Maastricht University

Al doende leert men : enkele studies naar aspecten van betrouwbaarheid en validiteit over de toetsing van vaardigheden

Citation for published version (APA):

van Luijk, S. J. (1994). *Al doende leert men : enkele studies naar aspecten van betrouwbaarheid en validiteit over de toetsing van vaardigheden*. Datawyse / Universitaire Pers Maastricht. https://doi.org/10.26481/dis.19941021sl

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

DOI:
10.26481/dis.19941021sl

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

Please check the document version of this publication:

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

Link to publication

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

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

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

www.umlib.nl/taverne-license

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

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 19 Sep. 2021

SAMENVATTING

Verspreid over de gehele wereld hebben observatietoetsen zich in de loop der jaren een belangrijke plaats weten te verwerven binnen een groot aantal medische opleidingen. Binnen het medisch onderwijs bestond de indruk dat alleen kennismeting niet voldoende informatie verschaft over de kwaliteit van de basisarts.
Vaardigheidstoetsen doen uitspraken over het vaardigheidsniveau van toekomstige artsen en dragen daarmee bij aan de bestaande meetinformatie omtrent de kwaliteit van artsen die de opleiding afronden. Bij de introductie van deze toetsen waren er vrijwel geen psychometrische gegevens beschikbaar over de eigenschappen van dergelijke toetsen. De afgelopen jaren is er echter een grote stroom onderzoeksgegevens beschikbaar gekomen over de vaardigheidstoets. Ook dit proefschrift draagt hieraan bij door het beschrijven van een aantal studies rondom de betrouwbaarheid en validiteit van de vaardigheidstoets (hoofdstuk zes tot en met negen). Deze empirische hoofdstukken worden voorafgegaan door een aantal hoofdstukken die bedoeld zijn om de gestelde onderzoeksvragen te plaatsen.

Hoofdstuk één geeft een nadere specificatie van het begrip vaardigheid. Hierbij wordt een onderscheid gemaakt tussen psychomotorische vaardigheden enerzijds en sociale vaardigheden anderzijds. Vervolgens wordt ingegaan op de argumenten die ten grondslag liggen aan de implementatie van deze vaardigheden in een apart "vaardigheidscurriculum" binnen de studie geneeskunde. Met name wordt aandacht besteed aan het leren van psychomotorische en sociale vaardigheden en de onderlinge integratie ervan binnen de specifieke context van het probleemgestuurd onderwijs.

Hoofdstuk twee geeft een beschrijving van de wijze waarop vaardigheden gemeten worden. Hierbij wordt grofweg een onderscheid gemaakt tussen de Objective Structured Clinical Examination (OSCE) en de Standardized Patient-based Test (SP-based test). Het hoofdstuk wordt afgesloten met literatuurgegevens omtrent betrouwbaarheid en validiteit van vaardigheidstoetsen. Geconcludeerd wordt dat de totale toetsbetrouwbaarheid met name afhankelijk is van het aantal verschillende inhoudelijke aspecten dat gemeten wordt. Dit is een punt van zorg bij observatietoetsen. De inhoudsvaliditeit van dergelijke toetsen wordt daarentegen vaak als een sterk punt beschouwd.

Hoofdstuk drie gaat in op de plaats van de vaardigheidstoets binnen het geheel van evaluatie-activiteiten in het probleemgestuurd onderwijs aan de Faculteit der Geneeskunde. Gedetailleerd wordt aangegeven hoe beoordelingen tot stand komen zowel met betrekking tot de vaardigheidstoets als ook met betrekking tot de verschillende andere evaluatie instrumenten aan de faculteit. Ingegaan wordt eveneens op de kwaliteitszorg van de afgenomen toetsen.
Vervolgens wordt uitgebreid ingegaan op de vorm en inhoud van de vaardigheidstoets. Gesteld wordt dat de vaardigheidstoets zoals die in Maastricht wordt afgenomen, zowel van de OSCE alsook van de SP-based test, zoals beschreven in hoofdstuk twee, delen heeft overgenomen. De vorm van de vaardigheidstoets in de eerste fase van de studie lijkt veel op de OSCE, terwijl de vaardigheidstoetsen in de tweede fase van de studie meer overeenkomen met de SP-based test.

Hoofdstuk vier gaat in op de ervaringen opgedaan met de vaardigheidstoets. Deze ervaringen zijn weergegeven op enquêtes die systematisch gedurende de afgelopen elf jaar zijn afgenomen bij studenten en observatoren. Over het algemeen laten de uitslagen van de enquêtes een redelijk stabiel beeld zien. De vaardigheidstoets wordt door studenten en observatoren als een relevante toets gezien voor de basisarts. Kritiek heeft men soms op het fragmentarische en gedetailleerde karakter van de inhoudelijke elementen waaruit de toets bestaat. Ook hebben observatoren de indruk dat het proces van de vaardigheid, door de grote hoeveelheid items die hierop betrekking hebben, meer wordt gewaardeerd in de toetsuitslag dan de uitkomst van de handeling. Het werken met gedetailleerde criterialijsten wordt over het algemeen minder gewaardeerd dan globale criterialijsten. Studenten lijken feedback van observatoren tijdens de toets erg op prijs te stellen.

