were referred to the community mental health centre. The cases concerned a 33 year-old married man with manic symptoms and delusional grandiosity, a 32-year-old unmarried man who heard voices and had a history of psychotic episodes, a 51-year-old divorced woman who felt depressed and had a history of depressive episodes, and a 27-year-old woman who became paranoid after using drugs. The cases were enacted by four simulated patients, each portraying one case. The simulated patients had considerable experience in role-playing psychiatric cases for education purposes. They were extensively trained to simulate the case with high fidelity and constant over time.

[2] The novices were for the most part the same Ss that participated in a study on the effects of training (Bögels, 1994)

Instruments

Diagnostic Interviewing Rating Scale for Mental Disorders

The DIRSM (Bögels et al., 1994) is a rating form to assess diagnostic interviewing skills of (future) mental health professionals. The DIRSM consists of two scales called "process skills" and "content skills". Process scores reflect ratings on 10 aspects of interpersonal and communication skills, such as asking questions, summarizing, thinking aloud and giving information. Content scores reflect ratings on 13 aspects of information-gathering ability, including history of the main problem, psychiatric examination, personal, social and family history. Content skills can furthermore be divided into two subscales: ability to gather diagnostic information, and ability to gather etiologic and contextual information. The rating form is guided by a manual in which each aspect is operationalized in behavioural terms, and the function of each aspect is described in detail. The aspects are scored on a behaviourally anchored 5-point scale ranging from "insufficient" (1) to "advanced" (5). Finally, raters are asked to provide global judgements of process and content interviewing quality of the interviewer, both on a 10-point scale. Compared to detailed checklists to measure medical history-taking, a rating form like the DIRSM has the advantage that different approaches of equivalent quality can be rewarded.

Recall

In order to assess memory for the clinical cases presented, Ss were instructed to "recall everything they remembered about the patient they had just interviewed". This is a well-known procedure to chart the mental representation of a case by the diagnostician (Schmidt & Boshuizen, 1993). These recalls were audio-taped. For each case a scoring list was developed to measure the amount of relevant patient information recalled. The lists were based on the case histories role-played by the simulated patients. Each list contained 30-40 items of relevant patient information concerning the categories 1) diagnosis, 2) differential diagnosis, 3) etiologic factors, 4) predisposing factors, and 5) help request and previous therapy. Each item was scored on a 2-point scale; 0=not mentioned, 1=mentioned by the interviewer.

Diagnostic accuracy

The quality of the diagnosis was measured by means of the Problem Solving Questionnaire (Kraan & Crijnen, 1987). The interviewer responded to 7 open-ended questions concerning the categories 1) (differential) diagnosis, 2) explanatory hypotheses, and 3) plans for further treatment. In order to rate the answers of the Ss, a scoring key was developed as follows. Five psychiatric experts were asked to study the written scripts of the cases and to complete the Problem Solving Questionnaire. Based on their reactions a list of accurate, semiaccurate

and wrong answers was made for each case. Accurate answers received 2 points, semi-accurate answers 1 point, and wrong answers 0 points.

Patient Satisfaction

The Patient Satisfaction with Communication Checklist (PSCC, Crijnen & Kraan, 1987) is a questionnaire to measure patient satisfaction related to physicians' communication. A shortened version of the PSCC (Kraan, Crijnen, Van der Vleuten, & Van Thiel, 1990), that consists of five subscales (facilitation, insight, intention to comply, disruption of communication, and reassurance) was modified for use in the psychiatric context. In our version the subscale insight was removed, since the subscale is in its present form not applicable in a psychiatric interview. In addition, the items from a short patient satisfaction questionnaire, developed to assess patient satisfaction in problem-based interviews (Bögels, Mourik, & Van der Vleuten, 1994), were added. A list of 18 statements resulted. Simulated patients indicated how strongly they agreed or disagreed with the statements, on a 5-point scale from 1 (totally disagree) to 5 (totally agree). An example of a statement is: "The interviewer encouraged me to talk about what's bothering me".

Reliability analysis of the present checklist reveals that the internal consistencies of the original subscales are low to moderate (facilitation $\alpha = .67$, compliance $\alpha = .58$, disruption of communication $\alpha = .36$, reassurance $\alpha = .68$), the internal consistency of the total score, however, is satisfactory ($\alpha = .82$).

Case fidelity

To control for the quality of cases, Ss were asked to indicate how realistic the patient portrayed the role on a 5-point scale (ranging from 1 = very unrealistic/low fidelity, to 5 = very realistic/high fidelity). The fidelity of the simulated patient cases was high, experts judged the role-plays as more realistic than students (resp. 4.3 and 3.9), this difference was however not significant (F (1,25) $= 2.6$, $p = .123$).

Procedure

Subjects were instructed to conduct a psychiatric interview with the four simulated patients, in order to make a diagnosis and management plan. For each interview the maximum time was 30 minutes, which is rather short for a psychiatric interview. It was assumed however, that differences between novices and experts with regard to diagnostic efficiency might become clearer with limited time available. The interviews were videotaped. Immediately after each interview Ss were asked to recall (on audio-tape) all information they remembered about the patient they had just interviewed. Recall time was unlimited, Ss, however, seldom took more than 5 minutes for recall. Between the interview and the recall Ss had to perform a half-minute counting task, in order to control for

recency effects (Peterson & Peterson, 1959). Finally, Ss completed the Problem Solving Questionnaire concerning the patient they had just seen. They were allowed to use the DSM-III-R to answer the questions about diagnosis. Time for filling in the questionnaire was limited to 10 minutes in order to prevent Ss from "studying" the DSM-III-R.

Scoring

Four students of Mental Health Sciences were asked to rate the skills of the interviewers. All raters were at the end of their studies and had completed their 6-month clinical clerkship in a mental health institute. Raters were extensively trained in using the DIRSM, to a satisfactory inter-rater agreement. Inter-rater reliability, based on the ratings of all four raters of 10 Ss interviewing different simulated patients, was 0.78 for process skills, and 0.86 for content-related skills (intraclass correlation, including differences in rater stringency as measurement error). Because prior research (Bögels et al., 1994) showed that one rater provides sufficient reliable ratings of process and content interviewing ability, at least if different raters are used per case, the videotaped interviews were scored by only one rater. Raters were nested within patients; rater 1 scored the interviews of all Ss with patient A, rater 2 scored the interviews of all Ss with patient B, and so on. Videotapes were rated in aselect order, and raters were not informed whether the interviewer was a novice or an expert. Audiotaped recall of all Ss was rated by one of the raters of the videotapes. The completed Problem Solving Questionnaires of the Ss were rated by another rater (also one of the four videotape raters), based on the scoring key.

Total sumscores and subscale sumscores across the four cases were calculated for all variables. Because recall had different maximum scores in different cases, a proportional score was calculated for this variable.

Results

A multivariate analysis of variance was carried out to test for overall differences between novices and experts on total scores of the five dependent variables (process skills, content skills, diagnostic accuracy, recall, and patient satisfaction). A significant overall difference was found between novices and experts in favour of the experts (F Hot. $(5,21)=10.9$, $p=.000$). Table 8.1 contains means, standard deviations, and univariate F-tests of total scores and subscale scores of the dependent measures.

Table 8.1. Mean scores, standard deviations between brackets and univariate *F*-test of novices ($N=15$) and experts ($N=12$) on all variables (total and subscale scores) across four cases

Variable		Novices		Experts		*F*-value
Process skills						
	-total	45.6	(4.2)	51.6	(4.4)	12.5**
Content skills						
	-total	37.5	(4.9)	47.6	(3.8)	34.1***
	-diagnosis	21.3	(2.6)	26.9	(2.2)	36.4***
	-etiology	11.3	(1.6)	13.6	(2.2)	10.2**
Recall						
	-total	27.2	(5.8)	36.3	(7.2)	13.1**
	-diagnosis	36.1	(8.1)	48.0	(8.1)	14.6**
	-differential diagnosis	11.5	(9.0)	12.4	(11.4)	.1
	-etiology	28.2	(10.3)	38.1	(10.1)	6.3*
	-predisposition	8.3	(7.0)	12.4	(11.5)	1.3
	-treatment	23.6	(10.5)	34.5	(17.6)	4.0
Diagnostic accuracy						
	-total	5.7	(1.6)	8.7	(1.7)	22.4***
	-diagnosis and d.d.	1.6	(0.4)	2.4	(0.8)	11.0**
	-explanatory hypotheses	2.3	(0.9)	3.4	(0.5)	14.6**
	-treatment plan	1.8	(0.9)	2.9	(0.7)	12.5**
Patient satisfaction						
	-total	62.9	(5.6)	66.1	(4.2)	2.7
	-facilitation	10.5	(0.9)	11.2	(0.8)	4.9*
	-compliance	14.5	(1.5)	15.5	(0.9)	3.9*
	-disruption	15.2	(1.0)	15.5	(0.6)	.8
	-reassurance	6.2	(1.0)	6.6	(0.9)	1.3

*** $p < .001$
** $p < .01$
* $p < .05$

Univariate analyses of variance show that experts outperform novices with respect to process as well as content interviewing skills (see Table 8.1). Inspecting the two subscales of content skills (ability to gather diagnostic information and ability to gather etiologic or background information) it is found that differences between experts and novices are highest with regard to the ability to gather diagnostic information.

In Table 8.2 the scores of novices and experts on each item of the DIRSM are presented, to show differences in interviewing behaviour between novices and experts in more detail.

Table 8.2. Mean scores and standard-deviations between brackets of novices (N=15) and experts (N=12) across four cases on the items of the DIRSM (item 1-10 are the process skills, item 11-23 are the content skills)

Process and Content Skills		Novices	Experts
1.	Attending	4.1 (.6)	4.5 (.5)
2.	Asking questions	4.3 (.4)	4.7 (.5)
3.	Concreteness	4.4 (.4)	4.8 (.4)*
4.	Reflection of feelings	4.0 (.4)	4.7 (.5)**
5.	Paraphrasing of content	3.2 (.4)	3.5 (.9)
6.	Summarizing	3.7 (.6)	3.9 (.8)
7.	Giving information	3.9 (.7)	4.7 (.3)**
8.	Asking feedback	3.6 (.7)	4.0 (.7)
9.	Thinking aloud	3.9 (.6)	4.6 (.4)**
10.	Structuring	3.6 (.5)	4.1 (.3)**
11.	Clarification of problems	3.4 (.4)	4.3 (.4)***
12.	Exploration of other problems	2.4 (.4)	3.4 (.7)***
13.	Reason for encounter, help request	2.7 (.7)	3.6 (.4)***
14.	History of main problem	2.6 (.6)	4.1 (.4)***
15.	Analysing causal factors	2.6 (.5)	3.3 (.8)*
16.	Exploration of duration and course	3.0 (.7)	3.5 (.5)*
17.	History of prior treatments	2.4 (.5)	2.9 (.4)*
18.	Psychiatric examination	2.4 (.9)	2.2 (.6)
19.	Examination of consequences for daily life	2.3 (.5)	2.8 (.4)**
20.	Personal history	2.2 (.7)	2.6 (.8)
21.	Assessment of premorbid personality	1.5 (.3)	2.0 (.6)*
22.	Exploration of social relations	2.4 (.7)	2.8 (.7)
23.	Exploration of occupational, financial, and housing conditions	2.6 (.4)	3.0 (.6)

*** $p < .001$
** $p < .01$
* $p < .05$

Experts perform better than novices on almost all items. Differences are highest with regard to the process skills reflection of feelings, giving information, thinking aloud and structuring. With regard to content skills, experts differ most from novices in clarification of the presenting problem, exploration of other problems, clarifying the reason for encounter and help request, analyzing the main problem, and analyzing causal factors. The interview patterns of novices and experts are much the same; both, for example, do not pay much attention to examination of the premorbid personality, relative to other behaviours, and both direct most attention to the clarification of the problems and symptoms.

Experts recall significantly more relevant patient information than novices. Analyses at the subscale level of this variable reveal that differences exist with

regard to the recall of diagnostic, etiologic, and treatment-relevant data. However, experts do not recall more differential-diagnostic information. Furthermore, they do not recall significantly more information on predisposing factors (personality and hereditary factors) than students. As the category predisposing factors was rather small (2 or 3 items) the latter finding may be attributed to the high variability in scores -as indicated by the high standard deviations- rather than to the fact that experts do not memorize this category of patient information.

Experts also perform better than novices on diagnostic accuracy. They produce higher-quality (differential) diagnosis, better analyses of important precipitating, predisposing, and maintaining factors (higher-quality explanatory hypotheses), and higher-quality plans for further management.

Somewhat unexpectedly, the simulated patients were overall not significantly more satisfied with the way the experts communicated with them than the way the students did. On subscale level, some differences were found in favour of the experts; experts received higher scores on "facilitation" and "intention to comply". Patients were well satisfied with both student and expert interviewers; interviewers received resp. 78% and 81% of the maximum satisfaction score.

Discussion

The aim of this study was to examine differences between novices and experts in the field of psychiatric interviewing and diagnosis, in order to obtain insight in what constitutes diagnostic expertise in psychiatry, and to test the validity of recently developed measures to assess diagnostic interview ability.

Experts performed significantly and substantially better than novices with respect to process and content-related interviewing. They recalled more relevant patient information and made more accurate diagnoses. These findings imply that the procedure applied, as well as the instruments used to measure different facets of diagnostic ability in psychiatry, can be considered sufficiently valid. The results of this study clearly contrast with similar studies of medical expertise, in which no differences were found between students and experts, probably resulting from the use of measures that reward thoroughness, rather than diagnostic efficiency (Norman, 1992; Norman, 1994; Swanson, 1992). The design of the present study took these negative findings into account by limiting the amount of time available for the tasks, and by using instruments that reward the collection of *relevant* information (rather than all information). In addition, a number of cases, rather than just one case, were presented to the Ss to control for possible effects of case specificity.

Although raters were blind to the nature of the Ss' level of expertise, it is possible that raters, while rating the videotaped interviews, were able to differ-

entiate experts from novices a priori, as a result of the differences in age or typical differences in behaviour. We do not know in what way this awareness could have influenced the results. Knowing that a subject is an expert could work both ways: one could be more lenient, or, on the contrary, more stringent. Moreover, taking into account the consistency of differences between experts and novices on the behavioural and on cognitive measures (age differences, for example, could not be detected in autiotaped recall or in written diagnostic accuracy) we do not expect that results were influenced by the problem of partial blinding.

There were large gender differences in the novice- and expert-group; novices were mainly women and experts were mainly men. These differences are, however, representative of the present situation in our country; experienced psychiatrists and psychologists who work as diagnosticians are mainly men, whereas the students who are at present educated for those professions are mainly women. The question is, whether these differences may have influenced the results. Gender differences have been found with regard to interpersonal behaviour in general. Several studies showed that women are rated as more empathic than men (DeWolfe, Jackson, & Winterberger, 1988; Fox, Gibbs, & Auerbach, 1985; Whitehead & Nokes, 1990), although in de study of DeWolfe, Jackson, and Winterberger (1988) women only showed higher emotional empathy, whereas men showed higher roletaking empathy. In a study directed to first-year medical students' communication skills during a simulated patient encounter, women students were rated as more empathic, warmer and competent than men students by the judges, whereas no differences were found in simulated patient satisfaction (Marteau et al., 1991). To our knowledge, there are no indications that gender differences influence diagnostic ability. Therefore, we do not expect that the study's main result, that experts outperformed novices on most measures, would change when more women experts or men students had participated. It may however be, that the lack of difference in simulated patient satisfaction between novices and experts has been affected by gender, since, of the used measures, patients' satisfaction may be most strongly influenced by the amount of empathy communicated by the interviewer.

What interviewing behaviours characterize the experts? With regard to the interviewing process, experts were better in providing their patients with information about the purpose of the interview, further treatment, and so on. They were also more outspoken with regard to their own thoughts, considerations and hypotheses during the interview. So, experts seem more skilled in the mutual exchange of information between helper and patient, whereas novices may concentrate primarily on collecting information from the patient. Experts also demonstrated higher-quality skills in structuring the interview, which may be an indication of their diagnostic efficiency. Considering the content-related interviewing skills, experts outperformed novices most in diagnostic information-gathering. However, novices followed the same interview pattern as experts;

paying relatively more attention to the problems and symptoms of the patient than to background issues like personal and social history. These results give important feedback to educators, responsible for the training of students in psychiatric interviewing. Although students, being extensively trained in psychiatric interviewing, are developing expert-like interviewing behaviour in general, they need more training in particular with regard to the gathering of diagnostic information. In addition, training needs to focus more on the mutual exchange of information between helper and patient, such as sharing information with the patient and thinking aloud.

