Selection for Medical School

Selection for medical school

The quest for validity

Sanne Schreurs

SELECTION FOR MEDICAL SCHOOL

The quest for validity

Sanne Schreurs

The research reported here was carried out at

in the School of Health Professions Education

in the context of the research school:

Interuniversity Center for Educational Research

Sanne Schreurs
Selection for medical school: the quest for validity

ISBN 978-94-028-1931-1

Cover design and lay-out: Lieke Moonen-Schreurs
Printing: Ipskamp Printing Maastricht

De uitgave van dit proefschrift is mede ondersteund door de Nederlandse Vereniging voor Medisch Onderwijs

Selection for medical school

The quest for validity

PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan Universiteit Maastricht

op gezag van de rector magnificus prof. dr. R.M. Letschert,

volgens besluit van het college van decanen

in het openbaar te verdedigen

op vrijdag 20 maart 2020 om 10.00 uur

door

Sanne Schreurs

Promotores
Prof. dr. M.G.A. oude Egbrink
Prof. dr. J.A. Cleland (University of Aberdeen, UK)

Copromotor
Dr. K.B.J.M. Cleutjens

Beoordelingscommissie
Prof. dr. S. Heeneman (voorzitter)
Prof. dr. F.J.M. Feron
Prof. dr. J.H.J.M. van Krieken (Radboud Universiteit Nijmegen)
Prof. dr. A.J.J.A. Scherpbier
Dr. K.M. Stegers-Jager (Erasmus Universiteit Rotterdam)

Paranimfen
Drs. Max Colombi
Drs. Margo Korpel

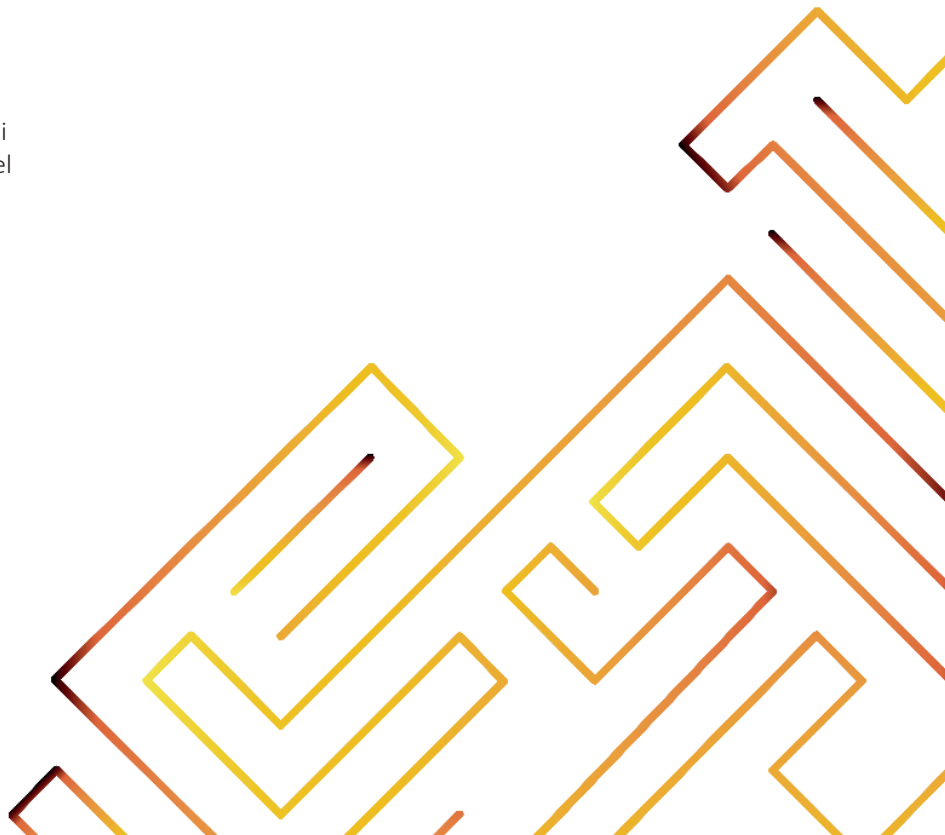
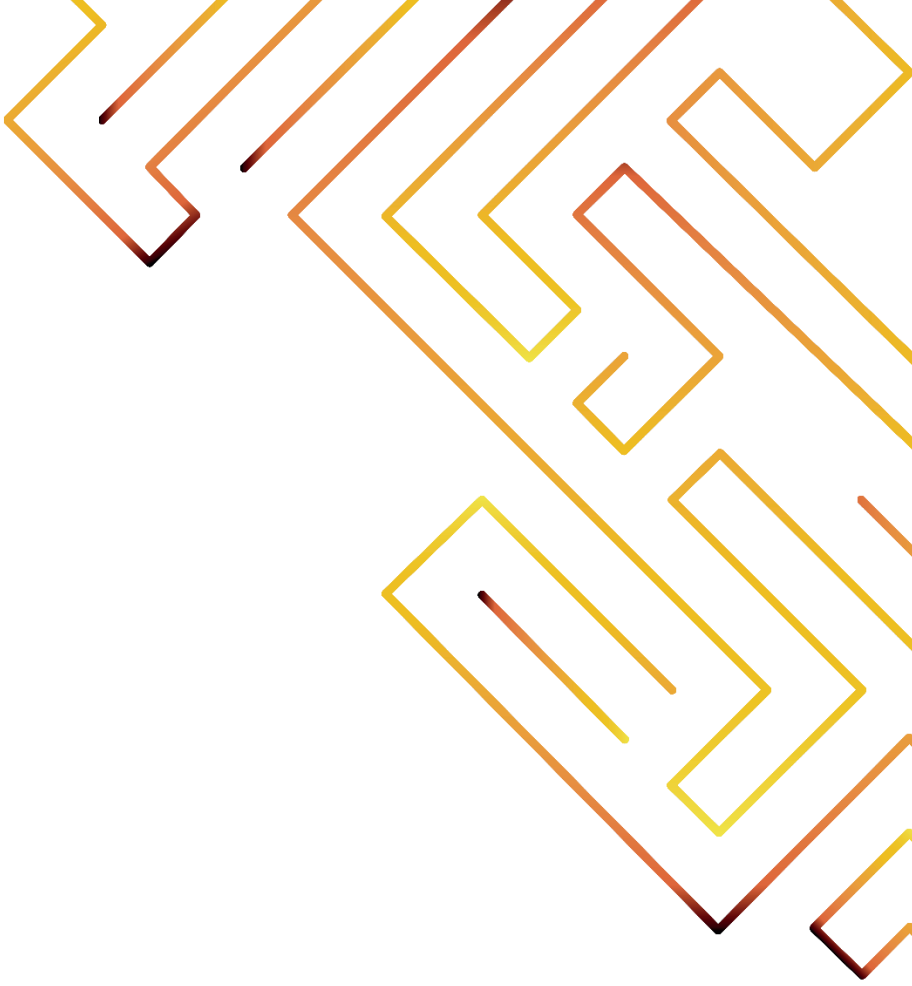


Table of contents



CHAPTER 1

General introduction

1.1 History of selection

Medical schools are responsible for educating good doctors who can provide optimal care to their patients. To this purpose, the schools must ensure educational quality while making cost-efficient decisions and adhering to society's needs (1, 2). An important first step in this process is the selection of the best suited students from a pool of applicants that is multiple times larger than the number of available study places (3). Therefore, in recent years, selection has become increasingly topical for medical educators throughout the world.