Hoofdstuk vijf geeft de vraagstellingen voor onderzoek weer, op basis van de informatie verstrekt in de eerste vier hoofdstukken.

Hoofdstuk zes bevat een inhoudelijke analyse van twee vaardigheidstoetsen. Hierbij zijn items uit de toets inhoudelijk geclassificeerd. Uit de classificatie blijkt dat er een scheve verdeling bestaat met betrekking tot de hoeveelheid items behorend bij bepaalde inhoudelijke categorieën van de onderzochte toetsen. Conform de waarnemingen van observatoren in hoofdstuk vier zijn items met een cognitief en procesmatig karakter aanzienlijk zwaarder in de toets vertegenwoordigd dan items met een produktmatig karakter. Studenten blijken items met een procesmatig en cognitief karakter ook beter te beheersen dan items met een produktmatig karakter. Gelet op de doelen van het vaardigheidsonderwijs in de verschillende jaren (hoofdstuk een) lijkt er een inhoudelijke discrepantie te bestaan tussen de inhoud van het onderwijs en de inhoud van de vaardigheidstoets in de hogere studiejaren. Inhoudelijke veranderingen in de vaardigheidstoets die meer aansluiten op de onderwijsdoelen zouden kunnen leiden tot grotere aantallen gezakte studenten en een lagere betrouwbaarheid voor toetsen.

Hoofdstuk zeven gaat in op het verschil in betrouwbaarheid tussen gedetailleerde en meer globale criterialijsten. In deze studie worden beide soorten lijsten door dezelfde observator gebruikt bij het beoordelen van vaardigheden bij studenten. Een controle studie wijst uit dat het invullen van dergelijke lijsten na elkaar door dezelfde observator nauwelijks invloed heeft op de beoordeling. Uit de vergelijking van de scores van studenten blijkt dat beoordelingen op basis van een gedetailleerde schaal tot meer extreme scores leidt dan beoordelingen op basis van een globale schaal. Ten aanzien van de betrouwbaarheid van beide soorten criterialijsten kan worden geconcludeerd dat er geen duidelijk verschil bestaat tussen globale en gedetailleerde criterialijsten op het niveau van de totale toetsbetrouwbaarheid. Ook ten aanzien van de inhoud van de verzamelde informatie komen beide beoordelingsvormen overeen. Gelet op de voorkeur van observatoren voor minder gedetailleerde criterialijsten (hoofdstuk 4) heeft deze bevinding belangrijke praktische consequenties.

Hoofdstuk acht heeft betrekking op de wijze waarop de normering voor de vaardigheidstoets kan worden vastgesteld. In aansluiting op hoofdstuk drie waar de huidige regelingen met betrekking tot de normering van de

vaardigheidstoets uitvoerig worden weergegeven, geeft hoofdstuk acht een aantal alternatieven voor het bepalen van de cesuur. Uit de resultaten blijkt dat de verschillende benaderingswijzen om tot een bepaalde cesuur te komen, enorme grote verschillen in zak/slaagpercentages teweeg kunnen brengen. De cesuur die bij de vaardigheidstoets wordt toegepast blijkt niet tot extreme zak/slaagpercentages te leiden. Geconcludeerd wordt dat het komen tot een bepaalde normstelling gebaseerd moet zijn op een zorgvuldige afweging van doelen die men beoogt te bereiken met toetsing.

Hoofdstuk negen heeft betrekking op de predictieve validiteit van de vaardigheidstoets voor het succesvol doorlopen van de klinische stages. Als referentiepunt wordt hierbij de toets voor het meten van algemene medische kennis, de voortgangstoets (hoofdstuk drie) gebruikt. Geheel tegen de verwachtingen in blijkt de vaardigheidstoets een slechtere voorspeller voor praktisch functioneren later in de studie dan de voortgangstoets. Dit kan liggen aan de inhoud van de vaardigheidstoets (hoofdstuk 6) maar ook aan de wijze waarop stagebeoordelingen tot stand komen. Dat de voortgangstoets een dergelijk grote predictieve waarde heeft voor het succesvol doorlopen van de stages eveneens opvallend. Geconcludeerd wordt dat nader onderzoek op dit terrein wenselijk is om de gegeven verklaringen voor deze onverwachte bevindingen te toetsen.