Experts generally recalled more relevant patient information than students. An exception was found with regard to the recall of *differential*-diagnostic information; experts and students reproduced an equal amount of differential-diagnostic information (12% of all predetermined relevant differential-diagnostic information). This finding is not atypical for case representation studies using recall as the dependent measure. According to Schmidt and Boshuizen (1993), experts use much of their knowledge in "encapsulated" mode. That is: they develop shortcuts in their thinking, embedding information acquired from a patient into fewer concepts of greater generality. Psychiatric experts possess knowledge about exclusion criteria associated with particular disorders (Kolodner, 1983), like "known organic factor" or "prominent hallucinations". If, however, these symptoms are rare in a certain clinical syndrome, experts may not recall *the absence* of these symptoms, as a result of this encapsulation.

As expected, experts' diagnoses of the cases and their plans for future management was clearly better than students'. Research into medical diagnostic expertise reveals that diagnostic accuracy is highly dependent on domain-specific knowledge and needs much exposure to a variety of clinical problems to develop (Schmidt, Norman, & Boshuizen, 1990). The students that participated in the experiment had received extensive theoretical instruction and psychiatric interview training with regard to affective and psychotic disorders, but were exposed to patients (real or simulated) only to a limited extent. Therefore, particularly clinical experience is what they may need to further improve their diagnostic accuracy.

Contrary to expectation, experts' more professional behaviour and better understanding of the patient's problems did not result in significantly higher overall patient satisfaction scores. Apart from gender differences, which were discussed above, three explanations can be offered for this finding. First, a simulated patient may not be able to experience problems in the same way as a real patient and, therefore, cannot experience in depth the comfort provided by the experienced interviewer. Secondly, experts as well as novices produced fairly high satisfaction scores in their patients, so failing to find a difference may be attributed to a ceiling effect in the data. Thirdly, patient satisfaction may not be a relevant outcome measure, because the purpose of a psychiatric interview is not so much to satisfy the patient, but to diagnose the problem and make a plan

for treatment. In psychiatry, establishing a good working relationship with a patient is rather a *precondition* for successful treatment, so a certain amount of satisfaction may be important. As the results showed, both groups seem to meet these preconditions. Although no differences were found between novices and experts on total patient satisfaction scores, some difference emerged in favour of the experts on two dimensions of patient satisfaction. The higher score on "facilitation" suggests that patients felt more stimulated by the experts to talk about their complaints, for voicing concerns and for asking any questions. The higher "compliance" score suggests that patients were more determinated to cooperate with the experts. It should however be noted that the internal reliability of the subscales is only moderate in the present study. Therefore, further research needs to be carried out on dimensions of patient satisfaction in the psychiatric context, before more definite conclusions can be drawn.

This study was one of the first efforts to investigate the nature of psychiatric diagnostic expertise. It is our hope that, through studies such as this, expertise in diagnostic interviewing in the domain of psychiatry will enter into the spotlight of the profession's attention. Patients whose problems have to be analyzed and understood, as well as students who need to acquire diagnostic (interviewing) skills, may benefit from further insights into what constitutes psychiatric expertise.

Chapter 9
Discussion

The previous chapters offered an introduction into the present situation with regard to the training in diagnostic interviewing for the mental health professions, a description of a training programme to teach diagnostic interviewing to undergraduate students, an empirical evaluation of the effects of the training programme, and studies with regard to the psychometric properties of authentic assessment of diagnostic interviewing performance. In this chapter, the most important findings of the different studies are summarized and their implications discussed. Directions for future research are outlined.

Case-specificity of diagnostic interviewing performance

Variability of performance across different cases, generally referred to as the "case-specificity of performance" (Elstein, Shulman, & Sprafka, 1978), is an important source of variance in virtually all assessments of clinical performance. Doctors' knowledge and skills are found to be content- or case-related to a large extent. Therefore, performance on one case is a poor predictor of performance on other cases. As a result, many cases need to be sampled in order to obtain a reproducible estimation of a doctor's skills.

It was hypothesized that diagnostic interviewing performance of mental health workers is also case- or content-specific. Two experiments were reported in this thesis in which the case-specificity of diagnostic interviewing performance of (future) mental health workers was studied. Two aspects of diagnostic interviewing were distinguished: *process* interviewing skills, reflecting the interpersonal and communication skills, and *content* interviewing skills, referring to the information-gathering ability of the interviewer. In a first study (Chapter 3) third-year students of Mental Health Sciences were instructed to conduct a diag-

nostic interview with four simulated patients who portrayed different mental health disorders from various diagnostic categories: anxiety disorders, psychotic disorders, and affective disorders. It was found that, consistent with previous findings from the medical domain, diagnostic interviewing performance on one case proved to be a poor predictor of performance on other cases. Content as well as process scores suffered from case-specificity. In a second study (Chapter 4), a comparable group of third-year students was subjected to a similar diagnostic interview test with four simulated patients. The range of diagnostic categories of the cases, however, was now restricted to affective and psychotic cases. The variability of content scores across cases was found to be smaller than in the first study. So, it appears that case-specificity of content scores is influenced by the range of cases of different diagnostic content. In a third study (Chapter 5) first-year students were instructed to interview four simulated patients who portrayed psycho-social problems. The case-specificity of their process skills scores was small, compared to the process scores of third-year students who interviewed cases portraying mental disorders. This finding may be explained by the fact that the interactional differences of the psychosocial cases were smaller than the interactional differences of the mental health cases (the difference in interactional style between a manic and a depressed patient may produce large variability in performance with regard to individual interviewers' process skills). Variability across cases was furthermore influenced by the time, available for each interview; case-specificity of process skills appears to increase in longer simulated patient encounters (Chapter 5). An explanation for this finding is that differences in communication style of the patient (for example talkative or silent) influence the interaction between interviewer and patient to a larger extent when more time is available. Individual differences in preference and proficiency to handle, for example, a talkative or silent patient (case-related performance) may become more prominent in longer interviews.

Generally, research on authentic assessment in medical education reports high proportions of variance resulting from case-specificity. Compared to those results, in our studies case-specificity accounts for less of the total variance. Two factors may contribute to this finding. First, we used a rating form to measure diagnostic interviewing performance, whereas in studies on authentic assessment of doctors, generally, checklists are used (Van der Vleuten & Swanson, 1991). Checklist scores seem more "anchored" to the content of the station (more case-specific) and therefore produce larger variability of scores across stations (van Luyk & Van der Vleuten, 1992). Secondly, we restricted the assessment to diagnostic interviewing, whereas in medicine often a wider range of skills are measured, like history-taking, physical examination, and laboratory skills. Consequently, the medical studies tend to measure case-specificity (variance caused by differences in performance across cases of different content) and task-specificity (variance caused by differences in performance across types of tasks) together, which may yield higher case-specificity variance.

Summarizing, our findings suggest that case-specificity is indeed an important source of variance which needs to be taken into account in measuring diagnostic interviewing performance in the mental health field, but case-specificity varies depending on the range of diagnostic categories included in a test, the degree to which patients' communication styles vary, the type of skills measured, and the time available for the interview. When using 4 to 8 cases (dependent on the range of diagnostic categories to be represented in the test) a reproducible estimate of the quality of diagnostic interviewing performance of individual interviewers can be made. Two cases will generally be sufficient to assess the quality of basic interviewing performance of beginning students, provided that raters are sufficiently trained.

Inter-rater reliability of diagnostic interview scores

In general, agreement between raters with regard to observational measures does not exceed .50, if measures are used that correct for coincidence agreement and include absolute differences between raters as measurement error, e.g. kappa or intraclass correlation (Swanson, 1992). The results of the different studies yielded quite variable inter-rater reliability estimates, ranging from .35 (Chapter 5) to .78 (Chapter 3) for process scores, with the same instrument. A number of factors may have been responsible for this variability in results: the amount of training the raters received, the homogeneity of the sample of raters, and the type of skill assessed. Inter-rater agreement with regard to process skills, when assessed separately from content skills, seems to be more difficult to achieve (Chapter 5).

A factor with regard to inter-rater reliability that deserves special attention is the scoring format. In this thesis, a rating form was used to assess diagnostic interviewing performance. This choice was based on validity arguments. First, communication skills are difficult to assess with checklists without trivializing the content. Secondly, expert raters tend to feel underestimated if they have to use checklists that are stringent and provide little room for expert judgement. Thirdly, checklists may have a negative impact on student learning. Studies from the medical school of the University of Limburg revealed that as a result of the fact that students started to memorize previously used checklists over the years, a deeper understanding of the skills appeared to be missing (Van Luyk, Van der Vleuten, & Van Schelven, 1990). Global ratings, on the other hand, tend to suffer more from inter-rater disagreement than analytic ratings or checklists (Streiner, 1985). Van Luyk and Van der Vleuten (1992) showed that inter-rater reliabilities for checklists are better than for global ratings, but checklists produce higher variability of persons across cases than rating scales, resulting in equally high reproducibility coefficients of rating-scores compared to checklist-

scores. This finding illustrates that inter-rater reliability needs to be considered together with other sources of measurement variance and measurement error, in order to make rational choices with regard to the design of authentic assessment. In educational practice, it may be more difficult to control inter-rater reliability than in experimental settings. As such, one needs to take into account that inter-rater reliability may prove to be low. There are at least three possibilities to handle this problem: use multiple ratings, use many cases and different raters per case (so that differences between raters are averaged out), or ask a second opinion or extend the number of cases in case of close pass-fail decisions.

Validity of diagnostic interviewing performance, measured under authentic assessment conditions

Having established that diagnostic interview performance can be assessed with sufficient reliability, the crucial question is whether the scores, obtained under authentic assessment conditions, are valid. Several validity issues were investigated. A first question concerns the construct validity of the skill-rating. Adequate construct validity is indicated if aspects of interviewing behaviour are assessed that contribute to the goals of a diagnostic interview: establishment and maintenance of a good working relationship, and collection of information to arrive at a diagnosis. Correlational analyses consistently indicated that skill-ratings were substantially related to relevant outcome-measures (Chapters 3 and 6); process skills were highly correlated to simulated patients' satisfaction, whereas content skills were highly correlated to the amount of relevant information given by the patient and diagnostic accuracy.

Secondly, the discriminant validity of the test was studied. It was assumed that, in order to be sensitive to growth in diagnostic interviewing skills, the test should demonstrate differences between experts and novices. Experts outperformed students on process as well as content skills (Chapter 8). This finding may seem self-evident. However, in previous studies on physicians' performance measured with authentic assessment, often no differences were found between medical students and experts, or students even outperformed experts (Norman, 1994). The fact that a rating form was used in our study, whereas checklists were used in the medical studies, may have contributed to our positive findings. Rating forms may, better than checklists, reward what skilled behaviour in fact is: performing some activity with flexibility and adaptation to the specific demands of that situation (Lowyck, 1991).

A third validity issue concerns an aspect of content validity: How authentic was the authentic assessment? Our data supported the authenticity of the cases and the role-playing of the simulated patients; subjects consistently indicated that the cases, as enacted by the simulated patients, were very realistic (Chapters 3,

5, and 8), and experts gave the simulated patients even higher fidelity scores than students did (Chapter 8). Various studies on the validity of simulated patients in the medical domain showed that doctors were unable to differentiate simulated patients from real patients (see, for example, Rethans, Drop, Sturmans, & Van der Vleuten, 1991a). So, there is evidence that well-trained role-players can credibly simulate medical as well as psychiatric cases.

Another aspect of content validity is the time given for a simulated patient encounter. In Chapter 5 this issue was investigated, by varying the length of the encounter for the same task; conducting a problem-based interview. The major finding was that absolute differences between persons (indicated by the desirable true person variance) tend to disappear when longer time is available. In line with this, various studies on medical expertise could only detect differences between expert clinicians and medical students, when very short time was given for the task (Muzzin, Norman, Faightner, Tugwell, & Guyatt, 1983). Summing up, there are several arguments to restrict time per single simulated patient encounter in competency decisions; short encounters reward efficiency, produce higher differences between examinees (are more sensitive), and are more feasible.

One important remaining validity issue is the representativeness of the interviewers' behaviour, displayed under authentic assessment conditions. A study of Rethans et al. (1991c) showed that what doctors do under authentic assessment conditions, when they are instructed to act to the best of their ability and have unlimited time, does not correlate with their behaviour in real life. This has been termed the difference between "competence" and "performance". It seems that not only time restriction, but also the instructions the subjects are given, are extremely important for the type of behaviour that is induced by authentic assessment. In accordance with this, Swanson (1992) argued that, since data-gathering activities with simulated patients are usually open-ended and uncued, this vagueness may result in differences in task perceptions or unintended task perceptions. So, high fidelity simulations may not necessarily produce high-fidelity behaviour. Therefore, it is desirable to give careful task instructions and to check how tasks are perceived.

Finally, for educational purposes, the "consequential validity" (Rettie, 1992) is perhaps the most important validity aspect of a test to use for educational purposes. What are the consequences of authentic assessment on teaching and learning? Although we have no data with regard to the consequential validity of the (diagnostic) interviewing skills assessment, we have the impression that it motivates students to practise their skills, that it directs teachers towards the goals of the training courses, and that it has a desirable influence on the development of training courses, as, in order to develop valid authentic assessments, the goals and content of the training need to be clear and explicit.

Effects of training in diagnostic interviewing

The effects of the training courses in basic and diagnostic interviewing skills were investigated in this thesis, using authentic assessment. The results are briefly summarized. First, the effects of a 5-session systematic training in basic interviewing and counselling skills were evaluated. The results were clear; the training was highly effective in teaching interviewing skills to students. After training, students attained levels comparable to those of professional helpers (Chapter 6). Next, the effects of a 6-session introductory training in the diagnostic interview of patients with affective disorders were investigated. Students' performance substantially improved as a result of the training. Furthermore, skills-acquisition appeared content-specific, that is, results hardly generalized to types of disorders, not dealt with during the training. The training suffered, however, from a loss of effect over time; performance dramatically decreased after a period of 6 months, in which no training took place. Finally, the results of a graduate 6-session training in diagnostic interviewing focusing on psychotic disorders were examined. Somewhat surprisingly, no significant improvements in skills were observed after that training (Chapter 7).

From these findings it can be concluded that basic interviewing and counselling (termed "process skills" in this thesis) can be easily acquired by beginning students who follow a short training, based on the micro-counselling method. Diagnostic interviewing with regard to neurotic disorders can also be effectively trained in a short period. The acquisition of diagnostic interviewing skills with regard to severe psychiatric disorders may, however, need another approach. To be more precise, at least two factors of the training design may be responsible for the lack of effect of the last training. The first factor concerns the complexity of the skills to be acquired. Psychotic patients often make great demands on the interpersonal qualities of the interviewer, as they, for example, mistrust the interviewer, are violent, or are difficult to keep in reality. The interviewer has to perform different tasks concurrently: observation of symptoms, asking questions to clarify the problems and symptoms, handling the patient, and crisis intervention. A training format of 6 sessions may, therefore, be too short to acquire the necessary skills. This explanation is supported by other research showing that complex skills generally need much training and practice to develop (Patrick, 1992). The second factor concerns the training methods applied. The first diagnostic interview training used a highly structured training format, following micro-training principles, whereas the second diagnostic interview training applied a less structured format; the training consisted mainly of live simulations in which a psychiatric nurse portrayed a psychotic patient. Standardized presentations of parts of illustrative patient interviews seem more effective than live, spontaneous demonstrations (Carrol & Monroe, 1979). Especially since the interviewer has so many tasks at the same time while inter-

viewing psychotic patients, these tasks may better be learned in isolation, rather than integrated by means of exciting, but difficult, simulations. Apart from the content of the training, a third factor may also partially account for the lack of effect of the last training; the phenomenon that learning is initially rapid, but more and more practice is required to achieve the same level of improvement later in training (Patrick, 1992). In other words: the more training students receive, the less the effect may be.

In line with other research showing that training in various skills suffers from a loss of effect over time (Baker & Daniels, 1989; Sanson-Fisher & Maguire, 1980), diagnostic interviewing skills seem to relapse without training or application possibilities. These through training freshly acquired skills probably have not yet reached the stage of automatism (or encapsulation) that characterizes experts' skills. According to Schmidt, Norman, & Boshuizen (1990), with extensive and repeated application and particularly through exposure to patient problems, novices' knowledge and skills will reorganize itself so that accessibility is assured. Activation, then, occurs smoothly and automatically. As long as this stage of automatism has not been achieved, skills may easily be pushed aside by new education activities that require attention, and may not automatically be re-activated afterwards. The question is, therefore, how the previously acquired skills can be retrieved or re-activated, since they may not have simply disappeared. Reynolds and Glaser (1964), for example, demonstrated that the "review" of complex training materials assists retention. Like the activation of prior knowledge, which is an essential part of problem-based education, the activation of previously acquired skills may need to be part of training education.

Directions for further research

A number of issues with regard to the assessment and training of diagnostic interviewing need further clarification.

Two aspects of diagnostic interviewing (process and content skills) were distinguished in this thesis. It still has to be investigated whether the two aspects are indeed distinguishable factors. Latent structural analysis may shed more light on this issue. However, more observations of different interviewers need to be sampled before such an analysis can be carried out.