Several developments have shaped the field of selection for medical schools in important ways. The first key development was the publication of the Flexner report in 1910, in which Abraham Flexner called upon medical schools to start recruiting students for a strong scientific basis, instead of based on who knew who (4). This led to the introduction of the first widespread selection requirement: cognitive or academic proficiency (5). A definition for cognitive ability that is still applied in medical education was found in a book from the field of psychology: "the repertoire of intellectual (or cognitive) skills available to the person at a particular point in time" (6). In clearer terms: a person's capacity to acquire, process and utilize knowledge. In the field of selection, cognitive ability is often operationalized as 'academic proficiency'.

The next important landmark for selection was an article by Papadakis et al., in which the authors showed that the majority of complaints about medical professionals were filed because of unprofessional behavior. Importantly, these unprofessional behaviors in clinical practice were often preceded by unprofessional behavior during medical studies (7), suggesting that it may be possible to predict future unprofessional behavior using previous behaviors (the principle of behavioral consistency) (8). There had been previous calls for including personal qualities in the admissions processes for medicine, for example in the Edinburgh declaration: "*In the selection of medical students, employ methods that go beyond intellectual ability and academic achievement, to include measures of personal qualities*" (9). However, it was the notion of behavioral consistency set forward by Papadakis (7), which eventually led to profound changes in the practice of selection: schools actually incorporating the assessment of 'non-academic' or 'non-cognitive' requirements in order to predict later non-academic/non-cognitive performance. The definition of 'non-cognitive' or 'non-academic' is a greater challenge than that of 'cognitive', as 'non-cognitive' and 'non-academic' only define what they are not: "not relating to or based on conscious intellectual activity". Therefore, throughout this dissertation, the term '(inter)personal skills' will be used to summarize behaviors not so much related to cognitive ability, and defined as "aptitudes one needs to effectively work together and communicate with others, but also personal aptitudes such as personality, coping with emotions and values".

Historically, 'cognitive' or 'academic' aptitudes have been contrasted with 'non-cognitive' and/or 'non-academic' aptitudes, seeing them as 'black and white'.

However, in the context of selection, many aptitudes that are measured do not necessarily fall into only one category (10, 11) and the aptitudes involved are entangled to such an extent that it makes more sense to place these aptitudes in a continuum ranging from cognitive to (inter)personal.

1.2 Selection tools

Along with the shift from solely cognitive to including more (inter)personal aptitudes to be assessed in selection procedures, the tools with which to measure specific aptitudes changed. Important stakeholders also affected the tools used in selection, shifting from a very specific focus on predictive value to more overarching ideas of validity, acceptability, suitability, feasibility, and so on. Many selection tools have been developed to measure cognitive or 'non-academic'/'non-cognitive' aptitude, or combinations thereof, as described in a number of reviews (e.g. 5, 8, 10, 12-14). One extensive review by Cleland et al (10) in 2012 led to an overview of evidence on the reliability, validity, candidate acceptability, costs, effects on widening access and susceptibility to coaching for specific tools, as summarized in Table 1.1. Below, an overview of different categories of selection tools with the underlying evidence is presented, based on information in the latest reviews (10, 13, 14).

Table 1.1: Evidence for well-established tools in selection, adapted from Cleland et al. (2012)

	Reliability	Validity	Candidate acceptance	Cost (school)	Cost (applicant)	Promotes WA	Coaching
Academic record	+	+	+	-	-	-	-
Traditional interviews	-	-	+	±	±	?	+
Multiple mini interviews	± / +	±	+	+	±	-	±
Situational judgement tests	+	± / +	± / +	±	-	±	- / ±
Aptitude testing	+	Var.	±	?	± / +	Var.	- / ±
Autobiographical submissions	-	-	+	±	-	-	+
References	-	-	+	±	-	-	N/A
Personality tests	+	±	- / ±	- / ±	± / +	?	± / +

WA = Widening Access; + = high; ± = moderate, - = low, ? = unknown; Var. = Various

Firstly, academic records, or Previous Academic Attainment (PAA), have a longstanding history of use (since around the time of Flexner's call), resulting in a widespread use in selection procedures all over the world, as well as an overwhelming amount of research on this tool. This is in contrast with some of the other tools used in selection procedures, which are newer and less often used and researched (see later). In the long period of time in which PAA's have been used, a bulk of evidence supporting its continued use has been gathered. PAA appears to be the most predictive tool available for the early years of medical school (15) and has been recommended for use in that context. However, there is also some research available that has shown PAA to be predictive of clinical performance: White et al. (16) showed that undergraduate GPA used in the selection for postgraduate medical education predicts

clinical performance. Possible downsides of PAA are potential biases against applicants with an immigration background or lower socio-economic status (SES). In addition, there appears to be an inflation of A-levels in the United Kingdom (among other countries; 8, 13). On the other hand, PAAs are relatively straightforward and inexpensive to implement (8).

Next, the category of interviews. There are many possible ways of interviewing. Cleland et al (10) used the traditional versus multiple mini interview distinction, but in more recent literature, three categories appear: the unstructured traditional interview, the structured interview and the multiple mini interview (MMI). In an unstructured interview, a candidate is interviewed by one or multiple interviewers, who, in general, have done little prior preparation and did not structure their questions, resulting in different questions posed to each candidate. Because of this lack of preparation and structure, these interviews have shown little or no reliability and validity. Moreover, there is evidence of bias (8). Structured interviews and MMIs, on the other hand, are effective ways to turn the ineffective unstructured interview into acceptable alternatives in terms of predictive value, political validity (i.e. acceptability of the selection tools/procedures for important stakeholders), reliability and fairness. A structured interview is a relatively long (single) interview usually involving several interviewers. These interviewers have a blueprint for each interview, asking similar questions to each applicant (8). In MMIs, each applicant passes through a number of independent stations with interviews, simulations and (sometimes) group exercises. There is a blueprint which states which stations are focusing on which specific skill or competency, and each station typically involves one interviewer and possibly one actor simulating a patient, peer, supervisor, etcetera. The biggest downside of MMIs is the fact that the logistics are complex and the costs are high (13).

Situational Judgement Tests (SJTs) are tests in which applicants are confronted with life-like, job-relevant scenarios, and have to respond to those scenarios by choosing an answer from a list of possible responses. There are many options when designing an SJT, for example with respect to manner of presenting the scenario (e.g. video-based or text-based), response format and response instructions (e.g. knowledge-based versus behavioral tendency; (8, 17-19). Research has provided positive results with regard to the use of SJTs in selection procedures, showing predictive and incremental value. Although SJTs can be costly to design, they can be used to assess many applicants at once, they are flexible and may be mapped to organizational values (13).