SUMMARY

Testing skills by standardized observation is a relatively new area in medical education. Starting in the late seventies, this method of testing is slowly becoming an integrated part of the assessment program in many medical schools. Hence, educational research with respect to the validity and reliability of this assessment instrument is relevant. Substantial work has already been carried out in this field of research. Nevertheless many questions remain unanswered.
This thesis discusses some questions related to the implementation of an observational test, called the Skills Test, in the Maastricht problem-based curriculum. Empirical research is preceded by four chapters concerning the role, position and experience of patient-based testing in the Maastricht medical school.

Chapter 1 describes the various definitions of a skill given in the literature and how the term skill is defined in the Maastricht educational program. The chapter also discusses how skills are taught in a curriculum. A description is given of the content and organization of the skillstraining program in the curriculum.

Chapter 2 provides an overview of observational testing in medical education. The Objective Structured Clinical Examination (OSCE) and the Standardized Patient-Based test are described, and differences between the two formats are discussed. Also problems of reliability and validity are explained.

Chapter 3 focuses on the assessment system implemented by the Maastricht medical school and the role of the Skills Test in the whole system. Compared to other curricula the Skills Test plays a more important role in decisions concerning graduation of students. In the

second part of this chapter, a more detailed description is given about the content and format of the Skills Test. Differences and similarities between the Skills Test on the one hand and the OSCE and Patient-Based tests on the other are analyzed. These differences, however small, are mainly related to differences in aims of the tests. The similarities focus mainly on the psychometric characteristics of the different test formats.

Chapter 4 reviews the experiences with Skills Testing during the last ten years based on results of numerous questionnaires. Both students and staff consider the Skills Test as relevant and important. However, there is also criticism. Students experience the test as very stressful. It seems that because "being observed" is stressful in itself. Sometimes, the behaviour of some observers is also a very stress-stimulating factor. The criticism of the faculty-observers is more related to the validity of the test. Especially the very detailed checklists are criticized. Also the relevance of items is discussed. Many observers have the feeling that students learn the tricks but do not understand the essence of practising skills.
A final experience is that skillstesting in the way it is implemented, is a dominant factor steering student-learning.

Chapter 5 summarizes the research questions derived from the information and experiences with observational testing described in the preceding chapters. The research questions deal with the content of the Skills Test, the specificity of items in the checklist, the different standard-setting methods en the predictive validity of the Skills Test.

Chapter 6 presents a study on the content of the Skills Test. This chapter investigates the criticism of observers that items concerning procedures of a skill and concerning knowledge are far more represented in the test than items concerning outcome of skills. If this is true, it makes the test less representative for future practice, creating a problem of validity.
To investigate this problem, the items of the Skills Test were divided in different content categories. The categories were related to items concerning knowledge of skills, items representing motoric performance, items concerning the process of a skill and items related to the outcome of skills. The criticism of observers appears to be true. The Skills Test is dominated by items which refer to knowledge and process. It also

appeared that items related to outcome show the lowest scores by students. Pass/fail decisions only based on outcome items would lead to more students failing compared to other categories. The different categories show moderate to high correlations. The reliability of the different categories is also assessed. Tests consisting of only process items are more reliable than tests consisting of outcome items.

Chapter 7 focuses on the value of global rating scales compared to the more detailed checklists. More specifically, this chapter describes the effects on scores and ranking of students using global rating scales and checklists. Also the results of students based on both formats are compared with the (global) impressions of the staff members and with the results of a knowledge test (progress test).
It appears that checklist scoring leads to higher scores for students including a greater standard deviation than scores based on global ratings. Related to pass/fail decisions, global ratings are milder for students than detailed ratings.
With respect to the reliability, the interrater reliability of checklist ratings is better. But despite this, and more importantly, the total-test reliability of both formats is more or less the same. Rater-errors at the station level apparently average out adequately across stations. The correlation between the different formats is high. Correlations with the progress test were low and the same for both formats.

Chapter 8 reviews some standard-setting methods. The advantages and disadvantages of the different methods are discussed. A number of standard-setting methods are applied on a sample Skills Test. The different methods lead to substantial differences in pass/fail decisions. Standard-setting methods related to the content of the test (domain-referenced) tend to be more severe than norm-referenced methods. It could be argued that the severity of the domain-referenced method is probably the result of the bias that judges only judge the content they are familiar with, and therefore tend to overestimate the competency of students.
In terms of severity, the standard-setting method which is used in the regular assessment program in Maastricht lies somewhere in between the norm- and domain- referenced methods.

Chapter 9 discusses the predictive validity of the results of the Skills Test and the knowledge test (progress test) in the first four preclinical years in relation to clinical ratings of students in years five and six. Contrary to expectations, the knowledge test has a better predictive validity than the Skills Test. A possible explanation is the dominance of knowledge testing as part of the clinical rating. An alternative explanation is that the content of the Skills Test does not cover clinical practice. It is argued that the first explanation is more likely.