With regard to the feasibility of authentic assessment of diagnostic interviewing, more research is needed related to strategies for improving efficiency and decreasing testing time without losing reproducibility and validity. One other way to accomplish this may be to divide the diagnostic interview into parts (for example into problem clarification, psychiatric examination, personal history, and analysis of present situation) and instruct interviewers to conduct each part

of the interview with a different simulated patient. In this way, time per interview can be shortened and interviewers are at the same time exposed to different cases. A disadvantage of such an approach is that the assessment does not approximate the professional task as well as a complete interview with one patient. In addition, case-specificity may increase since it then also includes task-specificity.

Taking into account the considerable costs of establishing sufficient reproducibility of authentic assessment, a relevant question is whether there are alternatives for authentic assessment. One alternative arises from our data. The problem-solving measure, used in various studies (e.g., diagnostic accuracy, problem analysis quality) consistently correlated highly with diagnostic interviewing skills. Moreover, the measure produced high(er) inter-rater reliability coefficients and high(er) reproducibility coefficients. Therefore, a cost-saving procedure might be to have students interview a simulated patient, but only assess their problem-solving of the case. However, a risk of such an approach may be that when behaviour is no longer rated, the behaviour might change in a non-desired direction (more outcome-directed and less patient-directed). As a consequence, the resulting problem-analyses scores may reward non-desired interview styles. On the other hand, combining problem-solving scores with patient satisfaction scores might resolve this problem. Research on the issue how rating influences behaviour is desired.

In order to gain insight into effective diagnostic interviewing strategies, the contribution of specific skills to outcome (adequate diagnosis and establishment of a working alliance) deserves further attention.

Considerable work needs to be done with regard to the training of (future) mental health professionals in diagnostic interviewing. We still know very little about effective training strategies to teach these skills to novices. For example, is a cognitive training approach, focusing on the application of diagnostic knowledge during the interview process, more or less effective than a behavioural training approach that focuses on the overt behaviour to be displayed during the interview? Secondly, there is an urgent need for empirical evaluation of training programmes in higher-order skills such as diagnostic interviewing. To our present knowledge, only five empirical studies yet exist that evaluate psychiatric interview training (of which our own study was the only one conducted with non-medical students), compared to hundreds of controlled studies on the effects of training in basic interviewing, medical interviewing, and counselling skills. Without empirical studies, the value of diagnostic interview training programmes in use remain questionable. Studies on the effects of diagnostic interview training of experienced mental health professionals are also important, because many mental health workers have not been adequately trained in diagnostic interviewing and psychiatric classification during their study. Thirdly, comparisons of the effects of training in the context of the curriculum with the effects of clinical clerkships are needed, to evaluate the cost and effectiveness of both approaches

to skills acquisition. Finally, in order to handle the phenomenon of relapse of skills, strategies need to be developed to efficiently retrieve or re-activate earlier acquired skills.

Summary

The training and assessment of diagnostic interviewing skills of (future) mental health professionals was investigated in this thesis. As is outlined in Chapter 1, the establishment of a new academic programme at the University of Limburg (called "Mental Health Sciences") led to the development of a training programme to teach diagnostic interview skills to students. The execution of this progamme was coupled with research to evaluate the effectiveness of the training courses and the appropriateness of the tests, used to assess students' performance.

In Chapter 2 recent developments with regard to the teaching of diagnostic interviewing skills in the mental health professions are discussed. Training methods to teach diagnostic interviewing are reviewed, and the (scarce) empirical evidence of effective interviewing styles is outlined. Based on an analysis of the tasks and skills of the diagnostic interviewer in mental health care, three aspects of diagnostic interviewing are distinguished: 1) process skills, reflecting the interpersonal and communication skills, 2) content skills, referring to the information-gathering ability, and 3) problem-solving skills, reflecting the ability to make an accurate diagnosis and treatment plan. A training curriculum is then described, developed to teach diagnostic interviewing to undergraduate students. The training curriculum is highly structured and simulated patients are utilized to introduce complex clinical problems and to evaluate students' emerging competencies. Student and trainer satisfaction with the program has been high over the past 6 year the course has existed in this format.

In Chapter 3 a behavioural test is evaluated, developed to assess the quality of diagnostic interviewing skills of (future) mental health professionals. Twenty graduate students of Mental Health Sciences interviewed four simulated patient, who portrayed mental health disorders from various categories. It was found that diagnostic interviewing skills can be reliably measured with respect to inter-rater reliability. However, interviewer performance on one case proved to be a poor predictor of performance on other cases. It was concluded that a large number of cases (10) is required to obtain reproducible scores of general diagnostic inter-

viewing ability. The validity of the test was supported by the correlational analyses. Process skills were strongly related to patient satisfaction, whereas content skills were related to the amount of relevant information given by the patient and the accuracy of the diagnostic formulation and treatment plan.

The reproducibility of diagnostic interviewing performance was again studied in Chapter 4. Subjects now consisted of 15 graduate students of Mental Health Sciences and 12 professional diagnosticians. In contrast with the study described in Chapter 3, the mental health disorders, portrayed by the simulated patient, were restricted to psychotic and mood disorders. The results of the study suggest that by means of four interviews (2 hours testing time), a reproducible estimation of psychiatric interviewing ability of the interviewers can be made. These results clearly contrast with a the results reported in Chapter 4, in which a larger number of cases were necessary to obtain reproducible results of general diagnostic interviewing ability. It seems that the restriction of range of the diagnostic categories of the simulated patients in this study yielded more reproducible scores. It is concluded that psychiatric interviewing performance can be reliably and feasibly assessed with the proposed procedure.

The purpose of Chapter 5 was to evaluate the effects of station length on the reproducibility and validity in a test for interviewing and counselling skills. Fifty first-year students Health Sciences were randomly assigned to one of three conditions: four 5-minute interviews, four 10-minute interviews, and four 20-minute interviews. Generalizability analyses showed that the highest reproducibility was obtained with four 10-minute stations. The lowest reproducibility was found with four 20-minute stations. General differences in interviewing performance between interviewers tend to reduce in 20-minute stations. Inter-rater reliabilities were rather low, and were not influenced by station length. Validity analyses revealed that students seem to do "more of the same" in longer stations; they used the same skills more frequently. The quality of the problem-analyses were better in 20- than in 10-minute interviews, but no differences in problem analysis quality were found between 5- and 10-minute interviews. It is concluded that 10-minute stations might be the optimal choice with regard to the assessment of interviewing and basic counselling skills.

Chapter 6 and 7 evaluate the effects of training courses in (diagnostic) interviewing. At the start of their study Mental Health Sciences, a 5-session training course in basic interviewing and counselling skills is organized. In Chapter 6, the effects of this training on students of Mental Health Sciences were evaluated by means of a time-series design. Large improvements in interviewing performance were found as a result of training. Furthermore, students had higher self-efficacy expectations after training. The post-training performance of students was compared with the performance of a group of professionals. Students attained skill-levels comparable to those of professional helpers. Simulated patients were equally satisfied with the communication of students and pro-

fessionals. However, professionals judged themselves as more interpersonally effective than students did.

In Chapter 7, the immediate and long-term effects of systematic training in diagnostic interviewing were investigated. Students of Mental Health Sciences were first trained in diagnostic interviewing with regard to affective disorders. Six months later, they received training focusing on psychotic disorders. By means of a quasi-experimental design effects of instruction on a random sample of 14 students were examined. At each moment of measurement students were instructed to interview four simulated patients. Two patients simulated an affective disorder and two others simulated a psychotic disorder. Four dependent variables were measured; process interviewing skills, content-related interviewing skills, recall of relevant patient information, and diagnostic accuracy. After the first training, students' performance substantially improved with regard to content-related interviewing skills and recall of relevant patient information. The effects of training were specific for affective disorders and hardly generalized to psychotic disorders. Performance decreased after the 6 months during which students received no training. The second training did not improve performance. It is concluded that diagnostic interviewing can be effectively trained in a short period with regard to affective disorders, but teaching students how to conduct a diagnostic interview with psychotic patients may need more practice.

The nature of diagnostic expertise in psychiatry is examined in Chapter 8. Novices (15 third-year students of Mental Health Sciences, trained in psychiatric interviewing) and experts (12 psychiatrists and clinical psychologists) each conducted diagnostic interviews with four simulated patients who portrayed different mental health disorders. It was found that experts performed better than novices with regard to process interviewing skills, content-related interviewing skills, diagnostic accuracy and recall of relevant patient information. Only on a measure of patient satisfaction, no differences were found between novices and experts. The results support the validity of the measures proposed to assess growth in diagnostic skills.

Samenvatting

Dit proefschrift handelt over de training en toetsing van anamnestische interviewvaardigheden van (toekomstige) hulpverleners op het terrein van de geestelijke gezondheidszorg. In de introductie (Hoofdstuk 1) wordt de aanleiding geschetst voor dit onderzoek: de oprichting van de afstudeerrichting Geestelijke Gezondheidkunde aan de Rijksuniversiteit van Limburg, Faculteit Gezondheidswetenschappen. Het aanleren van gespreks- en anamnesevaardigheden werd van meet af aan gezien als een belangrijk onderdeel van het curriculum. De ontwikkeling van trainingen op dit gebied werd vergezeld van onderzoek naar het effect van de trainingen, en onderzoek naar de betrouwbaarheid en validiteit van toetsen om het vaardigheidsniveau van de studenten te evalueren. De resultaten van dit onderzoek worden beschreven in dit proefschrift.

Hoofdstuk 2 biedt een overzicht gegeven van de literatuur op het gebied van training in het afnemen van een anamnese bij geestelijke gezondheidsproblemen. Verschillende trainingsmethoden worden besproken: microteaching, supervisie, en training met behulp van simulatiepatiënten. Ook wordt het (schaarse) empirische onderzoek naar de effectiviteit van verschillende interviewstijlen behandeld.

Op basis van een analyse van de taken en vaardigheden van de interviewer tijdens het anamnestisch interview van een patiënt met een geestelijk gezondheidsprobleem, worden een drietal aspecten aan de complexe vaardigheid anamnestisch interviewen onderscheiden: gespreksvaardigheden, anamnesevaardigheden, en cognitieve vaardigheden. Gespreksvaardigheden zijn interpersoonlijke en communicatieve vaardigheden, die nodig zijn om een goede werkrelatie met de patiënt te vestigen, zoals: het stellen van open vragen, reflecteren van gevoel, geven van informatie. Anamnesevaardigheden betreffen het vermogen van de interviewer om die informatie te verzamelen die nodig is voor de diagnostiek en behandeling. Cognitieve vaardigheden betreffen het vermogen van de interviewer om de verkregen informatie te ordenen, diagnostische hypothesen te formuleren en te toetsen, een adequate diagnose te stellen, en een behandelplan te maken.

Een trainingsprogramma in anamnestisch interviewen is ontwikkeld voor studenten Geestelijke Gezondheidkunde. Kenmerkende aspecten van dit

trainingsprogramma zijn: integratie van theoretisch onderwijs en vaardigheidsonderwijs, sterke structurering van de trainingen, en gebruik van simulatiepatiënten, zowel om complexe klinische problemen voor te leggen als om de vaardigheden van studenten te beoordelen. Uit programma-evaluaties blijkt dat studenten en docenten erg tevreden zijn met het trainingsonderwijs.

Hoofdstuk 3 handelt over het toetsen van anamnestische interviewvaardigheden. Twintig gevorderde studenten Geestelijke Gezondheidkunde werd gevraagd bij vier simulatiepatiënten een anamnese af te nemen. De patiënten speelden vier totaal verschillende beelden: paniekstoornis, sociale fobie, depressie, en psychotische stoornis. Met behulp van een scoringslijst werden de gespreks- en anamnesevaardigheden van de interviewers beoordeeld. De interbeoordelaarsbetrouwbaarheid was zeer hoog. De prestaties van de interviewer bij één patiënt bleken echter een slechte voorspeller voor prestaties bij andere patiënten, een fenomeen dat in de medische literatuur beter bekend is als "casus-specificiteit van klinische vaardigheid". Geconcludeerd wordt dat een flink aantal casus (ongeveer 10) nodig is om een betrouwbaar (reproduceerbaar) beeld van de anamnesevaardigheden van een interviewer te krijgen.

Correlationele analyses ondersteunden de validiteit van de gedragstest. Gespreksvaardigheden waren sterk gerelateerd aan patiëntsatisfactie, terwijl anamnesevaardigheden een sterke samenhang vertoonden met de hoeveelheid relevante informatie die was verzameld, en met de kwaliteit van de diagnose en behandelplan.

In Hoofdstuk 4 is de reproduceerbaarheid van meting van anamestische interviewvaardigheden opnieuw onderzocht. De proefpersonen bestonden ditmaal uit gevorderde studenten Geestelijke Gezondheidkunde (N=15) èn experts op het gebied van psychiatrische diagnostiek (N=12). Anders dan in Hoofdstuk 3, werden de diagnostische categorieën van de simulatiepatiënten beperkt tot twee groepen stoornissen: stemmingsstoornissen en psychotische stoornissen. Uit de resultaten blijkt dat met behulp van 4 casus (2 uur toetstijd) een reproduceerbare score van de anamnestische vaardigheden van een interviewer kan worden verkregen, zowel voor de studenten als voor de experts. Deze resultaten zijn veel gunstiger dan die, gevonden in Hoofdstuk 3. Dit lijkt het gevolg van de kleinere variatie in diagnostische categorieën van de casus in het huidige onderzoek, waardoor de casus-specificiteit lager uitvalt. Geconcludeerd wordt dat betrouwbare èn efficiënte meting van anamnestische interviewvaardigheid met behulp van de voorgestelde gedragstest mogelijk is.

Gegeven de bevinding dat een aantal casus (minimaal 4) nodig is om een betrouwbare schatting van een iemands vaardigheden te kunnen maken, wordt in Hoofdstuk 5 nagegaan in hoeverre de toetstijd per casus verkort kan worden, om zo de totale toetstijd van een vaardigheidstoets te beperken.

Vijftig eerstejaars-studenten Gezondheidswetenschappen werden random toegewezen aan één van drie condities: vier 5-minuten interviews, vier 10-minuten interviews, of vier 20-minuten interviews. De hoogste reproduceerbaarheid

werd gevonden bij 10-minuten interviews, de laagste reproduceerbaarheid bij 20-minuten interviews. Het lijkt erop dat verschillen tussen interviewers verminderen wanneer zij lang (20 minuten) de tijd hebben. Uit de validiteitsanalyses blijkt dat langere interviews niet veel meer informatie opleveren dan kortere interviews: interviewers zijn geneigd meer van dezelfde, toch al veel gebruikte, vaardigheden toe te passen, terwijl schaars gebruikte vaardigheden in korte gesprekken ook in langere interviews weinig gebruikt worden. Conform verwachting is de kwaliteit van interviewers' analyses van het probleem beter bij 20-minuten dan bij 10-minuten interviews, echter tussen 5- en 10-minuten interviews wordt geen verschil gevonden in de kwaliteit van de probleemanalyses. Op basis van deze resultaten wordt geconcludeerd dat 10-minuten interviews de optimale keus vormen voor het meten van gespreksvaardigheden.

In Hoofdstuk 6 en 7 worden de effecten van training in (anamnestisch) interviewen onderzocht. Bij aanvang van de studie Geestelijke Gezondheidkunde volgden de studenten een training van 5 sessies waarin op systematische wijze gespreksvaardigheden worden getraind. In hoofdstuk 6 wordt met behulp van een time-series design nagegaan wat tweedejaars-studenten leren van deze training. De resultaten geven aan dat grote vooruitgang wordt geboekt wat betreft de kwaliteit van de gespreksvaardigheden. Ook resulteert de training in hogere self-efficacy. De getrainde studenten werden vervolgens vergeleken met een representatieve referentiegroep van professionele hulpverleners. Na de training is het gespreksvaardigheidsniveau van studenten vergelijkbaar met dat van de hulpverleners. Ook is er geen verschil in simulatiepatiëntsatisfactie tussen studenten en hulpverleners. De enige maat waarop de experts het beter doen is self-efficay; experts hebben een hogere verwachting van hun interpersoonlijke effectiviteit dan studenten.

In Hoofdstuk 7 worden de korte en langere termijn-effecten van training in anamnestisch interviewen bestudeerd. Studenten Geestelijke Gezondheidkunde volgden eerst een training van 6 sessies in het afnemen van een anamnese bij patiënten met stemmingsstoornissen. Zes maanden later werden zij getraind (eveneens 6 sessies) in het afnemen van een anamnese bij psychotische patiënten. Bij een steekproef van 14 studenten werd nagegaan wat het effect van het gevolgde onderwijs was op de gespreksvaardigheden, anamnesevaardigheden, recall van relevante patiëntinformatie, en de kwaliteit van de diagnose en het behandelplan. Op elk meetmoment interviewden de studenten vier verschillende simulatiepatiënten, waarvan twee een stemmingsstoornis nabootsten, en twee anderen een psychotische stoornis.