The next group of tools are the aptitude tests. The evidence for aptitude tests is mixed, as this group consists of many different individual tests all of which focus on different aptitudes. Well-known examples are the Graduate Medical School Admissions Test (GAMSAT), Biomedical Admissions Test (BMAT) and UK Clinical Aptitude Test (UKCAT, now called University Clinical Aptitude Test [UCAT] to reflect that it is used in countries other than the UK). For some aptitude tests, like the UKCAT, relatively strong evidence exists, supporting their use in high-stakes selection contexts (20, 21). However, for others, no such support exists (8). The core problem with the

underperforming aptitude tests seems to be the lack of blueprinting to appropriate aptitudes. More information on this subject is provided in section 1.3 ('Selection content').

The next two tools, unstructured autobiographical submissions and references, are selection tools that have a longstanding history and are still used, mostly because applicants and/or assessors feel like they work since they are familiar to them. However, there is no evidence to support their use in a high-stakes situation like selection. Autobiographical submissions and references are inappropriate for high-stakes selection because they are likely to be biased: they are prone to both intentional and unintentional distortions. It is known that these tools very rarely predict any outcome as they are almost always very positive about applicants, making them indistinguishable, and they increase bias against lower SES applicants (8).

Lastly, personality tests: at this point, there is no clear image of how personality and medical education and practice are related. For example, extraversion seems to have a positive relationship with performance at the beginning at medical school, but a negative relationship has been found in later stages of medical school (8). In 2016, MacKenzie et al. conducted a large-scale study on the relationship between performance on non-cognitive selection tests (including personality assessments) and medical school exit assessments. They found limited predictive value of personality and incongruence between their results and theoretical expectations. Therefore, MacKenzie et al. warn against the use of personality tests in selection procedures (22). The finding of limited predictive value, with unknown relationships between different personality traits and the various stages of medical school, and concerns regarding 'faking good' resonates in meta-analyses (10, 13). Furthermore, a recent study showed that applicants engaged in substantial response distortion on personality tests used in high-stakes situations (23). At this point, it seems that personality cannot reliably and validly be used in the high-stakes context of selection.

All in all, there is no one 'holy grail': all tools have strengths and weaknesses. This means that a combination of tools should be used, in order for the different tools to compensate each other's weaknesses, in a manner much resembling programmatic assessment (24). However, this combination should not be done arbitrarily, but based on what was intended to be measured: the content should determine the tools (24, 25). Therefore, we now turn to a reflection on how to establish content of a selection procedure. After that, an example is discussed of how content can be combined with tools, keeping in mind their psychometric qualities as shown in Table 1.1.

1.3 Selection content

Many schools struggle with the question of what combination of tools to use to ensure that all desirable cognitive and (inter)personal qualities are assessed (26). Our observation is that, on a local level, the choice of selection tools is often rooted in tradition, resource concerns (money, man-power and logistics) and/or essential but narrow criteria, such as reliability and predictive validity (8, 10, 12, 27). Patterson, in

several different articles (e.g. 11, 12, 13, 15, 28), has set forward a possible solution: using a job analysis to determine which competencies are needed in the intended job and applying those in the selection procedure. The job content will, in turn, determine which tools should be used. Applying job analyses to establish content is currently seen as best practice in the field of selection (8).

In Patterson's recent book (8), Kerrin et al. (29) defined role analysis (used interchangeably with job analysis) as: "*a systematic process for collecting and analyzing job-relevant information; outputs of which provide a framework of the important KSAOs [knowledge, skills, abilities, and other characteristics] required for both selection into training and subsequent performance in clinical practice, and can be used to identify, and prioritize, role-specific selection criteria*" (Kerrin, 2018; page 141; 25). Preparing a selection procedure by conducting a job analysis beforehand results in an accurate blueprint with effective criteria, which will be the main determinant of the procedure's validity. This is in sharp contrast with previous research focusing on tools as the main determinants for reliability and validity. Kerrin et al. emphasize that the choice of a tool is important, but that this should be preceded by the development of a blueprint (29).

There are many different ways in which job analyses may be conducted, and as in many best practices, the preferred way is to combine different methods with different advantages and disadvantages. Examples of methods that can be used for a job analysis are: literature reviews, interviews, stakeholder consultations and behavioral observations. For more practical information on how to conduct a job analysis, the reader is referred to Kerrin et al.'s chapter (29). In the current thesis, a possible integration of the evidence related to the tools and the results of job analyses is set forward. I now turn to this example.

1.4 The Maastricht example

The current thesis revolves around research conducted in the medical curriculum at Maastricht University. Maastricht University has developed a selection procedure with the goal to address the aforementioned problem of combining tools of high psychometric value with content stemming from a job analysis. This procedure is an outcome-based approach to selection: defining the competencies of a 'good doctor' (i.e. job analysis) and using backward chaining (i.e. working backwards from the goal) to turn these competencies into a selection procedure (27, 30). Indeed, developing a multi-tool, outcome-based approach selection with a clear competency-blueprint is aligned with the global move towards competency-based approaches to preparing the next generation of health professionals (30, 31)

The competencies that should be assessed in the context of medical school selection can be derived from outcome frameworks. Outcome frameworks can be seen as extensive job analyses. The frameworks describe the competencies and expertise that medical students must achieve by graduation to ensure that they have acquired the basics for being good doctors and meeting patient/healthcare needs (examples of

outcome frameworks: 32, 33, 34). Different frameworks are used worldwide, but they share analogous objectives and differ mostly in level of detail, context and terminology (35). As a result of this commonality, backward chaining from one exemplary framework into an outcome-based selection procedure will be broadly relevant across medical schools. In the medical curriculum at Maastricht University, the outcome framework applied was the internationally-recognized competency framework CanMEDS (Canadian Medical Education Directives for Specialists; 34) and its Dutch derivative, the 'Raamplan Artsopleiding 2009' (33). They were seen as the outcomes of a job analysis. Therefore, the blueprint of the selection procedure was based on these frameworks.

Importantly, the context in which the selection procedure is applied should be taken into account, e.g. undergraduate versus graduate selection, learning environment, and other contextual factors of importance to the institution (see Figure 1.1). The outcome framework applied in the medical curriculum at Maastricht University describes the end terms of medical school, a level the applicants have not yet achieved. Therefore, a team of Subject Matter Experts (i.e. the selection committee, consisting of experts in education and medicine) translated the CanMEDS-competencies into so-called derived competencies applicants could already possess at bachelor entry-level, and which could be measured in a selection procedure. Another important piece of context at Maastricht University is the Problem-Based Learning (PBL) system which is applied throughout the university. Therefore, knowledge of and fit with PBL was also taken into account during the selection procedure.

The translation of the outcome framework into derived competencies, to make sure there was a fit with the context (undergraduate selection and PBL), took place by first thoroughly inspecting the CanMEDS competencies (i.e. Medical Expert, Communicator, Collaborator, Organizer [Leader in the 2015 edition], Health Advocate, Scholar and Professional; (34, 36). Hereafter, clinical and medical school related situations representative of these competencies were gathered to further inform the more concrete content of the blueprint. In an iterative process, the selection committee discussed the so-called derived competencies (i.e. the competencies resulting from the translation of the CanMEDS competencies), how they should be defined and to what extent they should be measured. The resulting derived competencies were: Transfer (i.e. knowledge and information integration), Textual skills, Reasoning, Communication, Collaboration, Organization, Medical and Societal Consciousness, Ethical awareness, Empathy and Reflection. The goal of the selection procedure for the medical curriculum at Maastricht University was to measure these derived competencies.