Na de eerste training verbeterden de studenten wat betreft anamnesevaardigheden en recall van relevante patiëntinformatie. Deze verbetering was "inhoudsspecifiek"; vooruitgang werd vooral geboekt op het gebied van stemmingsstoornissen en generaliseerde nauwelijks naar psychotische stoornissen. Er was echter sprake van een sterke terugval in vaardigheden na de 6 maanden waarin geen training plaatsvond. De training in het afnemen van een anamnese bij patiënten

met een psychotische stoornis leidde niet tot verbetering in prestaties. Geconcludeerd wordt dat anamnestische interviewvaardigheden in korte tijd effectief getraind kunnen worden wat betreft affectieve stoornissen, maar dat meer tijd en oefening nodig is om studenten te leren een anamnese af te nemen bij patiënten met een psychotische stoornis.

In Hoofdstuk 8 wordt de aard van diagnostische expertise op het gebied van de psychiatrie onderzocht. Novieten (15 gevorderde studenten Geestelijke Gezondheidkunde, getraind in anamnestisch interviewen en diagnostiek van mentale stoornissen) en experts (12 ervaren psychiaters en psychologen) interviewden vier simulatiepatiënten die verschillende psychiatrische stoornissen nabootsten.

Experts behaalden hogere scores dan studenten wat betreft de kwaliteit van hun gespreksvaardigheden, waarbij opvalt dat experts vooral beter waren in die vaardigheden die te maken hebben met de onderlinge uitwisseling van informatie tussen hulpverlener en patiënt (bijvoorbeeld geven van informatie en hardop denken) en het interview beter structureerden. Hoewel experts veel beter presteerden op het gebied van anamnesevaardigheden, volgden experts en getrainde studenten hetzelfde interview-patroon; zij besteedden relatief veel aandacht aan de problemen en symptomen en relatief weinig aandacht aan achtergrondgegevens zoals levensgeschiedenis en sociale context. Experts onthielden meer relevante patiëntinformatie uit het interview, en de kwaliteit van hun diagnoses en behandelvoorstellen was eveneens beter. Alleen op een meting van patiëntsatisfactie werd geen verschil gevonden tussen experts en novieten. Deze resultaten ondersteunen de discriminante validiteit van de instrumenten die ontwikkeld zijn om psychiatrische diagnostische expertise te meten.

References

American Psychiatric Association (1968). *Diagnostic and Statistical Manual of Mental Disorders* (2nd ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders* (Rev. 3rd ed.). Washington, DC: American Psychiatric Association.

Authier, J., & Gustafson, K. (1982). Microtraining: Focusing on specific skills. In E. K. Marshall & P. D. Kurtz (Eds.), *Interpersonal helping skills* (pp. 93-131). San Francisco: Jossey-Bass Inc. publishers.

Baker, S. B., & Daniels, T. G. (1989). Integrating research on the microcounselling program: A meta-analysis. *Journal of Counselling Psychology, 36*, 213-222.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioural change. *Psychological Review, 84*, 191-215.

Bandura, A. (1989). Regulation of cognitive processes through perceived self-efficacy. *Developmental Psychology, 25*, 729-735.

Barrows, H. S. (1971). *Simulated patients (programmed patients): The development and use of a new technique in medical education.* Springfield Il.: Charles C. Thomas.

Barrows, H. S. (1980). *Problem-based learning: An approach to medical education.* New York: Springer Publications.

Blok, G. A. & Bögels, S. M. (1990). Teaching professional skills in mental health care with simulated patients. In H. E. Klein (Ed.), *Problem solving with cases and simulations* (pp. 79-83). WACRA.

Bögels, S. M. (1992). Training in gespreksvaardigheden: Evaluatie van studieresultaten [Training in interviewing skills: Assessment of student learning]. In: van der Vleuten, C. P. M., Scherpbier, A. J. J. A., Pollemans, M. C. (Eds.), *Gezond Onderwijs* (pp 304-311). Houten: Bohn Stafleu van Loghum.

Bögels, S. M. (1994a). A structured training approach to teaching diagnostic interviewing. *Teaching of Psychology*, in press.

Bögels, S. M. (1994b). Training diagnostic interviewing to clincal psychology students: Immediate and long-term effects. Manuscript submitted for publication.

Bögels, S. M, & Kreutzkamp, R. (1990). Effecten van een training in basale gespreksvaardigheden [Effects of training in interviewing and counselling skills]. *Tijdschrift voor Onderwijsresearch, 15*, 201-214.

Bögels, S. M., Mourik, T. G. C. van, & Vleuten, C. P. M. van der (1994). Authentic assessment of interviewing and counselling skills: Effect of testing time on reproducibility and validity. *Teaching and Learning in Medicine* (in press).

Bögels, S. M., Vleuten, C. P. M. van der, Blok, G., Kreutzkamp, R., Melles, R., & Schmidt, H. G. (1994). The diagnostic interview in mental health care: Assessment and validation of process and content-related interviewing skills. Manuscript submitted for publication.

Borus, J. F., & Yager, J. (1986). Ongoing evaluation in psychiatry: The first step toward quality. *American Journal of Psychiatry, 143*, 1415-1419.

Bowman, F. M., Goldberg, D. P., Millar, T., Gask, L., & McGrath, G. (1992). Improving the skills of established general practitioners: The long-term benefits of group teaching. *Medical Education, 26*, 63-68.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: The American College Testing Program.

Carrol, J. G., & Monroe, J. (1980). Teaching clinical interviewing in the health professions. *Evaluation of the Health Professions, 3*, 21-45.

Carrol, J. G., & Monroe, J. (1979). Teaching medical interviewing: A critique of educational research practice. *Journal of medical Education, 54*, 498-500.

Collins, H. (1985). Changing order. Chapter 3: Replicating the TEA-laser: Maintaining Scientific Knowledge, pp. 51-78. London: Sage.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin Company.

Cox, A., Holbrook, D., & Rutter, M. (1981). Psychiatric interviewing techniques: VI. Experimental study: Eliciting feelings. *British Journal of Psychiatry, 139*, 144-158.

Cox, A., Hopkinson, K., & Rutter, M. (1981). Psychiatric interviewing techniques: II. Naturalistic study: Eliciting factual information. *British Journal of Psychiatry, 138*, 283-291.

Cox, A., Rutter, M., & Holbrook (1981). Psychiatric interviewing techniques: V. Experimental study: Eliciting factual information. *British Journal of Psychiatry, 138*, 283-291.

Crijnen, A. A. M., & Kraan, H. F. (1987). Measuring patient satisfaction with the communication. In *The Maastricht History-taking and Advice Checklist: Studies of instrumental utility* (eds H.F. Kraan & A.A.M. Crijnen), pp. 145-171. Ph-d thesis, Amsterdam: Velden van den Hazelaar.

Cronbach, L.J., Gleser, C.G., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioural Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley & Sons.

Elstein, A., Shulman, L. &, Sprafka, S. (1978). *Medical Problem Solving*. Cambridge, MA: Harvard University Press.

DeWolfe, T. E., Jackson, L. A., & Winterberger, P. (1988). A comparison of moral reasoning and moral character in male and female incarcerated felons. *Sex-roles, 18,* 583-593.

DiMatteo, M. R., & DiNicola, D. D. (1982). *Achieving patient compliance*. New York: Pergamon.

Diseker, R. A., & Michielutte, R. (1981). An analysis of emapthy in medical students before and following clinical experience. *Journal of Medical Education, 56,* 1004-1009.

Engler, C. W., Saltman, C. A., Walker, M. L., & Wolf, F. M. (1981). Medical skills acquisition and retention of communication and interviewing skills. *Journal of Medical Education, 56,* 572-579.

Elstein, A., Shulman, L., & Sprafka, S. (1978). *Medical Problem Solving*. Cambridge, MA: Harvard University Press.

Evans, B. J., Stanley, R. O., Burrows, G. D., & Sweet, B. (1989). Lectures and skills workshops as teaching formats in a history-taking skills course for medical students. *Medical Education, 23,* 364-370.

Evans, B. J., Stanley, R. O., Mestrovic, R., & Rose, L. (1991). Effects of communication skills training on students' diagnostic efficiency. *Medical Education, 25,* 517-526.

Ford, J.D. (1979). Research on training counsellors and clinicians. *Review of Educational Research, 49,* 87-130.

Fox, M., Gibbs, M., & Auerbach, D. (1985). Age and gender dimensions of friendship. *Psychology of Women Quarterly, 9,* 489-501.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39,* 139-202.

Friedman, J. M. H. (1989). Structured interviews: The expert's vantage. In S. Wetzler & M. M. Katz (Eds.), *Contemporary approaches to psychological assessment: Vol 1. Clinical and experimental psychiatry* (pp. 83-98). New York: Brunner/Mazel.

Fyffe, A. E. & Oei, T. P. S. (1979). Influence of modelling and feedback provided by the supervisor in a microskills training program for beginning counsellors. *Journal of Clinical Psychology, 35,* 651-656.

Gallagher, M. S. & Hargie, O. D. W. (1989). An investigation into the validity of role play as a procedure for counsellor skill assessment. *British Journal of Guidance and Counselling, 17,* 155-165.

Gask, L., Goldberg, D., Lesser, A. L., & Millar, T. (1988). Improving the psychiatric skills of the general practice trainee: an evaluation of a group training course. *Medical Education, 22,* 132-138.

Gask, L., McGrath, G., Goldberg, D., & Millar, T. (1987). Improving the psychiatric skills of established general practitioners: Evaluation of group teaching. *Medical Education, 21,* 362-368.

Gelder, M., Gath, D., & Mayou, R. (1989). *Oxford textbook of psychiatry (2nd ed.).* Oxford: Oxford University Press.

Gercema, A. J. (1992). Het categoriseren van consultopgaven naar moeilijkheidsgraad: De Amsterdam Clinical Challenge Scale (ACCS) [Classifying consult tasks to degree of difficulty: The Amsterdam Clinical Challenge Scale (ACCS)]. *Bulletin Medisch Onderwijs, 11,* 14-21.

Ginsberg, G. L. (1985). Diagnosis and psychiatry: Examination of the psychiatric patient. In H. I. Kaplan & B. J. Sadock (Eds.), *Comprehensive textbook of psychiatry/IV* (pp. 482-494). Baltimore: Williams & Wilkins.

Glass, G., McGaw, B. & Smith, M.L. (1981). *Meta-analysis in social research.* Beverly Hills CA: Sage.

Goldberg, D. P., Steele, J. J., Smith, C., & Spivey, L. (1980). Training family doctors to recognise psychiatric illnesses with increased accuracy. *The Lancet,* 521-523.

Goldberg, D., & Huxley, P. (1980). *Mental illness in the community: The pathway to psychiatric care.* London: Tavistock Publications.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychology, 39,* 93-104.

Groot, A. D. de (1974). To what purpose, to what effect? Some problems of method and theory in the evaluation of higher education. In W. A. Verreck (Ed.), *Methodological problems in research and development in higher education* (pp. 16-44). Amsterdam: Swets & Zeitlinger.

Hersen, M. & Turner, S. M. (1985). *Diagnostic interviewing.* New York, Plenum Press.

Hill, C. E., Charles, D., & Reed, K. G. (1981). A longitudinal analysis of changes in counselling skills during doctoral training in counselling psychology. *Journal of Counselling Psychology, 28,* 428-436.

Hopkinson, K., Cox, A. & Rutter, M. (1981). Psychiatric interviewing techniques: III. Naturalistic study; eliciting feelings. *British Journal of Psychiatry, 138,* 406-415.

Ivey, A. (1971) *Microcounselling: Innovations in interviewing training.* Springfield, Il.: Thomas.

Jaynes, S., Charles, E., Kass, F., & Holzman, S. (1979). Clinical supervision of the initial interview: Effects on patient care. *American Journal of Psychiatry, 11,* 1454-1457.

Kolodner, J. L. (1983). Towards an understanding of the role of experience in the evolution from novice to expert. *International Journal of Man-Machine studies, 19,* 497-518.

Kraan, H. F., & Crijnen, A. A. M. (1987). *The Maastricht History and Advice Checklist: Studies of instrumental utility*. Thesis. Amsterdam: Velden van den Hezelaer.

Kraan, H. F., Crijnen, A. A. M., Vleuten, C. P. M. van der, & Imbos, T. (1992). Comparison of self and expert evaluation of medical interviewing skills: A multitrait-multimethod approach. *Proceedings of the conference on Research in Medical Education (RIME)* (pp.181-187).

Kraan, H. F., Crijnen, A. A. M., Vleuten, C. P. M. van der, & Van Thiel, J. (1990). The medical interview: How reliable is its evaluation by patients. *Int. Conference on Communication in Health Care*, Oxford.

Kraan, H. F., Crijnen, A. A. M., Vries, M. W. de, Zuidweg, J., Imbos, T., & Vleuten, C. P. M. van der (1990). To what extend are medical interviewing skills teachable? *Medical Teacher, 12,* 315-328.

Kurtz, P. D., Marshall, E. K. & Banspach, S. W. (1985). Interpersonal skill training research; a 12 year research and analysis. *Counsellor Education and Supervision,* 249-263.

Lang, G., & Molen, H. T. van der (1984). *Psychologische gespreksvoering*. Baarn: Nelissen B.V.

Lang, G., & Molen, H. T. van der (1992). *Methodiek van gesprekstraining*. Baarn: Nelissen B.V.

Lang, G., & Molen, H. T. van der with Trower, P. & Look, R. (1990). *Personal conversations: Roles and skills for counsellors*. London: Routledge.

Langsley, D. G., & Hollender, M. H. (1982). The definition of a psychiatrist. *American Journal of Psychiatry, 139,* 81-85.

Lieberman, S., Cobb, J. & Jackson, C.H. (1989). Studying the 'Grammar of Psychotherapy' course using a student and control population. *British Journal of Psychiatry, 155,* 842-845.

Lovett, L. M., Cox, A., & Abou-Saleh, M. (1990). Teaching psychiatric interview skills to medical students. *Medical Education, 24,* 243-250.

Lowyck, J. (1991). Vaardigheden: Inleiding op een thema. *Tijdschrift voor Hoger Onderwijs*, 2, 70-76.

Luyk, S. J. van, & Vleuten, C. P. M. van der (1992). A comparison of checklists and rating scales in performance-based testing. In Hart, I. R., Harden, R. M., & Marchais, J. des (Eds.), *Current developments in assessing clinical competence* (pp. 357-362). Ottawa: Can Heal Publications.

Luyk, S. J., Vleuten, C. P. M. van der, & Schelven, S. M. van (1990). The relationship between content and psychometric characteristics in performance-based testing. In Bender, W., Hiemstra, R. J., Scherpbier, A. J. J. A., & Zwierstra, R. P. (Eds.), *Teaching and assessing clinical competence.* Groningen: Boekwerk Publ.

MacKinnon, R. A., & Michels, R. (1971). *The psychiatric interview in clinical practice.* Toronto: Saunders.

Marmar, C. R. (1990). Psychotherapy process research: progress, dilemmas and future directions. *Journal of Consulting and Clinical Psychology, 3,* 265-272.

Marshall, E. K., Charping, J. W. & Bell, W. J. (1979). Interpersonal skills training: A review of the research. *Social work Research and Abstracts, 15,* 10-16.

Marteau, T. M., Humphrey, C., Matoon, G., Kidd, J., Lloyd, M. & Horder, J. (1991). Factors influencing the communication skills of first-year clinical medical students. *Medical Education, 25,* 127-134.

McGuire, P. (1982). Psychiatrists also need interview training (Comments). *British Journal of Psychiatry, 141,* 423-424.

McCready, J. R., & Waring, E. M. (1986). Interviewing skills in relation to psychiatric residency. *Canadian Journal of Psychiatry, 31,* 317-322.

Merluzzi, T. V. & Glass, C. R. (1981). Cognitive assessment. New York: guilford Press.

Mumford, E., Schlesinger, H., Cuerdon, T. & Scully, J. (1987). Ratings of videotaped simulated patients and four other methods of evaluating psychiatric clerkship. *American Journal of Psychiatry, 3,* 316-322.

Muzzin, L. J., Norman, G. R., Feightner, J. W., Tugwell, P., & Guyatt, G. (1983). Expertise in recall of clinical protocols in two specialty areas. *Proceedings of the 26nd Conference on Research in Medical Education (RIME)* (pp. 122-127). Washington DC.

Newble, D. I., & Jaeger, K. (1983). The effect of assessment and examinations on the learning of medical students. *Medical Education, 20,* 162-165.

Nicholi, A. M. (1978). History and mental status. In A. M. Nicholi, A. M. (Ed.), *Harvard guide to modern psychiatry (pp. 25-40).* Cambridge: Harvard University Press.

Norcini, J. J., Stillman, P. L., Regan, M. B., Sutnick, A. L., Williams, R. G., & Friedman, M. (1992, April). Scoring and standard setting with standardized patients. In B. Dawson-Saunders (Chair), *Standard setting for clinical competence examinations.* Symposium conducted at the annual meeting of the American Educational Research Association, San Fransisco.

Norman, G. R., Neufeld, V., Walsh, A., Woodward, C. & McConvey, G. (1985). Measuring physicians' performance by using simulated patients. *Journal of Medical Education, 60,* 925-934.

Norman, G. R. (1992). Authentic assessment: Lessons from the professions. Paper presented at the *Annual Meeting of the American Educational Research Association*, San Francisco, April 20-24.

Norman, G. R. (1994). Age, education, and expertise. In preparation.

Norman, G. R., Brooks, L. R., & Allen, S. W. (1989). Recall by expert medical practitioners and novices as a record of processing attention. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15*, 1166-1174.

Norman, G. R., Tugwell, P. & Feightner, J. W. (1982). A comparison of resident performance on real and simulated patients. *Journal of Medical Education, 57*, 708-715.

Nurcombe, B., & Fitzhenry-Coor, I. (1982). How do psychiatrists think? Clinical reasoning in the psychiatric interview: A research and education project. *Australian and New Zealand Journal of Psychiatry, 16*, 13-24

Nurcombe, B. & Gallagher, R. M. (1986). *The clinical process in psychiatry: Diagnosis and management planning*. New York: Cambridge University Press.

O'Toole, W. M. (1979). Effects of practice and some methodological considerations in training counselling interviewing skills. *Journal of Counselling Psychology, 26*, 419-426.

Patel, V. L., & Groen, G. J. (1986). Knowledge-based solution strategies in medical reasoning. *Cognitive Science, 10*, 91-116.

Patrick, J. P. (1992). *Training: Research and practice*. London: Academic.

Pendleton, D., Schofield, T., Tate, P., & Havelock, P. (1984). *The consultation: An approach to learning and teaching*. Oxford: Oxford University Press.

Peterson, V. L., & Peterson, M. J. (1959). Short term retention of individual verbal items. *Journal of Experimental Psychology, 58*, 193-198.

Pollock, D. C., Shanley, D. F., & Byrne, P. N. (1985). Psychiatric interviewing and clinical skills. *Canadian Journal of Psychiatry, 30*, 64-68.

Preven, D. W., Krajic Kachur, E., Kupfer, R. B., & Waters, J. A. (1986). Interviewing skills of first-year medical students. *Journal of Medical Education*, 61, 842-844.

Pritchard, M. J. (1981). Temporal reliability of a questionnaire measuring psychological responses to illness. *Journal of Psychosomatic Research, 25*, 63-66.

Reiser, D. E. (1984). The psychiatric interview. In: *Goldman, H. G. (ed.), Review of general psychiatry* (pp. 197-205). Los Altos, CA: Lange.

Rethans, J., & Boven, C. van (1987). Simulated patients in general practice: A different look at the consultation. *British Medical Journal, 294*, 809-812.

Rethans, J., Sturmans, F., Drop, R., & Vleuten, C. van der (1991a). A method for introducing standardized (simulated) patients into general practice consultations. *British Journal of General Practice, 41*, 94-96.

Rethans, J., Sturmans, F., Drop, R., & Vleuten, C. van der (1991b). Assessment of general practitioners by the use of standardized (simulated) patients. *British Journal of General Practice, 41,* 97-99.

Rethans, J., Sturmans, F., Drop, R., Vleuten, C. van der, & Hobus, P. (1991c). Does competence of physicians predict their performance? A direct comparison between examination setting and actual practice using undetected standardized (simulated) patients. *British Medical Journal, 303,* 1307-1380.

Rettie, C. S. (1992). Authentic assessment in the professions: An introduction. Presented at the Annual Meeting of the *American Educational Research Association*. San Francisco, april 20-24.

Reynolds, J. H. & Glaser, R. (1964). Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Experimental Psychology, 55,* 297-308.

Riccardo, V. M. & Kurtz, S. M. (1983). Methods for teaching and assessing communication in health care. *Communication in health care (appendix B).* Springfield, Ill.: Charles C. Thomas.

Rimaldi, J. A. (1961). The test of diagnostic skills. *Journal of Medical Education, 36,* 73-79.

Robinson, V. M. J., & Halliday, J. (1987). A critique of the microcounselling approach to problem understanding. *British Journal of Guidance and Counselling, 15,* 113-124.

Robinson, V. M. J. & Halliday, J. (1988). Relationship of counsellor reasoning and data collection to problem-analysis quality. *British Journal of Guidance and Counselling, 16,* 50-62.

Rosenthal, R. H., & Akiskal, H. S. (1985). Mental Status Examination. In M. Hersen & S. M. (Eds.), *Diagnostic interviewing* (pp. 25-52). New York: Plenum.

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

Rutter, M., Cox, A., Egert, S., Holbrook, D. & Everitt, B. (1981). Psychiatric interviewing techniques: IV. Experimental study: Four contrasting styles. *British Journal of Psychiatry, 138,* 456-465.

Salvendy, J. T. (1987). Supervision of the initial interview: A choice of methods. *Journal of Psychiatric Education, 11,* 121-126.

Sanson-fisher, R., Fairburn, S., & Maguire, P. (1981). Teaching skills in communication to medical students-a critical review of the methodolgy. *Medical Education, 15,* 33-37.

Sanson-Fisher, R., & Maguire, P. (1980). Should skills in communicating with patients be taught in medical schools? *The Lancet,* 523-526.

Sanson-Fisher, R., & Poole, A. D. (1980). Simulated patient and the assessment of medical students' interpersonal skills. *Medical Education, 14,* 249-253.

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On the origin of intermediate effects in clinical case recall. *Memory and Cognition, 21,* 338-351.

Schmidt, H. G., & Boshuizen, H. P. A. (1992). Encapsulation of biomedical knowledge. In D. A. Evans & V. L. Patel (Eds.). *Advanced models of cognition for medical training and practice.* New York, NY: Springer Verlag.

Schmidt, H. G., Norman, G. R. & Boshuizen, H. P. A. (1990). A cognitive perspective on Medical Expertise: Theory and implications. *Academic Medicine, 65,* 611-621.

Scott, N. C., Donnely, M. B., & Hess, J. W. (1975). Changes in interviewing styles of medical students. *Journal of Medical Education, 50,* 1124-1126.

Shea, S. C. (1988). *Psychiatric interviewing: The art of understanding.* Philadelphia: Saunders.

Shea, S. C., & Mezzich, J. E. (1988). Contemporary psychiatric interviewing: New directions for training. *Psychiatry, 15,* 385-397

Siassi, I. (1984). Psychiatric Interview and Mental Status Examination. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 259-275). New York: Pergamon.

Shatzer, J. H., Darosa, D., Colliver, J. A., & Barkmeier, L. (1993). Station-length requirements for reliable performance-based examination scores. *Academic Medicine, 68,* 224-229.

Shatzer, J. H., Wardrop, J. L., Bible, R. G., Williams, R. G., Carlson, M. T., & Witkovsky, M. T. (1992). Station-length requirements for SP exams: Their influence on generalizability estimates. In Hart, I.R., Harden, R.M. & des Marchais, J. (Eds.), *Current Developments in Assessing Clinical Competence.* Ottawa: Can-Heal Publications.

Shatzer, J. H., Wardrop, J. L., Williams, R. G., & Hatch, T. F. (1994). The generalizability of performance on different station length standardized patient cases. *Teaching and Learning in Medicine,* in press.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989) Generalizability theory. *American Psychologist, 6,* 922-932.

Stillman, P., Swanson, D., Regan, M.B. et al. (1991). Assessment of clinical skills of residents utilizing standardized patients: A follow-up study and recommendations for application. *Annals of Internal Medicine, 114,* 393401.

Stone, G. L. (1982). Evaluating the effectiveness of skills training programs. In Marshall, E. K. & Kurtz, P. D. & associates (Eds.), *Interpersonal helping skills* (pp. 447-467). San Francisco: Jossey-Bass publishers.

Streiner, D. L. (1985). Global rating scales. In V. R. Neufeld & G. R. Norman (Eds.) *Assessing clinical competence* (pp. 119-141). New York: Springer.

Sullivan, H. S. (1954). *The Psychiatric interview.* London: Tavistock Publication.

Swanson, D. B. (1992). Authentic assessment of clinical skills: A psychometric perspective. Paper presented at the *Annual Meeting of the American Educational Research Association.* San Francisco, april 20-24.

Tanner, L. A., & Silverman, G. (1981). A teacher's guide to teaching medical interviewing. *Medical Education, 15*, 100-105.

Thiel, J. van (1992). Assessment of medical interviewing skills: Generalizability of scores using the MAAS-R2. Paper presented at the *Fifth Ottowa International Conference of Clinical Competence,* Dundee, sept. 1-3.

Thiel, J. van, Kraan, H. F., & Vleuten, C. P. M. van der (1991). Reliability and feasability of measuring interviewing skills using the revised Maastricht History and Advice Checklist. *Medical Education, 25,* 224-229.

Thompson, M. G. G. (1979). *A residents's guide to psychiatric education.* New York: Plenum.

Turner, S. M., & Hersen, M. (1985). The interviewing process. In M. Hersen & S. M. Turner (Eds.), *Diagnostic interviewing* (pp. 3-23). New York: Plenum.

Vleuten, C. P. M. van der, & Swanson D. B. (1990). Assessment of clinical skills with standardized patients: state of the art. *Teaching and Learning in Medicine, 2*, 58-76.

Vleuten, C. P. M. van der, & Swanson D. B. (1990). Assessment of clinical skills with standardized patients: state of the art. *Teaching and Learning in Medicine, 2*, 58-76.

Whitehead, M. M., & Nokes, K. M. (1990). An examination of demographic variables, nurturance, and empathy among homosexual and heterosexual Big Brother/Big Sister volunteers. *Journal of Homosexuality, 19,* 89-101.

Appendix I

Handleiding beoordeling van het anamnestisch interview in de geestelijke gezondheidszorg

Deze handleiding betreft een instrument waarmee getrainde observatoren de anamnestische interviewvaardigheid van studenten en hulpverleners in de geestelijke gezondheidszorg kunnen meten. Het gaat daarbij om het afnemen van een anamnese bij patiënten met geestelijke gezondheidsproblemen.

Anamnestisch interviewen is een complexe vaardigheid waaraan verschillende aspecten onderscheiden kunnen worden. Ten eerste is het vermogen om relevante informatie te verzamelen belangrijk, het gaat hier om de *inhoud* van het interview. Ten tweede zijn de gespreksvaardigheden van de hulpverlener (HV) van belang, het gaat hier om het *proces* en de *vorm* van het interview. Het beoordelingssysteem bevat deze beide aspecten.

Deel 1 van deze handleiding bevat de beschrijving van de gespreksvaardigheden-items. De beschrijvingen zijn deels gedestilleerd uit het boek Psychologische Gespreksvoering van Lang en Van der Molen, 1984. Bij de ontwikkeling van deel 1 heeft Marie Snoek voorts een bijdrage geleverd. Deel 2 bevat de beschrijving van inhoudelijke items, en is mede geïnspireerd door verschillende versies van de Maastrichtse Anamnese- en Advies Scoringslijst (MAAS) (Kraan & Crijnen, 1987, Van Thiel, Van der Vleuten & Kraan, 1992). Tot slot bevat de handleiding een observatieformulier, waar de observatoren hun beoordelingen kunnen vastleggen.

Bij het observeren gaat het vooral om het gedrag van de hulpverlener.

Deel 1: Gespreksvaardigheden bij de anamnese

De beoordeling van de gespreksvaardigheden bestaat uit twee onderdelen. Tijdens het observeren van het interview wordt elke interventie geturfd op zogenaamde scoringstabellen. Daarna wordt, terugkijkend op het gesprek, een totaal-oordeel per vaardigheid gegeven.

Turven op de scoringsbalken.
Het turven op de scoringsbalken dient in principe tijdens het zien en horen van het interview te gebeuren. Probeer zoveel mogelijk interventies te turven, d.w.z.:
- Bepaal onder welke vaardigheid 1 t/m 10 een interventie valt.
- Turf op de balk alle interventies die onder deze vaardigheid vallen; zet het streepje globaal onder 1, 2, 3, 4 of 5 overeenkomstig de mate waarin de interventie correct resp. incorrect is volgens de criteria van kwaliteit en functionaliteit van de betreffende vaardigheid.
- Houd globaal bij aan welke criteria van een bepaalde vaardigheid wel of niet wordt voldaan door bij de betreffende criteria een + of - te noteren.

Met betrekking tot de criteria die een indicatie vormen voor de functionaliteit van de interventies vergt dit turven nog enige toelichting. Het criterium 'waar nodig' dient als volgt te worden beoordeeld:
+ een vaardigheid wordt toegepast en is op dat moment ook geïndiceerd;
– een vaardigheid wordt toegepast, maar is op dat moment niet geïndiceerd.

Voor het criterium 'juist gedoseerd' geldt het volgende:
+ de mate waarin de vaardigheid wordt toegepast bevordert een soepel verloop van het gesprek;
– de mate waarin de vaardigheid wordt toegepast belemmert een soepel verloop van het gesprek (d.w.z. te veelvuldig of spaarzaam toegepast).

Beoordelen per vaardigheid.
Op grond van het turven wordt per vaardigheid een totaal oordeel gegeven op de 1-2-3-4-5-schaal over de kwaliteit en functionaliteit van de interventies, waarbij 1=ruim onvoldoende, 2=onvoldoende, 3=middelmatig, 4=voldoende, 5=goed.

Hieronder worden een aantal vaardigheden beschreven die nodig zijn om een probleemverhelderd of anamnestisch gesprek te voeren. Van elke vaardigheid wordt aangegeven wat het doel is, wanneer toepassing ervan is geïndiceerd is en wat de criteria voor juist gebruik zijn.

1. Aandachtgevend gedrag

Aandachtgevend gedrag behoort tot de nonselectieve luistervaardigheden en bestaat uit verbale en nonverbale gedragingen zoals gelaatsuitdrukkingen, oogcontact, hoofdknikken en ondersteunende handgebaren, 'hummen', korte stiltes (i.t.t. stiltes die voortvloeien uit onzekerheid of afwezigheid), en kleine tussenwerpsels. *Doel* ervan is een situatie te creëren waarin de CL gestimuleerd wordt vrijuit en in alle rust over zijn problemen te praten. De nonverbale en verbale aanmoedigingen mogen niet zo vaak voorkomen of onrustig zijn dat zij de CL afleiden van zijn verhaal.

Criteria voor aandachtgevend gedrag zijn:
- *ontspannen, toegenegen lichaamshouding;*
- *geïnteresseerde gelaatsuitdrukking;*
- *stimulerend oogcontact;*
- *kleine nonverbale aanmoedigingen zoals hoofdknikken en ondersteunende handgebaren;*
- *kleine verbale aanmoedigingen zoals hummen en kleine tussenwerpsels;*
- *korte, aandachtige stiltes;*
- *waar nodig gegeven te worden (om de CL te stimuleren verder te praten)*
- *juist gedoseerd te worden (d.w.z. stimulerend, maar niet storend)*

2. Vragen stellen

Vragen stellen behoort tot de selectieve vaardigheden. Een belangrijk onderscheid is dat tussen gesloten en open vragen. Een *open vraag* sluit geen antwoordcategorie uit en laat de CL een grote vrijheid in de formulering van zijn antwoord. Open vragen hebben tot *doel* de CL ruimte te geven om vanuit zijn eigen referentiekader te praten en zelf richting en inhoud van het gesprek te bepalen. Open vragen kunnen variëren in de hoeveelheid ruimte die een CL wordt geboden. Vergelijk: "Waar wilt u over praten?" en "Hoe ging het afgelopen week met de collega's op uw werk?" De ruimte die een open vraag biedt, dient zodanig gekozen dat het de CL stimuleert tot (verdere) exploratie. *Scoor een open vraag alleen onder 'open vragen' wanneer het een openingsvraag betreft of een vraag over een nieuw onderwerp. Verdere open vragen binnen het reeds aangesneden onderwerp zijn 'doorvragen' en worden gescoord onder concretiseringen.* Het stellen van open vragen is *geïndiceerd* wanneer de HV de CL exploratief tegemoet treedt, dus met name tijdens de probleemverheldering, exploratieve periodes tijdens de speciële anamnese, en exploratieve periodes tijdens de biografische en sociale anamnese. De vragen mogen de informatiestroom of de gedachtengang van de CL echter niet te zeer onderbreken. Open vragen naar een nieuw onderwerp zijn geïndiceerd wanneer een onderwerp naar het idee van de HV en de CL voldoende is geëxploreerd.