The next step was to decide how these competencies could be assessed in a selection procedure. As stated before, there is no tool that can be seen as the 'holy grail' of selection. All tools have their advantages and disadvantages, and therefore it is paramount to combine multiple assessments (i.e., tools), scored by multiple assessors, preferably at multiple independent points in time (30, 31). Therefore, a two-phase

selection procedure was devised, consisting of three tools, as assessed by members of the selection committee (8 Subject Matter Experts).

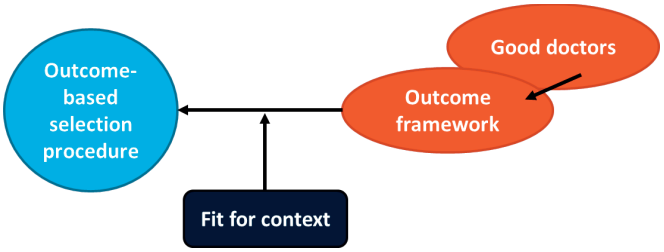


Figure 1.1: Visual representation of the use of backward chaining from the desired end goal ('good doctors') to create an outcome-based selection procedure.

The first round focuses on both general academic (including PAA) and more (inter)personal (distinguishing skills; e.g. defensible experience with Collaboration or Organization) aptitude, as well as fit with and knowledge about PBL. This information is gathered in a portfolio format, as this is a flexible format also applied throughout the medical curriculum at Maastricht University (although in a different manner). Furthermore, applicants had to be able to fill it out at home in their own pace. The portfolio used in the selection procedure for the medical curriculum at Maastricht University is different from the autobiographical submissions explained by Cleland (10), because of the pre-structured and factual nature of the Maastricht portfolio and the high burden of proof associated with it. Furthermore, the first round is seen as a broad brush. About double the amount of students to be admitted to medical school through selection is admitted to the second round of the selection procedure.

The second round in itself consists of two tests, one focused mainly on several more (inter)personally loaded competencies and one focused more broadly on all 'derived competencies'. For the more (inter)personal competencies, the SJT-format was chosen, as SJTs seem capable of measuring these types of competencies (37, 38) and they have shown high reliability and moderate to high validity (see Table 1.1) and predictive value in actual clinical practice (39), while also showing a moderate 'promotion' of widening access (10, 13). Video-based SJTs have been shown to make the situation more believable and acceptable to the applicants and therewith appear to make the results more valid (12); therefore, a Video-based SJT (V-SJT) was designed, assessing Collaboration, Medical and Societal Consciousness, Ethical awareness, Empathy and Reflection. In order to also assess the more cognitive side of the competency-spectrum, an aptitude test was included in the second round. As stated before, aptitude tests have obtained varying results in terms of reliability and validity. Importantly, the evidence supporting the reliability and validity of the test increases if there is a specific and clear blueprint to be followed, based on thorough job analysis (29). As there was a clear blueprint for the selection procedure for the medical curriculum at Maastricht University, an aptitude test (the Written Aptitude Test; WAT) was devised, measuring Transfer, Textual skills, Reasoning, Organization, Medical and Societal Consciousness, Ethical awareness and Reflection. As the majority of the

assignments in the SJT as well as the WAT include open-ended questions, the competency Communication was assumed to play a role in any and all assignments, and was not assessed independently.

The selection procedure as described above is new in several ways. A job analysis formed the base for a detailed blueprint. Importantly, the job analysis the blueprint was based on is also the framework to which the entire medical curriculum at Maastricht University was blueprinted. Although other universities also employ procedures which are outcome-based, the idea of purposeful and consistent backward chaining of those outcomes and considering the context before creating the procedure is still relatively new. Furthermore, the tools were chosen based on the content they were intended to measure as well as their evidence-base, and methodically combined into a selection procedure. An important addition, from a research point-of-view, is the situation in the Netherlands. When Maastricht University started selecting for its medical curriculum in 2011, not all of its students could be admitted through selection due to legal restrictions; a national weighted lottery was in place as well. This lottery was weighed based on the applicants' high school Grade Point Average (GPA): applicants with a GPA ≥8 were always admitted, while applicants with GPAs of 7.5 until 8, 7 until 7.5, 6.5 until 7, and 6 until 6.5 were admitted in the ratio 9:6:4:3. Applicants who were rejected in the selection procedure automatically participated in the national weighted lottery procedure, providing them with a 'second chance' to be admitted to medical school. There was also a possibility of entering the national weighted lottery without first having participated in the selection procedure. This means that our student cohorts starting in 2011, 2012 and 2013 (from 2014 on, all cohorts were fully selected) consisted of selected students, students who were rejected in the selection procedure but got in through the national lottery, and students who did not participate in the selection procedure but still got in through lottery (see Figure 1.2).

Because the manner in which the selection procedure was devised was new, an extensive investigation was needed to find out whether it fulfilled our expectations: is this setup resulting in the selection of the most promising students and starting doctors, does the procedure really measure what we intended it to measure, and is it cost-effective? The most fundamental research question is that of validity: is the level of validity evidence for this selection setup sufficiently high for high-stakes decisions related to admission to medical school? In the Standards for educational and psychological testing, issued by the AERA, APA and NCME (40), validity is defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests". Therefore, we now turn to the role of validity in selection research.

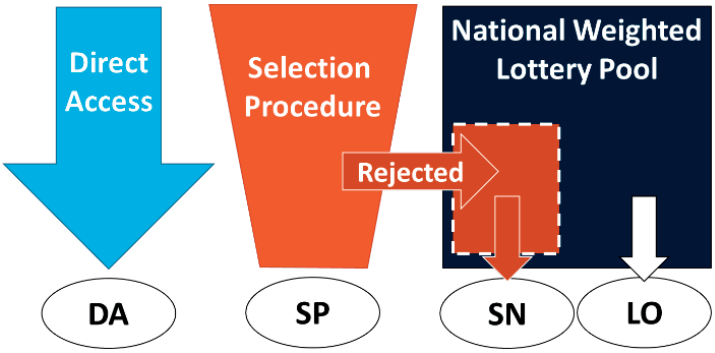


Figure 1.2: Routes of admission to the medical curriculum at Maastricht University in 2011-2013. SP = Selection Positive (i.e. selected students), SN = Selection Negative (i.e. rejected in the selection procedure but admitted through national weighted lottery) and LO = Lottery (i.e. did not participate in the selection procedure but admitted through the national weighted lottery).