Open vragen dienen:
- *duidelijk geformuleerd te worden;*
- *slechts naar een onderwerp tegelijk te verwijzen;*
- *er mogen geen vooronderstellingen in doorklinken;*
- *juist gedoseerd te zijn (d.w.z. stimuleren de CL tot verdere exploratie storen de CL niet in zijn verhaal)*

Een *gesloten vraag* beperkt de antwoordmogelijkheden van de CL. Het antwoord wordt voor een groot deel bepaald door de inhoud van de vraag. De CL kan in het algemeen volstaan met het beamen danwel ontkennen. Nadelen van gesloten vragen zijn:
a) dat de CL wordt beperkt in zijn uitingen;
b) dat ze vaak suggestief zijn omdat ze voortkomen uit het referentiekader van de HV, en
c) dat de CL steeds kortere antwoorden gaat geven zodat het gesprek het karakter van een kruisverhoor gaat krijgen.

In de probleemverhelderende fase is een spaarzaam gebruik van gesloten vragen dan ook aan te raden. In de speciële anamnese en psychiatrische anamnese i.e.z. is het vaker geïndiceerd gesloten vragen te stellen. Gesloten vragen zijn *geïndiceerd* wanneer de HV iets specifieks wil weten (eenvoudige of feitelijke informatie) of als de CL weerstanden heeft om een bepaald onderwerp te bespreken.

Gesloten vragen dienen:
- *duidelijk geformuleerd te zijn;*
- *slechts naar één onderwerp tegelijk te verwijzen;*
- *zowel in formulering als toon zo min mogelijk suggestief te zijn;*
- *spaarzaam gedoseerd te zijn (d.w.z. alleen wanneer de HV iets specifieks wil weten).*

3. Concretiseren (doorvragen)

Concretiseren of doorvragen behoort tot de selectieve luistervaardigheden en wil zeggen: de CL zo nauwkeurig en precies mogelijk over zijn problemen laten vertellen. De HV stelt vragen als: "Wat gebeurde er precies?" "Wie zijn 'ze'?" "Wanneer voelde u zich zo?" Concretiseren kan plaatsvinden m.b.t. verschillende aspecten van het probleem: bijv. de situatie, gedrag, gedachten, gevoelens. Concretiseringen dienen bij voorkeur plaats te vinden in de vorm van open vragen. *Doel* van concretiseren is om algemene, onpersoonlijke en impliciete uitspraken en verhalen van de CL te verduidelijken om zo een scherper en gedifferentieerder inzicht in de problematiek te krijgen. Concretiseringen zijn *geïndiceerd* wanneer de CL vage uitspraken doet en verduidelijking hiervan nodig is om een redelijk precieze voorstelling van de problemen te maken. Concretiseringen zijn nodig om een onderwerp dat de CL al naar voren heeft gebracht verder te verduidelijken. De HV mag echter niet zover gaan dat zij de CL laat uitweiden over irrelevante details of dat het gesprek het karakter van een kruisverhoor krijgt.

Concretiseringen dienen:
- *bij voorkeur open gesteld te worden;*
- *aan te sluiten op wat de CL zojuist heeft verteld;*
- *zo nodig op verschillende aspecten van de situatie betrekking te hebben;*
- *waar nodig gesteld te worden (d.w.z. bij vage uitspraken);*
- *juist gedoseerd te zijn (d.w.z. niet te gedetailleerd).*

4. Parafraseren van inhoud

Parafraseren behoort tot de selectieve vaardigheden. Het betekent kort en in eigen woorden weergeven van de belangrijkste inhoudelijke informatie die de CL juist tevoren heeft gegeven. Het parafraseren kan verschillende *doelen* dienen:
a) tonen van begrip
b) nagaan of je de CL hebt begrepen
c) de informatie van CL preciezer weergeven
Wil een HV laten merken dat hij de CL begrijpt, dan dient de parafrase op bevestigende toon te worden uitgesproken. Wil de HV nagaan òf hij de CL heeft begrepen, dan dient CL de mogelijkheid te krijgen tot correctie en dient de parafrase op veronderstellende toon te worden uitgesproken. Parafrasen zijn *geïndiceerd* wanneer CL iets vertelt wat blijkbaar belangrijk voor hem is door dit met grote nadruk of juist aarzeling uit te spreken. Parafrasen dienen de CL te stimuleren door te praten, maar zij mogen niet zovaak voorkomen dat zij de gedachtengang of de informatiestroom van de CL onderbreken.

Parafrasen dienen:
- *kort en duidelijk de essentie weer te geven van hetgeen de CL het laatst gezegd heeft;*
- *neutraal te zijn (d.w.z er mogen geen vooronderstellingen in doorklinken)*
- *met de juiste intonatie te worden uitgesproken (bevestigend of veronderstellend);*

- *waar nodig gegeven te worden (d.w.z. geen relevante zaken blijven liggen);*
- *juist gedoseerd te zijn (d.w.z. er wordt voldoende begrip getoond maar het verhaal van CL wordt niet onderbroken).*

5. Reflecteren van gevoel

Reflecteren van gevoel behoort tot de selectieve luistervaardigheden en betekent het kort in eigen woorden weergeven van het belangrijkste gevoel (of gevoelens) dat in de woorden van de CL doorklinkt. De gevoelens kunnen betrekking hebben op personen (de CL zelf, anderen, de HV) of op gebeurtenissen, zij kunnen in het verleden gespeeld hebben of actueel zijn, zij kunnen enkelvoudig zijn (bijv. verdrietig) of meervoudig (verdrietig en opgelucht tegelijk), zij kunnen verbaal in al dan niet bedekte termen worden geuit, maar vaker worden zij nonverbaal geuit (lichaamshouding, gelaatsuitdrukking, stemvolume, intonatie, etc.).

Het reflecteren van gevoel dient dezelfde *doelen* als het parafraseren van de inhoud (tonen van begrip, nagaan of je de CL begrijpt, preciezer weergeven wat de CL bedoelt). Echter, door een parafrase toont de HV dat hij probeert te begrijpen wat de CL zegt, door een gevoelsreflectie toont de HV hoe de CL zich voelt of voelde. Meer dan een parafrase dient een gevoelsreflectie om ervoor te zorgen dat een CL zich begrepen, geaccepteerd en veilig voelt zodat hij gestimuleerd wordt zijn gevoelens te uiten en zich er meer bewust van te worden. Als zodanig draagt het reflecteren van gevoel bij uitstek bij tot de probleemverheldering. Nuanceerende empathische reacties, d.w.z. reflecties van onderliggende gevoelens, de verborgen gevoelstoon, gevoelens 'tussen de regels door', worden ook gescoord onder de vaardigheid 'gevoelsreflecties'. Omdat de HV hier interpreteert, dienen deze reflecties bij uitstek tentatief gebracht te worden.

Wanneer *geïndiceerd?* Gevoelsreflecties dienen minimaal zo vaak gegeven te worden dat geen belangrijke gevoelens blijven liggen en dat er voldoende begrip wordt getoond.

Reflecties van verbaal en nonverbaal geuite gevoelens dienen:
- *kort en duidelijk de essentie weer te geven van de gevoelens die uit de woorden of het gedrag van de CL blijken;*
- *in eigen woorden gesteld te zijn;*
- *inhoudelijk correct de door de CL geuite gevoelens te benoemen;*
- *qua intensiteit overeen te stemmen met de door de CL geuite gevoelens*
- *geen waarde-oordelen te bevatten;*
- *zonodig de context (wanneer, t.a.v. wie of wat) te benoemen;*
- *de juiste intonatie te hebben (d.w.z. bevestigend bij duidelijke gevoelens en veronderstellend wanneer de HV nagaat of hij de CL heeft begrepen);*
- *waar nodig gegeven te worden (d.w.z. geen belangrijke gevoelens blijven liggen);*
- *juist gedoseerd te zijn (d.w.z. er wordt voldoende begrip getoond, het verhaal wordt niet onderbroken).*

6. Samenvatten

Samenvatten behoort tot de selectieve vaardigheden. Een samenvatting is een in eigen woorden uitgesproken weergave van de belangrijkste elementen van minimaal 3 uitspraken van de CL. Het verschil tussen samenvatten en parafraseren en reflecteren is hoofdzakelijk daarin gelegen dat in een samenvatting uitingen van de CL over een langere tijdsperiode worden weergegeven. Een samenvatting kan de volgende *doelen* dienen:
a) nagaan of je de CL goed begrepen hebt;

b) de CL stimuleren tot verdere exploratie van zijn gedachten of gevoelens;
c) structuur aanbrengen in het gesprek door de hoofdzaken te ordenen. Samenvattingen zijn *geïndiceerd*:
- na een verwarde woordenstroom;
- als de CL klaarblijkelijk alles verteld heeft wat voor hem van belang is;
- aan het einde van de bespreking van een bepaald onderwerp (bijv. een fase van de anamnese), bij de overgang naar een volgend onderwerp;
- aan het einde van een gesprek en aan het begin van een vervolggesprek.

Samenvattingen mogen echter niet het verhaal van de CL storen of voortkomen uit onzekerheid of verlegenheid van de HV.

Samenvattingen dienen:
- *kort en duidelijk de hoofdzaken te ordenen van wat de CL heeft gezegd;*
- *inhoudelijk juist te zijn;*
- *door intonatie en pauze de CL de gelegenheid te geven op de samenvatting te reageren;*
- *op het juiste moment gegeven te worden.*

7. Geven van informatie

Het geven van informatie is een regulerende vaardigheid. Een HV dient op verschillende momenten in een gesprek informatie te geven aan de CL. Informatie wordt verstrekt t.b.v. de CL. Noodzakelijk is dat de CL de informatie kan begrijpen en onthouden. Informatie wordt beter begrepen als ze concreet is, in begrijpelijke taal (geen vaktermen en moeilijke woorden) en aangepast aan het begripsvermogen van de CL. Informatie spreekt meer aan wanneer ze aansluit bij de belevingswereld van de CL, bijv. door voorbeelden uit het verhaal van de CL te gebruiken. Informatie wordt beter onthouden wanneer ze in kleine hoeveelheden wordt aangeboden (2 of 3 zinnen). Het geven van informatie is *geïndiceerd*:
- aan het begin van het gesprek (HV stelt zich voor, geeft doel en werkwijze aan, noemt tijdsduur, verwijst naar informatie die hij reeds heeft van CL, legt gebruik cassetterecorder of iets dergelijks uit);
- bij het begin van een nieuwe fase van de anamnese kondigt de HV de fase aan, legt het doel ervan uit en de werkwijze;
- bij vragen van de CL over de klachten en de behandeling[1].
- aan het eind van het gesprek (HV sluit gesprek af, geeft informatie over de verdere voortgang, noemt afspraken).

Informatie die wordt gegeven dient:
- *duidelijk en concreet te zijn;*
- *in begrijpelijk taal, en aangepast aan het begripsvermogen van de CL;*
- *zo mogelijk aan te sluiten bij de belevingswereld van de CL;*
- *in kleine hoeveelheden te worden aangeboden;*
- *waar nodig gegeven te worden (d.w.z. bij begin, bij overgangen, bij vragen CL, bij het eind);*

[1] Indien de CL vragen stelt over vragen van de HV (bijv. "Waarom wilt u dit weten?"), wordt het antwoord van de HV hierop gescoord onder item 9 (hardop denken), het gaat dan immers om het verduidelijken van het referentiekader van de HV.

- *juist gedoseerd te zijn (CL mag niet overspoeld worden met informatie het gaat immers om zijn verhaal, maar dient wel voldoende geïnformeerd te worden over de gang van zaken).*

8. Vragen om feedback

Met vragen om feedback wordt bedoeld: het vragen om reacties van de CL nadat de HV informatie heeft gegeven.
Vragen om feedback hebben verschillende *doelen*:
a) onderzoeken of de CL de informatie gehoord en begrepen heeft;
b) nagaan hoe de CL over de informatie denkt (is CL het met de informatie eens, zitten HV en CL op dezelfde golflengte);
c) het bevorderen van samenwerking.
Vragen om feedback kunnen gesloten of open gesteld worden; gesloten indien de HV om akkoord wil vragen (bij een contract) of op begrip wil controleren. Open indien HV geïnteresseerd is in CL's mening, eigen gedachte of reacties op hetgeen HV gezegd heeft. Beide vormen kunnen in een gesprek voorkomen. Als de HV alleen gesloten feedback vragen stelt, kan het gebruik van de vaardigheid een plichtmatig karakter krijgen. Als de HV alleen open feedback vragen stelt, kan dit de zitting vertragen en de efficiëntie van het gesprek verminderen. De HV heeft een geïnteresseerde houding t.a.v. feedback van de CL en laat merken deze feedback belangrijk te vinden. Vragen om feedback is *geïndiceerd* nadat de HV informatie heeft gegeven. Vragen om feedback dient in ieder geval te gebeuren bij de opening van het gesprek (contract!), wanneer de HV informatie over de klacht, of advies geeft, wanneer de HV reageert op vragen van de CL en aan het eind van het gesprek. Ook dient de HV passend te reageren op spontane feedback
van CL.

Vragen om feedback dienen:
- *gesloten of open gesteld te worden, afhankelijk van het doel van de vraag;*
- *blijk te geven van werkelijke interesse van HV in feedback van CL;*
- *waar nodig gesteld te worden (d.w.z. bij opening, bij geven van informatie, aan het eind van het gesprek en reacties op spontane feedback van CL);*
- *juist gedoseerd te zijn (d.w.z. niet bij elke zin om feedback vragen maar ook niet teveel informatie tegelijkertijd geven zonder tussentijds om feedback te vragen).*

9. Hardop denken

De HV spreekt haar gedachten hardop uit. Niet alleen de uiteindelijke conclusies, maar ook de observaties, gedachten en overwegingen van de HV worden aan de CL meegedeeld. Hardop denken heeft verschillende *doelen*: het bevordert de openheid en duidelijkheid van de kant van de HV, het werkt angstreducerend, het stimuleert de CL zelf tot hardop denken en het bevordert de samenwerkingsrelatie. M.b.v. hardop denken kan bovendien informatie worden uitgelokt waarmee het zich ontwikkelende begrip van de HV van het probleem wordt getoetst.
Hardop denken is *geïndiceerd*:
- wanneer de HV een vraag over een nieuw onderwerp wil stellen of een vraag of opmerking uit haar eigen referentiekader wil maken. De HV maakt de vraag **inzichtelijk** (het aankondigen van een nieuwe fase van de anamnese wordt bij item 8 (geven van informatie) gescoord);
- wanneer de HV gedachten heeft die haar blokkeren in het luisteren naar de CL;
- ter bevordering van de samenwerkingsrelatie, om de CL een actievere rol te laten spelen.

Hardop denken dient:
- *open en duidelijk te gebeuren;*
- *CI inzicht te geven in gedachten en overwegingen van de HV;*
- *waar nodig te gebeuren (d.w.z. wanneer de HV blokkeert, bij het inzichtelijk maken van vragen uit zijn referentiekader, voor het bevorderen van de samenwerking).*

10. Structureren

Structureren is een regulerende vaardigheid[2]. Met structureren wordt bedoeld het ordenen en leiden van het gesprek. Structurerende reacties hebben tot *doel*: de lijn van het gesprek vast te houden, het gesprek te ordenen en het gesprek efficiënt te laten verlopen (d.w.z. binnen de afgesproken tijd zoveel mogelijk van de doelen van de HV te bereiken). De HV kan structureren door: terug te komen op eerder (half) besproken onderwerpen, de CL af te kappen wanneer hij zijpaden inslaat en terug te brengen op de vraag of het onderwerp, de CL niet te laten doorpraten over onderwerpen waarvoor het belang, volgens de observator, niet duidelijk is.

De structurerende reacties dienen duidelijk te zijn voor de CL, d.w.z. de HV kapt bijv. zodanig af dat de CL begrijpt waarom hij afgekapt wordt en zich niet bestraft voelt. De HV dient flexibel te zijn in haar structureringen, d.w.z. niet te rigide vast te houden aan een van te voren opgesteld programma en soepel in te kunnen spelen op wijzigingen die door de CL impliciet dan wel expliciet naar voren worden gebracht. De HV dient een goede balans te hebben tussen volgende, exploratieve reacties en structurerende, directieve reacties. Structurerende reacties komen voort uit het referentiekader van de HV. Het is belangrijk dat de CL dat referentiekader kent.

Structurerende reacties zijn *geïndiceerd* wanneer:
- de CL en/of HV afdwaalt van het onderwerp;
- de CL uitwijdt over irrelevante details;
- de voor het gesprek gestelde doelen niet gehaald dreigen te worden (deze doelen dienen echter wel realistisch te zijn).

Structureringen dienen:
- *het interview te ordenen en sturen;*
- *duidelijk en inzichtelijk te zijn voor de CL;*
- *waar nodig gegeven te worden (d.w.z. wanneer men afdwaalt, of doelen niet gehaald dreigen te worden;*
- *juist gedoseerd te zijn (d.w.z. niet rigide, maar flexibel, maar ook niet te weinig waardoor de tijdsplanning in het gedrang komt).*

Totaaloordeel vaardigheden

Geef tot slot een subjectief eindoordeel over de kwaliteit en functionaliteit van de gespreksvaardigheden van de HV, in de vorm van een rapportcijfer (1-10).

---

[2] De vaardigheid structureren is een samengestelde vaardigheid. D.m.v. bijv. samenvattingen, hardop denken, informatie geven, kan de HV structureren. Deze reacties kunnen dus dubbel gescoord worden.

Deel 2: Handleiding beoordeling inhoud van de anamnese

Naast de beoordeling van de gespreks**vaardigheden** van de hulpverlener die een anamnestisch interview houdt, dient de **inhoud** van het anamnestisch interview beoordeeld te worden. De onderwerpen die de HV aansnijdt dienen inhoudelijk relevant te zijn voor het doel van het interview; een diagnose stellen in DSM-III termen, een beeld krijgen van de achtergronden en oorzaken van de problemen (etiologie), en het beleid te bepalen m.b.t. verdere diagnostiek en/of behandeling. Het anamnestisch interview in de GGZ wordt ingedeeld in 4 fasen:

a) Probleemverhelderen en speciële anamnese
b) Psychiatrische anamnese
c) Biografische anamnese
d) Sociale anamnese

a) Tijdens de probleemverheldering wordt het eerste contact met de cliënt gelegd. Door middel van open uitnodigingen probeert de HV een beeld te krijgen van hoe de CL zijn klachten omschrijft en beleeft. Het referentiekader van de CL staat hier centraal. In de speciële anamnese wordt de hoofdklacht verder uitgediept. De HV stelt hierbij vragen uit haar eigen referentiekader. Het probleemoplossend denken van de HV wordt op gang gebracht. De HV genereert hypothesen over de diagnose o.g.v. verbale en nonverbale informatie en toetst die aan de hand van nieuwe vragen.
b) Tijdens de psychiatrische anamnese wordt psychiatrische symptomatologie nagevraagd. Ook hier staat het referentiekader van de HV centraal. Een aantal verschijnselen betreffende CL's conditie, zoals oriëntatie en waarneming, worden onderzocht. Deels worden deze verschijnselen nagevraagd, deels ook worden ze geobserveerd. Fase a) en b) kunnen elkaar deels overlappen.
c) Tijdens de biografische anamnese wordt de levensgeschiedenis van de CL in kaart gebracht. De HV tracht een beeld te krijgen van de lichamelijke, psychische, sexuele en sociale ontwikkeling. Doel hiervan is het huidige disfunctioneren te begrijpen in relatie tot etiologische factoren.
d) In de sociale anamnese probeert de HV een beeld te krijgen van de aard en omvang van het sociale netwerk van de CL, en zijn maatschappelijke situatie. Doel is erachter te komen hoe gezond de CL sociaal en maatschappelijk functioneert en hoeveel steun hij heeft in zijn omgeving.

Gezien de beperkte tijd die de HV voor het anamnestisch interview heeft in testsituaties (een half uur tot een uur) is het ondoenlijk voor de HV alle fasen uitgebreid te behandelen. Het beoordelingsinstrument van het interview is hier dan ook op aangepast.

Bij de beoordeling van de inhoud is het *niet* de bedoeling dat ook de fasering van het interview wordt beoordeeld, het gaat er puur om of en hoe uitgebreid bepaalde onderwerpen door de HV behandeld zijn. De volgorde waarin de items besproken worden doet er dus hier niet toe. De fasering van het interview is een vaardigheid die gescoord wordt onder de gespreksvaardigheden 'geven van informatie' en 'structureren'. De beoordeling van de inhoud van de anamnese bestaat uit twee gedeelten. Tijdens het observeren van het interview houdt de beoordelaar bij in hoeverre elk item door de HV behandeld wordt. Wanneer een onderdeel voldoende behandeld is, maakt de beoordelaar het betreffende rondje (0) op de beoordelingslijst zwart. Na afloop van het interview wordt een totaal oordeel gegeven over de mate en kwaliteit van behandeling van elk item door de HV, waarbij de punten 1 t/m 5 hieronder beschreven zijn.

11. Probleemverheldering

De HV exploreert het aanmeldingsprobleem van de CL, zodanig dat beiden een duidelijk en concreet beeld krijgen van het probleem, zoals de CL het presenteert en formuleert. Dit houdt in dat de HV vraagt hoe het probleem eruit ziet, wanneer, waar, en bij wie het zich vooral voordoet, en hoe CL het probleem zelf beleeft. Het referentiekader van de CL staat hierbij centraal.

1. *Het aanmeldingsprobleem wordt niet verhelderd.*
2. *Het aanmeldingsprobleem wordt nauwelijks verhelderd, blijft erg vaag.*
3. *Het aanmeldingsprobleem wordt gedeeltelijk verhelderd.*
4. *Het aanmeldingsprobleem wordt duidelijk.*
5. *Het aanmeldingsprobleem wordt uitstekend verhelderd.*

12. Exploratie van (mogelijke) andere klachten of problemen.

De HV vraagt na of er nog andere problemen zijn, en gaat hier verder op in. Als CL spontaan andere problemen naar voren brengt, gaat de HV hier ook verder op in.

1. *HV vraagt niet naar andere problemen en gaat ook niet in[3] op 'hints' van CL. De HV heeft alleen maar aandacht voor dat ene probleem.*
2. *HV gaat nauwelijks in op andere klachten of problemen.*
3. *HV vraagt kort naar andere problemen of klachten, zonder hier verder uitgebreid op in te gaan, of vraagt er niet naar maar gaat wel op CL in, wanneer hij andere problemen spontaan naar voren brengt.*
4. *HV vraagt naar andere problemen èn gaat hier verder op in.*
5. *HV vraagt uitgebreid naar andere problemen en gaat hier uitvoerig op in. HV is sensitief voor 'hints' van CL m.b.t. andere problemen.*

13. Aanleiding om hulp te zoeken en hulpvraag

De HV vraagt wat de aanleiding was dat de CL *juist nu* hulp vraagt voor zijn problemen. Deze vraag kan ook meer zicht geven op de lijdensdruk van dit moment. De HV gaat tevens na wat de hulpvraag van de CL is. HV vraagt naar wensen en verwachtingen ten aanzien van mogelijke hulp. Zowel aanleiding als hulpvraag zijn van belang bij het uitstippelen van het verdere diagnostiek- en behandelbeleid bij de CL.

1. *HV vraagt niet naar de aanleiding en niet naar de hulpvraag. Als CL hier spontaan iets over zegt gaat hij er niet verder op in.*
2. *HV vraagt summier naar of gaat summier in op of de aanleiding of de hulpvraag.*
3. *HV vraagt summier naar of gaat summier in op èn de aanleiding èn hulpvraag, of op één van beide onderwerpen uitgebreider.*
4. *HV vraagt summier naar of gaat summier in op aanleiding en uitgebreid op hulpvraag, of andersom.*

---

[3] Met "gaat in op" wordt steeds bedoeld dat de HV d.m.v. aandachtgevend gedrag, parafraseren, reflecteren van gevoel, concretiseren of samenvatten aan de CL laat merken dat hij hetgeen de CL hier spontaan over meedeeld gehoord heeft, en/of stimuleert de HV de CL om hier meer over te zeggen.

5. *HV vraagt naar de aanleiding en de hulpvraag, en krijgt van beide onderwerpen een goed beeld.*

14. Uitdiepen van de hoofdklacht

Uitdiepen van de hoofdklacht gebeurt vanuit het referentiekader van de HV. Deze informatie wordt ingewonnen in het kader van het probleemoplossingsproces, dat wil zeggen het genereren en toetsen van hypothesen door de HV. De HV stelt verdere vragen over de aard en intensiteit van de hoofdklacht, en vraagt daarbij naar symptomen die het probleem begeleiden (bijvoorbeeld: aan een CL die klaagt over somberheid vraagt de HV of hij ook is afgevallen de laatste tijd. Of aan een CL die de straat niet op durft vraagt de HV welke lichamelijke gevoelens hij heeft als hij op straat loopt. Aan een CL die last heeft van vreetbuien vraagt de HV of hij ook laxantia gebruikt).

1. *HV stelt geen uitdiepende vragen over de klacht, gaat niet verder dan datgene wat CL zelf aangeeft over de klacht.*
2. *HV stelt nauwelijks uitdiepende vragen over aard, intensiteit en begeleidende symptomen.*
3. *De hoofdklacht wordt gedeeltelijk uitgediept door de HV.*
4. *De HV diept de hoofdklacht voldoende uit.*
5. *HV diept de hoofdklacht op uitstekende wijze uit, vraagt door op de aard en intensiteit van de hoofdklacht vanuit zijn eigen referentiekader, en vraagt naar (bijna) alle belangrijke begeleidende symptomen.*

15. Ontstaansfactoren

HV vraagt hoe de klacht ontstaan is en welke omstandigheden en gebeurtenissen een rol speelden bij het ontstaan van de klacht. Daarnaast vraagt de HV welke ideeën CL heeft over de oorzaken van de klacht. Met deze laatste vraag krijgt de HV tevens een idee van het ziekteinzicht en oordelingsvermogen van CL. Dit is zowel van belang bij het stellen van de diagnose (bijv. het ontbreken van een kritisch oordelingsvermogen bij paranoïde toestandsbeeld) als bij de beeldvorming van de persoonlijkheid van CL en behandelingsmogelijkheden.

1. *HV stelt geen vragen over het ontstaan, en gaat niet in op ongevraagde mededelingen hierover van CL.*
2. *HV krijgt een zeer onvolledig beeld van de onstaansfactoren, door of summier te vragen naar ontstaan, of summier naar eigen ideeën van CL over oorzaken, of summier in te gaan op spontane mededelingen van CL over ontstaan.*
3. *HV krijgt een matig beeld van de ontstaansfactoren, door of te vragen naar ontstaan, of naar eigen ideeën van CL over oorzaken, of in te gaan op spontane mededelingen van CL over ontstaan.*
4. *HV vraagt naar hoe de klachten ontstaan zijn en vraagt naar of gaat in op eigen ideeën van CL over oorzaken.*
5. *HV vraagt uitgebreid naar hoe de klachten ontstaan zijn, welke factoren ermee samenhingen, en welke ideeën CL over oorzaken heeft.*

16. Verloop

HV gaat na hoelang de klacht duurt, en hoe de klacht zich vanaf het moment van ontstaan heeft ontwikkeld tot nu. HV gaat na of de klacht in bepaalde periodes verergerd, verminderd of afwezig was en welke factoren hiermee samenhingen.

1. *HV vraagt niet naar verloop, en gaat niet in op spontane mededelingen hierover.*
2. *HV vraagt summier en zeer onvolledig naar verloop, of gaat summier in op spontane mededelingen hierover.*
3. *HV vraagt naar en gaat in op verloop, doch is duidelijk onvolledig, door periodes over te slaan en niet te zoeken naar factoren die samenhingen met verergering of vermindering.*
4. *HV vraagt naar verloop en gaat deels na welke factoren samenhingen met vermindering of verergering. HV slaat geen grote periodes over.*
5. *HV vraagt uitgebreid naar verloop, gaat na wanneer klachten verminderden of verergerden en wat hiermee samenhing.*

17. Invloed op dagelijks leven

Dit item betreft de gevolgen van de klacht op het dagelijks leven van CL. Het gaat hierbij om de gevolgen op het *gedrag* van CL en om de *reactie* van de omgeving op de klacht. De mogelijke functie van de klacht kan hierdoor duidelijk worden.

1. *HV vraagt niet naar en gaat niet in op de invloed van de klacht op het leven van CL.*
2. *HV vraagt summier naar de invloed van de klacht op het leven van CL, of gaat summier in op spontane mededelingen hierover door CL.*
3. *HV krijgt een matig beeld van de invloed van de klacht op het leven van CL, door òf te vragen naar de invloed, òf in te gaan op de invloed van de klacht op het leven van CL.*
4. *HV vraagt naar en gaat in op de invloed van de klacht op het leven van CL.*
5. *HV gaat op uitstekende wijze na wat de invloed van de klacht is op het leven van CL.*

18. Behandeling en zelfhulp

Met dit item wordt nagegaan of Cl eerder behandeling heeft gehad voor deze klachten, het gaat hierbij om behandeling in de breedste zin des woords: medische, paramedische, psychiatrische en psychologische behandeling. Het gebruik van medicatie voor de klachten, en alcohol of drugs, valt eveneens onder dit item. Ook wordt gevraagd wat CL zelf al heeft geprobeerd te doen aan de klachten. Van alle 'behandelingen' wordt nagegaan wat het effect hiervan op de klacht is.

1. *HV vraagt niet naar en gaat niet in op welke 'behandeling' dan ook.*
2. *HV vraagt summier naar of gaat summier in op enige behandeling.*
3. *HV vraagt naar en gaat in op enige behandeling, maar is hierin duidelijk onvolledig.*
4. *HV vraagt naar en gaat in op verschillende behandelvormen en het effect.*
5. *HV behandelt het onderwerp behandeling uitgebreid, vraagt naar zelfbehandeling en behandeling door anderen en gaat de effecten na.*

19. Psychiatrische anamnese in engere zin

Met dit item onderzoekt de HV een aantal verschijnselen met betrekking tot CL's conditie.
Oriëntatie. HV gaat na of CL goed georiënteerd in tijd, plaats en persoon is; d.w.z. HV onderzoekt of CL weet welke dag en tijd van de dag het is, waar hij is, en wie hij tegenover zich heeft.
Aandacht en geheugen. HV gaat na of er sprake is van concentratieproblemen, sufheid of slaperigheid. Tevens wordt nagegaan of er sprake is van inprentingstoornissen, korte- of lange termijn geheugenstoornissen.
Waarneming. Met dit item onderzoekt HV waarnemingsstoornissen. HV vraagt of CL wel eens dingen ziet, hoort, ruikt of voelt die anderen niet zien, horen, ruiken of voelen. Nagegaan wordt of hier sprake is van echte hallucinaties of van pseudo-hallucinaties of illusionaire vervalsingen (bijv. "Hoort u echt een stem of is het net alsof u een stem hoort?").
Denken. Met dit item onderzoekt HV gedachtenstoornissen, naar inhoud en vorm.
- Inhoud: Is er sprake van een waan? Zijn er suïcidegedachten of sterk schuldbeladen gedachten? Zijn er obsessies (onvrijwillige steeds terugkerende irrationele gedachten)?
- Vorm: Is er sprake van traagheid van denken, wijdlopigheid, of versneld denken?

Stemming. Stemming is de subjectieve emotie die door CL wordt ervaren gedurende een langere periode. Er kan sprake zijn van een depressieve stemming, een euforische stemming, een geïrriteerde stemming of een angstige stemming. Ook kan er sprake zijn van 'afwezigheid van gevoel'.

1. *Geen van de onderwerpen uit de p.a. worden door de HV expliciet behandeld, of een onderwerp op minimale wijze.*
2. *Twee van de onderwerpen uit de p.a. worden door de HV gedeeltelijk behandeld, of een onderwerp uitvoeriger.*
3. *Drie van de onderwerpen worden door de HV gedeeltelijk behandeld, of twee onderwerpen uitvoeriger.*
4. *Minimaal vier onderwerpen worden door de HV gedeeltelijk behandeld, of drie uitvoerig.*
5. *Minimaal vier van de onderwerpen worden uitvoerig behandeld, de HV neemt een uitstekende p.a. af.*

20. Biografie

HV vormt zich een goed beeld van de levensgeschiedenis van de CL in verschillende periodes, de vroege jeugd, kindertijd, puberteit en volwassenheid. HV gaat na hoe CL zich in verschillende opzichten (psychisch, sociaal, intellectueel, sexueel) heeft ontwikkeld. HV stelt vragen als: "Beschrijf eens het gezin waar je in bent opgegroeid?" "Hoe ging het op school?" "Hoe waren de contacten met leeftijdsgenoten?" Ook vraagt HV naar vergelijkbare klachten bij familieleden.

1. *HV vraagt niets over de levensgeschiedenis.*
2. *HV vraagt alleen summier over de levensgeschiedenis in het kader van de klachtontwikkeling en ontstaansfactoren, en beperkt tot die periode waarin de klacht zich ontwikkelde, òf exploreert zeer summier andere periodes, doch krijgt nauwelijks een beeld van de levensgeschiedenis van CL.*
3. *HV exploreert gedeeltelijk de levensgeschiedenis, en krijgt een matig beeld van de ontwikkeling van CL.*

4. *HV exploreert de levensgeschiedenis van CL over minimaal 2 periodes en minimaal 2 gebieden redelijk.*
5. *HV exploreert de levensgeschiedenis uitgebreid over minimaal 3 periodes en minimaal 3 gebieden.*

21. Persoonlijkheid

Dit item betreft de exploratie van persoonlijkheidseigenschappen en zelfbeeld door de HV. Onder persoonlijkheidseigenschappen wordt verstaan trekken die CL kenmerken *over langere tijd*, bijvoorbeeld precies, verlegen in contacten, extravert, zorgelijke instelling, perfectionistisch, losbandig, etc. Door vragen te stellen als: "Hoe vindt u het om een fout te maken op uw werk?", "Als u een probleem hebt, bespreekt u dat dan met anderen of lost u het bij voorkeur zelf op?" kan de HV zich een beeld vormen van de persoonlijkheid van CL. Het zelfbeeld wordt geëxploreerd door CL te vragen een beschrijving te geven van wat voor persoon hij is, en/of hem te vragen naar zijn goede en slechte eigenschappen. Eigenschappen die samenhangen met de klachten en niet voordien al aanwezig waren, zijn geen persoonlijkheidseigenschappen.