1.5 The role of validity in selection

Previous research on validity in selection has focused mainly on the predictive validity (to what extent a predictor is able to predict an outcome) and incremental validity of one tool over another (whether one tool has predictive value over and beyond another, whether they add value to each other; 20, 41). These are very narrow considerations of validity; meanwhile, the field of validity has progressed towards 'modern' validity theories (42). Lately, more and more calls have been published asking for research using this modern type of validity. Kulasegaram, for example, stated that "Increasingly, the concept of validity is being used in sophisticated ways in health professions education" (43). This was echoed by Kreiter: "measuring [motivation, persistence, self-esteem] and other non-academic dimensions with a high-stakes assessment will require new types of validity evidence to establish whether it meets standards of fairness that are required for all psychological tests that are used to make high-stakes admission decisions" (44). However, in selection, these calls have not yet been answered.

The modern validity theories Kulasegaram and Kreiter refer to are Kane's validity framework (45, 46), Messick's 'validity of psychological assessment' framework (47), 'the Standards of Educational and Psychological Measurement' (40), and Downing's framework on 'meaningful interpretation of assessment data' (48). Although the focus of these validity frameworks can be somewhat different, they mostly overlap conceptually (42, 49). Royal (42) looked at the overlap between these modern validity frameworks, and found four main similarities between them. First, validity refers to inferences, not instruments. This implies that an instrument cannot be studied in one context and be generalized to another without consideration of the differences between the contexts. For example, if we find our selection procedure to be useful for the medical curriculum at Maastricht University, this does not mean it can be applied at our School of Business and Economics unaltered. Instead, validity can only be investigated using an instrument in a specific context and then carefully generalized to

similar contexts, but never without thorough consideration. Second, the uniform conceptualization of validity. Previously, validity was fragmented and there were many different 'kinds' of validity. In modern validity theory, the old 'construct validity' is seen as the overarching type of validity, and evidence pertaining to all other 'types' of validity may be used to support the general construct validity. Third, validity is a continuum. Previously, validity was seen as something an instrument either possesses or it does not. The modern theories have not only moved from the instrument to the inference, but also state that there can be any amount of support in terms of validity. In other words, validity is not black-and-white, but a "continuum onto which cumulative evidence is weighed and judged to support an inference" (38). Importantly, different aspects of validity may also be supported to a greater extent than others, and different researchers may find different results. This leads to the fourth similarity, validation is an ongoing process. "Recognizing the complex nature of latent trait measurement, most validity theorists contend that validation is an ongoing process because multiple factors … are subject to change" (38). Therefore, validation is now defined as a 'never-ending process', in which evidence is gathered, revisited, combined and replicated continually. Applying these new theories on validity will expand the theory-base of selection, which also has been called for recently (13, 44).

Looking at the specific validity theories, there is almost complete overlap between Downing (48) and the Standards (40), and both closely follow Messick's ideas (47) on validity. Instead of general sources of evidence for validity, Kane focused on how evidence should be prioritized within a validity argument (50): focus on the weakest assumption in the hypothesis (in this case the validity argument) and gather evidence supporting that assumption. On their own, each individual validity framework proved insufficiently practical to investigate our selection procedure thoroughly. Therefore, we synthesized them and supplemented this synthesis with the COSMIN checklist (51) and the related but more elaborate book 'Measurement in medicine' (52) to create an integrated framework applicable to selection research. Our aim with this synthesis was to provide a comprehensive understanding of validity that is relatively easily applied to selection. Downing attempted to pull the abovementioned contemporary views on validity into the medical education assessment research (48). Therefore, this framework was deemed most applicable to selection in the medical education context, and our synthesis of the different frameworks was grounded in Downing's work.

The synthesized validity framework is shown in Figure 1.3. Supportive to this figure is a toolbox which contains manners in which to support the sources of evidence for validity, which is set forth in the Appendix (at the very end of this dissertation). This toolbox is intended to be easily applicable; any researcher investigating validity should be able to choose and apply the tools most appropriate for their question. Of course, not all tools are appropriate for every study or assessment, so evidence should be prioritized, as set forth by Kane (45, 50).

As shown in Figure 1.3, before any source of validity can be investigated, the **proposed use** of the assessment (for the current dissertation, the selection procedure for the medical curriculum at Maastricht University) should clearly be defined. For evidence based on the **content** of the measurement under study, information should be gathered showing that the test domain actually reflects the whole of the domain that is to be measured. Scoring integrity should be considered in order to gather evidence based on **response processes**. This is a relatively broad source of evidence, it contains information on how data is handled, how scores are computed and combined, but also evidence showing that in order to respond to items, the competencies the item is supposed to measure are in fact needed to successfully complete the item (for this, process measures can be used, such as think-aloud protocols, eye-tracking and trace data). Psychometric evidence concerning, for example, reliability and dimensionality may be gathered in support of **internal structure**. Next, evidence that links the test performance to other performances and (preferably) real-world performance supports the construct validity with **relations to other variables**. Lastly, evidence related to the **consequences** of the measurement may be assessed, which is mostly defined as the implications the measurement has for the different stakeholders.

1.6 The quest for validity in selection

In this dissertation, we took on the quest for validity for a new setup of selection for medical school. The selection procedure employed for the medical curriculum of Maastricht University is relatively new, which means we needed to find out whether it fulfilled our expectations. Therefore, as mentioned before, the current thesis is centered around the validation process of this new selection setup. Herein, not only the more established subjects of effectiveness and incremental value are assessed, but also the heavily under-researched topics of validity based on content, internal structure and consequences. The steps taken on this quest (i.e. the studies that were conducted) are shortly introduced below.

Combined, the empirical chapters of this thesis looked at four of the five sources of evidence for validity, according to the model explained above (see Figure 1.3). In **chapter two** and **three**, the relation between the applicants' performance during the selection procedure and their performance as a student throughout the first (pre-clinical) and second (clinical) phase of medical school is investigated, thus gathering information of the selection procedure's relation to other variables. In both chapters, the unique situation in the Netherlands, in which decentralized (i.e. university-specific) selection procedures ran in parallel with a national GPA-weighted lottery procedure (see Figure 1.2), was exploited. In chapter two, the selected (selection positive or SP) students were compared to the rejected (selection negative or SN) students on cognitive, (inter)personal, mixed and general outcomes throughout the first three years of medical school (i.e. the pre-clinical bachelor program). In chapter three, the performance of SP and SN students in the clinical master phase of medical school (assessed using the CanMEDS roles) is compared. Hence, this study investigates whether the alignment of selection with the curriculum and outcome framework

actually results in students performing better on the required competencies in the clinical workplace.

In chapter four, two sources of evidence for validity of selection were investigated: the content and internal structure. The main goal of this study was to investigate whether internal structures of the tests in the second round of the selection procedure (the V-SJT and the WAT) reflect the content that was intended to be measured. Content and internal structure were taken together as they are intrinsically related to a large extent. The manner in which the content was defined and converted into a test defines the internal structure. To gather evidence relating to the content of the procedure, the manner in which the selection procedure as a whole was created is explicated. To assess the internal structure of the selection tools applied in the selection procedure for the medical curriculum at Maastricht University, well-known test theories were considered to be insufficient. Therefore, a new theory, Cognitive Diagnostic Modeling, was applied to investigate whether the competencies that each item was intended to measure were indeed measured, as reflected by the performance of applicants in those items.