1. *HV vraagt niet naar zelfbeeld en vraagt niet naar en gaat niet in op persoonlijkheidseigenschappen.*
2. *HV vraagt niet naar zelfbeeld en vraagt zeer summier naar of gaat zeer summier in op persoonlijkheidseigenschappen.*
3. *HV vraagt naar zelfbeeld of vraagt gedeeltelijk naar of gaat gedeeltelijk in op persoonlijkheidseigenschappen.*
4. *HV vraagt naar zelfbeeld en vraagt naar bepaalde persoonlijkheidseigenschappen, of gaat in op spontane mededelingen van CL hierover.*
5. *HV vraagt naar zelfbeeld en vraagt uitgebreid naar persoonlijkheidseigenschappen, en gaat in op spontane mededelingen van CL hierover.*

22. Sociale situatie

Dit item betreft de exploratie van sociale contacten door de HV. HV dient zowel de aanwezigheid als de kwaliteit van sociale contacten na te vragen. Het gaat hierbij om relaties binnen het huidige gezin (de primaire groep) en daarbuiten. HV krijgt zo een beeld van hoe CL sociaal functioneert en hoeveel sociale steun hij in zijn omgeving heeft. Dit kan van belang zijn bij het bepalen van het verdere beleid.

1. *HV stelt geen vragen over de sociale situatie en gaat niet in op mededelingen van CL hierover.*
2. *HV stelt zeer summier vragen over de sociale situatie of gaat zeer summier in op spontane mededelingen van CL hierover.*
3. *HV exploreert gedeeltelijk de sociale situatie, door of de aanwezigheid van sociale contacten na te gaan zonder een beeld te krijgen van de kwaliteit, of de kwaliteit van enkele sociale relaties na te gaan, zonder een beeld te krijgen van andere sociale contacten.*
4. *HV exploreert zowel de aanwezigheid als kwaliteit van een aantal sociale contacten.*
5. *HV exploreert de aanwezigheid en kwaliteit van sociale contacten, zowel binnen als buiten de primaire groep.*

23. Maatschappelijke situatie en bezigheden

Onder maatschappelijke situatie en bezigheden wordt hier verstaan de werksituatie, andere bezigheden (vrije tijdsbesteding), financiële situatie, en woonsituatie. Bij dit item dient HV na te vragen hoe CL op zijn werk en in zijn vrije tijd functioneert, in hoeverre hij in staat is zichzelf te verzorgen (zowel financieel als anderszins), en hoe tevreden hij is met zijn maatschappelijke situatie en bezigheden.

1. *HV vraagt naar geen van deze onderwerpen.*
2. *HV vraagt naar een van deze onderwerpen (of gaat in op spontane mededelingen van CL hierover) doch gaat niet in op satisfactie.*
3. *HV vraagt naar of gaat in op spontane mededelingen van CL over meerdere onderwerpen doch gaat niet in op satisfactie, of vraagt naar of gaat in op spontane mededelingen van CL over een onderwerp en gaat ook in op satisfactie.*
4. *HV vraagt naar of gaat in op spontane mededelingen van CL over meerdere onderwerpen en gaat in op satisfactie.*
5. *HV vraagt naar of gaat in op spontane mededelingen van CL over de meeste onderwerpen en gaat in op satisfactie.*

Totaaloordeel inhoud

Geef een subjectief eindoordeel m.b.t. de kwaliteit van de inhoud van het anamnestisch interview (rapportcijfer 1-10).

BEOORDELINGSLIJST GESPREKSVAARDIGHEDEN BIJ DE ANAMNESE

1. Aandachtgevend gedrag

1 2 3 4 5

K. ontspannen toegenegen houding
geinteresseerde gelaatsuitdrukking
stimulerend oogcontact
aanmoedigende gebaren
verbale aanmoedigingen
aandachtige stilte
F. waar nodig
juist gedoseerd

2. Vragen stellen

1 2 3 4 5

K. duidelijk geformuleerd
enkelvoudig
neutraal
F. open vragen bij exploratie
gesloten vragen wanneer HV iets specifieks wil weten
juist gedoseerd

3. Concretiseren (doorvragen)

1 2 3 4 5

K. bij voorkeur open gesteld
aansluitend op wat CL net heeft verteld
F. waar nodig (vage uitspraken, verdere verduidelijking)
juist gedoseerd

4. Parafraseren

1 2 3 4 5

K. duidelijk en kernachtig
in eigen woorden
inhoudelijk juist
juiste intonatie
F. waar nodig
juist gedoseerd

5. Reflecteren van gevoel

1 2 3 4 5

K. duidelijk en kernachtig
in eigen woorden
inhoudelijk + intensiteit juist
verbale en nonverbale gevoelens
onderliggende gevoelens
juiste intonatie
F. waar nodig (belangrijke gevoelens worden gereflecteerd)
juist gedoseerd

6. Samenvatten

1 2 3 4 5

K. duidelijk en kernachtig
inhoudelijk juist
F. waar nodig (als CL alles verteld heeft over een onderwerp,
overgangen, einde gesprek)
juist gedoseerd

7. Geven van informatie

1 2 3 4 5

K. duidelijk en concreet
in kleine stukjes
afgestemd op CL
F. waar nodig (introductie, overgangen, afsluiting, vragen CL)
juist gedoseerd

8. Vragen om feedback

1 2 3 4 5

K. gesloten of open vragen
geinteresseerd
F. waar nodig (bij geven van informatie, bij spontane feedback CL)
juist gedoseerd

9. Hardop denken

1 2 3 4 5

K. open en duidelijk
geven CL inzicht in gedachten HV
F. waar nodig (vragen uit referentiekader HV, blokkade HV, bevorderen SW)
juist gedoseerd

10. Structureren

1 2 3 4 5

K. ordenend en sturend
duidelijk voor CL
F. waar nodig (afdwalen, wanneer doelen niet gehaald dreigen te worden)
juist gedoseerd

11. **Probleemverheldering** 1 2 3 4 5
duidelijke en concreet beeld van aanmeldingsprobleem O , beleving CL O

12. **exploratie andere problemen** 1 2 3 4 5
vragen naar O , ingaan op O

13. **aanleiding komst en hulpvraag** 1 2 3 4 5
vraagt naar aanleiding komst O, hulpvraag O

14. **uitdiepen hoofdklacht** 1 2 3 4 5
aard O , intensiteit O , begeleidende symptomen O O O O O O

15. **ontstaansfactoren** 1 2 3 4 5
hoe ontstaan O , samenhangende factoren O , ideeën CI O

16. **verloop** 1 2 3 4 5
duur O , ontwikkeling O , factoren bij vermindering of verergering O

17. **invloed op dagelijks leven** 1 2 3 4 5
op gedrag CL O , reactie omgeving O

18. **behandeling** 1 2 3 4 5
eerdere behandeling(en) O , zelfhulp O , medicatie O , alc. O , drugs O
effect O

19. **psychiatrische anamnese i.e.z.** 1 2 3 4 5
oriëntatie tijd O , plaats O , persoon O
aandacht en geheugen aandacht O , geheugen O
waarneming hallucinaties O , soorten O , echt of pseudo O
denken inhoud: waan O , suïcide, schuld O , obsessies O , préoccupaties O
vorm: traagheid, wijdlopigheid, snelheid O
stemming O

20. **biografie** 1 2 3 4 5
genogram O,
vroege jeugd O , kindertijd O , puberteit O , volw. O
psychisch O , sociaal O , intellectueel O , sexueel O

21. **persoonlijkheid** 1 2 3 4 5
persoonlijkheidseigenschappenO O O, zelfbeeld O

22. **sociale situatie** 1 2 3 4 5
aanwezigheid O , kwaliteit O
binnen gezin O , daarbuiten O

23. **maatschappelijke situatie** 1 2 3 4 5
werksituatie O , vrije tijdsbesteding O , zelfverzorging O , satisfactie O

Kwaliteit en functionaliteit van de gespreksvaardigheden (1-10):

Inhoudelijke kwaliteit (1 - 10):

Appendix II

Probleemoplossingsvragenlijst[1] (Problem Solving Questionnaire)

Instructie: Wilt u de onderstaande vragen zo volledig mogelijk invullen m.b.t. de patiënt die u zojuist heeft geïnterviewd? Voor het beantwoorden van de vragen heeft u een kwartier de tijd. U mag hierbij de DSM-III-R gebruiken.

1. Diagnostische classificatie van de hoofdklacht.
Noteer de meest waarschijnlijke diagnose, in DSM-III-R termen. U hoeft geen codes te geven, alleen de omschrijving (als er meerdere hoofdklachten zijn geeft u meerdere diagnoses, indien het DSM-III-R systeem dit toestaat).

2. Differentiaaldiagnose.
Noem maximaal 2 diagnoses die u minder waarschijnlijk acht, maar ook overweegt. Noteer ze in volgorde van waarschijnlijkheid (1= meest waarschijnlijk, 2= minder waarschijnlijk).

3. Recente beïnvloedende factoren.
Beschrijf de twee, in uw ogen, meest belangrijke recente factoren die een rol spelen in het ontstaan van het probleem (bijv. psychosociale stress, life-events, somatische factoren, etc.)

4. Reeds langer bestaande beïnvloedende factoren.
Welke, reeds langer bestaande factoren precipiteerden het huidige probleem en/of hielden het probleem in stand? Noem de 2 belangrijkste.

5. Predisponerende factoren.
Welke persoonlijkheids- of andere kenmerken van de cliënt zijn naar uw idee van belang bij de ontwikkeling van het probleem? Noem er 2.

6. Diagnostiek.
Beschrijf de stappen die u zou ondernemen m.b.t. verdere diagnostiek (maximaal 2 stappen).

7. Behandeling.
Beschrijf welke behandeling of combinaties van behandelingen u overweegt. Beargumenteer uw voorstel.

---

[1] Deze lijst is gebaseerd op de Problem Solving Questionnaire van Kraan en Crijnen (1987).

Appendix III

Patiëntsatisfactie-vragenlijst (Patient Satisfaction with Communication Checklist)[1]

1.	De interviewer toonde begrip voor mijn gedragingen en gevoelens	1	2	3	4	5
2.	De interviewer gaf mijn uitspraken precies weer zoals ik ze bedoelde	1	2	3	4	5
3.	De interviewer gebruikte gewoon nederlands	1	2	3	4	5
4.•	De interviewer ging te snel van het ene naar het andere onderwerp over	1	2	3	4	5
5.	De interviewer vroeg of ik accoord was met de verdere gang van zaken	1	2	3	4	5
6.	Mijn ongerustheid is door het gesprek met de interviewer afgenomen	1	2	3	4	5
7.	De interviewer heeft gevraagd of ik de uitleg en/of adviezen begrepen heb	1	2	3	4	5
8.	De interviewer maakte mij het praten makkelijker	1	2	3	4	5
9.	Ik werk zeker aan de voorgestelde procedure mee	1	2	3	4	5
10.•	De vragen en opmerkingen van de interviewer maakten mij ongerust en zenuwachtig	1	2	3	4	5
11.	De interviewer maakte het mij mogelijk om alles te vragen wat ik wilde	1	2	3	4	5
12.	De afspraken die de interviewer met mij gemaakt heeft kan ik navertellen	1	2	3	4	5
13.	Ik voelde mij op mijn gemak gesteld door de interviewer	1	2	3	4	5
14.	De interviewer gaf mij de ruimte mijn verhaal te vertellen	1	2	3	4	5
15.	De interviewer ging in op hetgeen ik tijdens het gesprek naar voren bracht	1	2	3	4	5
16.	De interviewer had aandacht voor de verschillende aspecten van mijn probleem	1	2	3	4	5
17.	De interviewer drong tijdens het gesprek door tot de kern van het probleem	1	2	3	4	5
18.	De interviewer gaf blijk een goed inzicht te hebben in mijn problemen	1	2	3	4	5

[1] Deze lijst is een op de GGZ aangepaste versie van de Patient Satisfaction with Communication Checklist van Crijnen & Kraan (1987). Van de met • aangeduide items worden de scores omgekeerd.

Dankwoord

Allereerst wil ik mijn beide promotoren, Henk Schmidt en Cees van der Vleuten, bedanken voor het meedenken over de opzet en analyse van de verschillende onderzoeken en hun commentaar op de hoofdstukken van het proefschrift. Tessa van Mourik, die eerst als student-assistent, daarna als onderzoeksstagiaire, en tot slot als docent, bij diverse onderzoeken betrokken was, bedank ik voor haar hulp en stimulerende samenwerking.

Een groot aantal mensen (studenten, docenten, en hulpverleners) ben ik erkentelijk voor het beoordelen van videobanden: Berien Bakker, Geke Blok, Brigitte Boon, Jolanda Coolen, Nicole van Gastel, Els Graafsma, Marie-anne Haenen, Tanja van den Hogen, Jannette Hommes, Reinier Kreutzkamp, Rutger Lulofs, Reinhilde Melles, Heidi Mertens, Helga Nauta, Marleen Rooijmans, Miranda Vermeeren, en Raymond Ubbink.

Zonder alle namen te noemen bedank ik de hulpverleners uit GGZ-instellingen die bij hebben gedragen aan het maken van antwoordsleutels door het analyseren van casuïstiek. De studenten en hulpverleners, die bereid waren om als proefpersonen voor de camera's de simulatiepatiënten te interviewen, ben ik zeer erkentelijk voor hun deelname.

Gerry van Wunnik wordt bedankt voor zijn assistentie bij het maken van de video-opnames. Bert Kerkhofs, Diana Riksen en Erik Schouten hebben geholpen bij het rekenwerk, waarvoor mijn dank. Herro Kraan bedank ik voor het beschikbaar stellen van video-banden. Marie-Louise Schreurs was niet direct bij het onderzoek betrokken, maar haar wil ik bedanken voor de prettige samenwerking bij het vaardigheidsonderwijs. Petry Thiemann bedank ik voor de uitstekende wijze, waarop zij het vaardigheidsonderwijs steeds organiseerde. Mijn dank gaat verder uit naar al diegenen die als simulatiepatiënt op voortreffelijke wijze psychische problematiek hebben uitgebeeld. Marcel van den Hout bedank ik voor de voorwaarden die hij geschapen heeft om dit proefschrift af te kunnen ronden, nadat ik bij de vakgroep DEP werd aangesteld. Marja Hortulanus ben ik dankbaar voor het verzorgen van de lay-out van het proefschrift. Marjorie Mulders heeft het proefschrift gecorrigeerd op de Engelse taal. Rudolf Ponds bedank ik voor het maken van de grafieken.

Dankwoord

Curriculum vitae

Susan Bögels werd geboren op 25 november 1960 te Geleen. Na het eindexamen Atheneum-A te Delfzijl in 1979 studeerde zij tot 1986 Psychologie aan de Rijksuniversiteit Groningen. In mei 1986 behaalde zij het doctoraalexamen psychologie, met als afstudeerrichting klinische psychologie en bijvak kinderpsychiatrie. Sinds juni 1986 is zij als docent verbonden aan de Rijksuniversiteit Limburg. Zij hield zich van 1986 tot 1992 (aanvankelijk bij het Skillslab, vanaf 1989 bij de vakgroep Onderwijsontwikkeling en Onderwijsresearch) bezig met de ontwikkeling van en onderzoek naar vaardigheidsonderwijs op het gebied van de Geestelijke Gezondheidkunde. Daarnaast is zij vanaf 1988 therapeutisch werkzaam bij het Dwarsverband Academisering van de RIAGG Maastricht. In 1992 behaalde zij het gewoon lidmaatschap van de Nederlandse Vereniging voor Gedragstherapie en werd zij geregistreerd als basis-psychotherapeut. Vanaf november 1992 werkt zij als universitair docent bij de vakgroep Differentiële en Experimentele Psychologie (Rijksuniversiteit Limburg), waar zij naast het geven van onderwijs onderzoek doet naar het ontstaan en de behandeling van sociale fobie. Tevens is ze als gastdocent verbonden aan de RINO Zuid-Nederland in het kader van de post-academische opleiding tot gedragstherapeut.

Datawyse | Universitaire Pers Maastricht
ISBN 90 5278 154 0