In chapter five, the focus shifted towards the consequences of selection as a source of validity, specifically looking at the cost-effectiveness of the procedure. In this study, an analysis of costs versus benefits of the selection procedure contrasted with the lottery procedure was conducted. In short, all economic costs and benefits were gathered for the decentralized (effortful) selection procedure as well as the relatively inexpensive lottery procedure (a GPA-based process), as these two procedures would be the most likely options for admission to medical school in the Netherlands. Costs were defined as the costs that need to be borne by the medical school. The benefits were defined as possible decreases in later costs (due to drop-out, repetition of courses or resource-intensive tests). These costs and benefits were then taken together and compared for the two distinct admissions procedures.

In chapter six, a commentary to an article by Jonathan Foo and colleagues (53) is included. Foo et al. concluded that the cost evaluation studies conducted in the field of Health Professions Education (HPE) are rare and often of substandard quality. Importantly, the amount of studies (relative to the total amount of studies within HPE in general) has not increased over the last 15 years, and the quality did not improve either. Furthermore, studies looking at cost are already rare, but studies also including value are almost non-existent. In order for stakeholders to make well-founded decisions, this information is paramount. We looked at this situation within the context of selection for medical school, where the situation of cost evaluations is equally dire, and looked at the implications for selection research and the role of validity.

Lastly, in chapter seven, all findings are discussed in terms of the validity framework established in this General Introduction and compared to other results in the field of

selection so far. Importantly, implications and suggestions for the practice and future research of selection in the Netherlands as well as internationally are debated.

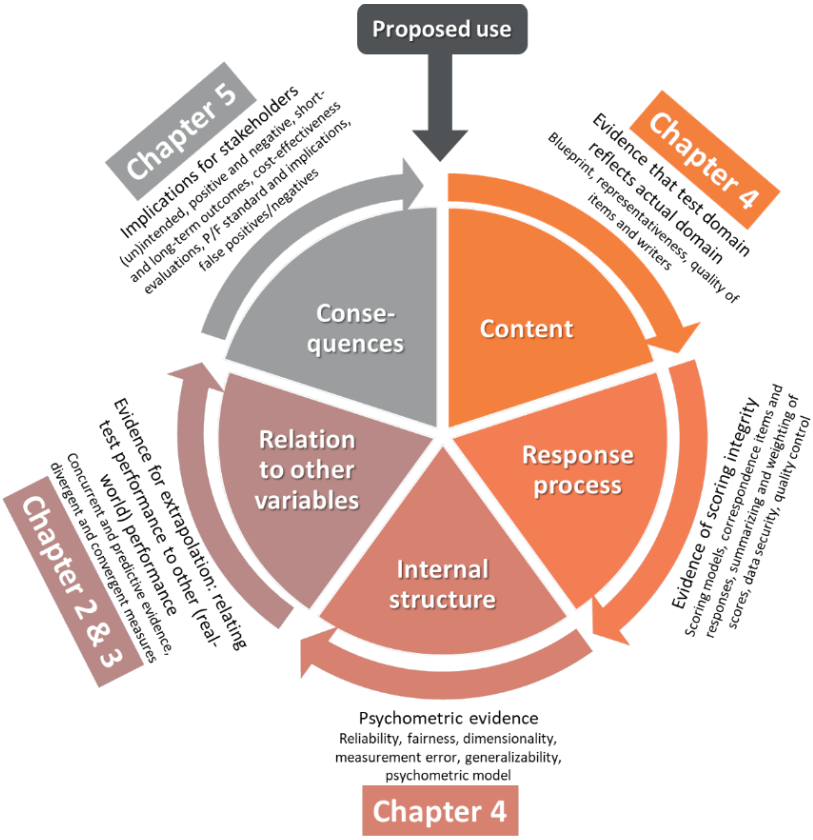


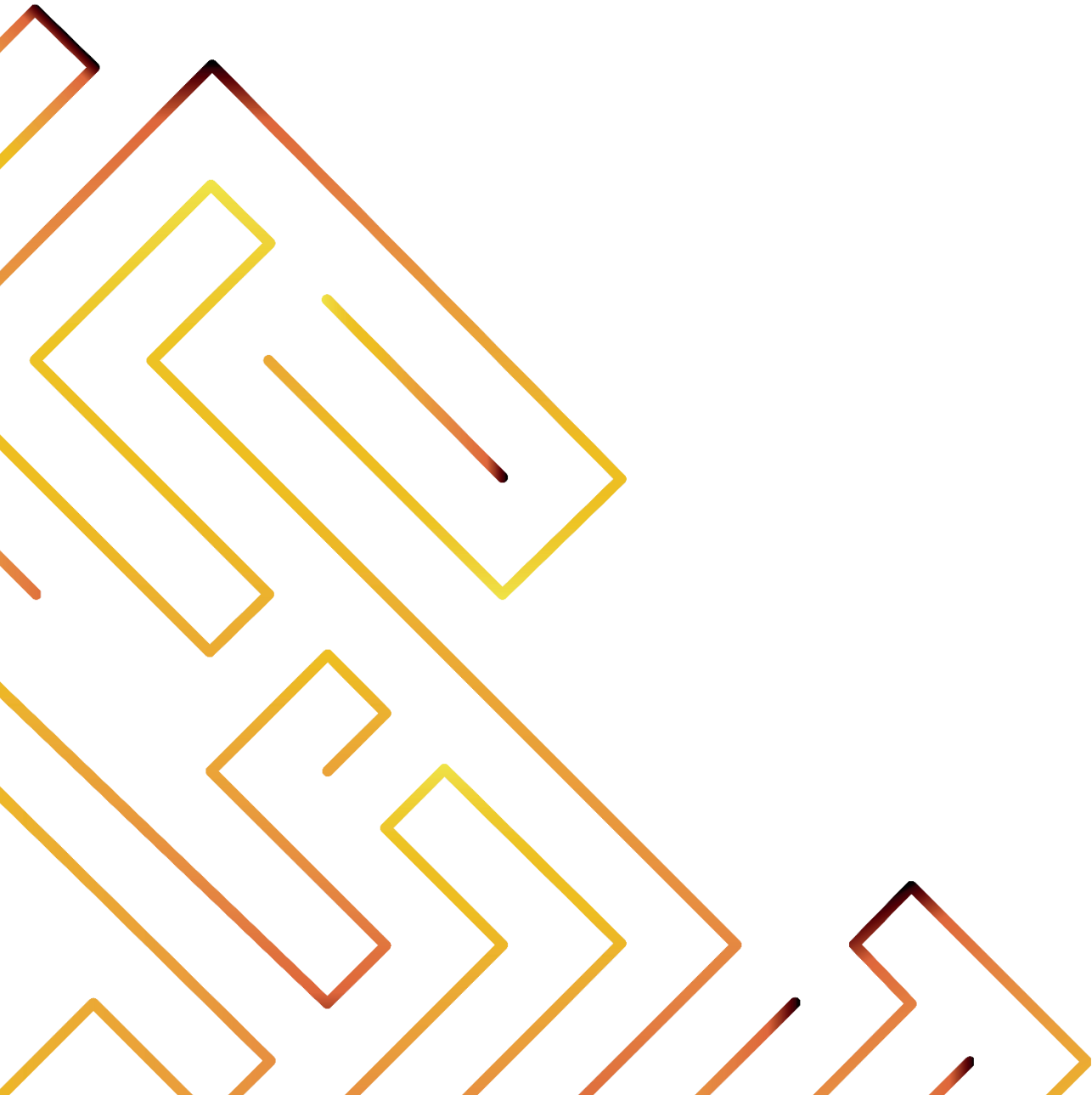
Figure 1.3: Overview of the validity framework used in this dissertation, grounded in the Downing framework, and concretized and adapted for use on selection for medicine

1.7 References

1. Burch VC. Medical school admissions: where to next? Adv Health Sci Educ Theory Pract. 2009;14(2):153-7.
2. Whitehouse C. Pre-medicine and selection of medical students. Med Educ. 1997;31 Suppl 1:3-6.
3. Lucieer SM. Selecting students for medical education: Exploring novel approaches [Dissertation]. Rotterdam: Erasmus University Rotterdam; 2016.
4. Flexner A. Medical education in the united states and canada: A report to the carnegie foundation for the advancement of teaching. New York City: The Carnegie Foundation for the Advancement of Teaching; 1910.
5. Siu E, Reiter HI. Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. Adv Health Sci Educ Theory Pract. 2009;14(5):759-75.
6. Humphreys LG. Intelligence: Three kinds of instability and their consequences for policy. Intelligence. 1989:193-216.
7. Papadakis MA, Teherani A, Banach MA, Knettler TR, Rattner SL, Stern DT, et al. Disciplinary action by medical boards and prior behavior in medical school. N Engl J Med. 2005;353(25):2673-82.
8. Patterson F, Zibarras L, editors. Selection and Recruitment in the Healthcare Professions: Research, theory and practice. Cham, Switzerland: Springer Nature Switzerland AG; 2018.
9. Roddie IC. The Edinburgh Declaration. The Lancet. 1988;332(8616):908.
10. Cleland J, Dowell J, McLachlan J, Nicholson S, Patterson F. Identifying best practice in the selection of medical students (literature review and interview survey). https://www.gmc-uk.org/-/media/about/identifyingbestpracticeintheselectionofmedicalstudentspdf51119804.pdf; 2012.
11. Patterson F, Ferguson E, Knight AL. Selection into medical education and training. In: Swanwick T, editor. Understanding Medical Education2013.
12. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ. 2016;50(1):36-60.
13. Patterson F, Roberts C, Hanson MD, Hampe W, Eva K, Ponnamperuma G, et al. 2018 Ottawa consensus statement: Selection and recruitment to the healthcare professions. Med Teach. 2018;40(11):1-11.
14. Prideaux D, Roberts C, Eva K, Centeno A, McCrorie P, McManus C, et al. Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):215-23.
15. Patterson F, Cleland J, Cousans F. Selection methods in healthcare professions: where are we now and where next? Adv Health Sci Educ Theory Pract. 2017;22(2):229-42.
16. White CB, Dey EL, Fantone JC. Analysis of factors that predict clinical performance in medical school. Adv Health Sci Educ Theory Pract. 2009;14(4):455-64.
17. de Leng WE, Stegers-Jager KM, Born MP, Themmen APN. Integrity situational judgement test for medical school selection: judging 'what to do' versus 'what not to do'. Med Educ. 2018;52(4):427-37.
18. De Leng WE, Stegers-Jager KM, Born MP, Themmen APN. Influence of response instructions and response format on applicant perceptions of a situational judgement test for medical school selection. BMC Med Educ. 2018;18(1):282.
19. De Leng WE, Stegers-Jager KM, Husbands A, Dowell JS, Born MP, Themmen APN. Scoring method of a Situational Judgment Test: influence on internal consistency reliability, adverse impact and correlation with personality? Adv Health Sci Educ Theory Pract. 2017;22(2):243-65.
20. Tiffin PA, Mwandigha LM, Paton LW, Hesselgreaves H, McLachlan JC, Finn GM, et al. Predictive validity of the UKCAT for medical school undergraduate performance: a national prospective cohort study. Bmc Med. 2016;14(1):140.
21. MacKenzie RK, Cleland JA, Ayansina D, Nicholson S. Does the UKCAT predict performance on exit from medical school? A national cohort study. Bmj Open. 2016;6(10):e011313.
22. MacKenzie RK, Dowell J, Ayansina D, Cleland JA. Do personality traits assessed on medical school admission predict exit performance? A UK-wide longitudinal cohort study. Adv Health Sci Educ Theory Pract. 2017;22(2):365-85.
23. Anglim J, Bozic S, Little J, Lievens F. Response distortion on personality tests in applicants: comparing high-stakes to low-stakes medical settings. Adv Health Sci Educ Theory Pract. 2018;23(2):311-21.
24. Van Der Vleuten CPM, Schuwirth LWT, Driessen EW, Govaerts MJB, Heeneman S. Twelve Tips for programmatic assessment. Med Teach. 2015;37(7):641-6.

25. Pearce J, Jackel B. SJT, MCQ, ETC… The worrying conflation of format and content. Med Educ. 2018;52(9):993-.
26. Cleland J, Dowell J, Nicholson S, Patterson F. How can greater consistency in selection between medical schools be encouraged? A project commissioned by the Selecting for Excellence Group (SEEG)2014. Available from: http://www.medschools.ac.uk/SiteCollectionDocuments/Selecting-for-Excellence-research-Professor-Jen-Cleland-et-al.pdf
27. Wilkinson TM, Wilkinson TJ. Selection into medical school: from tools to domains. BMC Med Educ. 2016;16(1):258.
28. Patterson F, Ferguson E, Thomas S. Using job analysis to identify core and specific competencies: implications for selection and recruitment. Med Educ. 2008;42(12):1195-204.
29. Kerrin M, Mossop L, Morley E, Fleming G., Flaxman C. Role Analysis: The Foundation for Selection Systems. In: Patterson F, Zibarras L, editors. Selection and Recruitment in the Healthcare Professions: Research, theory and practice. Cham, Switzerland: Springer Nature Switzerland AG; 2018.
30. Frank JR, Snell L, Englander R, Holmboe ES, Collaborators I. Implementing competency-based medical education: Moving forward. Med Teach. 2017;39(6):568-73.
31. Holmboe ES, Sherbino J, Englander R, Snell L, Frank JR, Collaborators I. A call to action: The controversy of and rationale for competency-based medical education. Med Teach. 2017;39(6):574-81.
32. Holmboe ES, Edgar L, Hamstra S. The Milestones Guidebook. In: ACGME, editor. Chicago, IL.2016.
33. Van Herwaarden CLA, Laan RFJM, Leunissen RRM. Raamplan Artsopleiding 2009. NFU, editor. Utrecht2009.
34. Frank JR. The CanMEDS 2005 physician competency framework: Better standards, better physicians, better care. 2005. Available from: http://www.ub.edu/medicina_unitateducaciomedica/documentos/CanMeds.pdf.
35. Hautz SC, Hautz WE, Feufel MA, Spies CD. Comparability of outcome frameworks in medical education: Implications for framework development. Med Teach. 2015;37(11):1051-9.
36. Frank JR, Snell LS, Sherbino J. The draft CanMEDS 2015 physician competency framework - series ii. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2014.
37. Patterson F, Ashworth V, Kerrin M, O'Neill P. Situational judgement tests represent a measurement method and can be designed to minimise coaching effects. Med Educ. 2013;47(2):220-1.
38. Patterson F, Ashworth V, Zibarras L, Coan P, Kerrin M, O'Neill P. Evaluations of situational judgement tests to assess non-academic attributes in selection. Med Educ. 2012;46(9):850-68.
39. Cousans F, Patterson F, Edwards H, Walker K, McLachlan JC, Good D. Evaluating the complementary roles of an SJT and academic assessment for entry into clinical practice. Adv Health Sci Educ Theory Pract. 2017;22(2):401-13.
40. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, United States of America: American Educational Research Association; 2014.
41. McManus IC, Dewberry C, Nicholson S, Dowell JS. The UKCAT-12 study: educational attainment, aptitude test performance, demographic and socio-economic contextual factors as predictors of first year outcome in a cross-sectional collaborative study of 12 UK medical schools. Bmc Med. 2013;11:244.
42. Royal KD. Four tenets of modern validity theory for medical education assessment and evaluation. Adv Med Educ Pract. 2017;8:567-70.
43. Kulasegaram K. Use and ornament: expanding validity evidence in admissions. Adv Health Sci Educ Theory Pract. 2017;22(2):553-7.
44. Kreiter CD. A research agenda for establishing the validity of non-academic assessments of medical school applicants. Adv Health Sci Educ Theory Pract. 2016;21(5):1081-5.
45. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 2015;49(6):560-75.
46. Kane MT. Current concerns in validity theory. Journal of Educational Measurement. 2001;38(4):319-42.
47. Messick S. Validity of Psychological-Assessment - Validation of Inferences from Persons Responses and Performances as Scientific Inquiry into Score Meaning. American Psychologist. 1995;50(9):741-9.
48. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-7.

49. St-Onge C, Young M, Eva KW, Hodges B. Validity: one word with a plurality of meanings. Adv Health Sci Educ Theory Pract. 2017;22(4):853-67.
50. Kane MT. An Argument-Based Approach to Validity. Psychological Bulletin. 1992;112(3):527-35.
51. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19(4):539-49.
52. De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide: Cambridge University Press; 2011.
53. Foo J, Cook DA, Walsh K, Golub R, Abdalla ME, Ilic D, et al. Cost evaluations in health professions education: a systematic review of methods and reporting quality. Med Educ. 2019.



CHAPTER 2

Selection into medicine: The predictive validity of an outcome-based procedure

Schreurs S, Cleutjens KB, Muijtjens AMM, Cleland J, oude Egbrink MGA. Selection into medicine: the predictive validity of an outcome-based procedure. *BMC Medical Education*. 2018; 18(1):214

Abstract

Background

Medical schools must select students from a large pool of well-qualified applicants. A challenging issue set forward in the broader literature is that of which cognitive and (inter)personal qualities should be measured to predict diverse later performance. To address this gap, we designed a 'backward chaining' approach to selection, based on the competencies of a 'good doctor'. Our aim was to examine if this outcome-based selection procedure was predictive of study success in a medical bachelor program.

Methods

We designed a multi-tool selection procedure, blueprinted to the CanMEDS competency framework. The relationship between performance at selection and later study success across a three-year bachelor program was examined in three cohorts. Study results were compared between selection-positive and selection-negative (i.e. primarily rejected) students.

Results

Selection-positive students outperformed their selection-negative counterparts throughout the entire bachelor program on assessments measuring cognitive (e.g. written exams), (inter)personal and combined outcomes (i.e. OSCEs). Of the 30 outcome variables, selection-positive students scored significantly higher in 11 cases. Fifteen other, non-significant between-group differences were also in favor of the selection-positives. An overall comparison using a sign test indicated a significant difference between both groups ($p < 0.001$), despite equal pre-university GPAs.

Conclusions

The use of an outcome-based selection approach seems to address some of the predictive validity limitations of commonly-used selection tools. Selection-positive students significantly outperformed their selection-negative counterparts across a range of cognitive, (inter)personal, and mixed outcomes throughout the entire three-year bachelor in medicine.

2.1 Background

As there are many more applicants than places, medical schools need to select students from a large pool of suitably qualified candidates. Schools must also ensure they admit those candidates most likely to succeed and, crucially, become good doctors [1–3]. A number of important issues influence selection for admission [3, 4]. One of these is ensuring that selection tools assess the attributes considered important by key stakeholders, including patients. Traditionally, selection into medical school was solely based on prior academic attainment. Currently, there is increasing recognition that broader criteria are required, as there is more to being a capable medical student or doctor than academic performance [5–7]. Most medical schools now aim to select applicants who are both academically capable and also possess (inter)personal skills befitting a career in medicine, such as team-working and communication skills [8, 9].

Developing a selection procedure that can fairly and accurately discriminate between applicants, based on academic as well as (inter)personal criteria, is challenging [10–13]. Many schools struggle with the question of what combination of tools to use to ensure that all desirable academic and (inter)personal qualities are assessed [14]. Our observation is that, on a local level, the choice of selection tools is often rooted in tradition, resource concerns and/or essential but narrow criteria, such as psychometric qualities [1, 2, 15]. In addition, different selection tools are better at predicting different outcomes. For example, tools measuring cognitive abilities (e.g. Grade Point Average, GPA) seem better at predicting academically-loaded assessments in the earlier years of medical school [2, 16], whereas '(inter)personal' assessments (e.g. Multiple Mini Interviews, MMIs, and Situational Judgement Tests, SJTs) seem better at predicting more clinically-oriented performance in the later years of medical education [1]. Cognitive and (inter)personal assessments have been integrated in some tools, but the predictive value of these integrated tools is moderate at best [1, 2, 9].

One potential way to address the aforementioned dilemmas is to develop a more holistic and outcome-based approach to selection into medical school. One way of doing this is to define the competencies of a 'good doctor' and use these as the basis of a selection procedure [15, 17]. These competencies can be derived from outcome frameworks, which describe the competencies and expertise that medical students must achieve by graduation to ensure that they have acquired the basics for being good doctors and meeting patient/healthcare needs (examples of outcome frameworks: 18–20). Different frameworks are used worldwide, but they share analogous objectives and differ mostly in level of detail, context and terminology [12]. As a result of this commonality, 'backward chaining' (i.e. working backwards from the goal) from one exemplary framework into an outcome-based selection procedure will be broadly relevant across medical schools. Furthermore, the context in which the selection procedure is applied should be taken into account, e.g. undergraduate versus graduate selection, learning environment, and other contextual factors of

importance to the institution (see Figure 2.1). The proposed procedure is in line with recently stated developments in competency-based medical education, where it is paramount to combine multiple assessments by multiple assessors. Indeed, developing a multi-tool, outcome-based approach selection blueprinted to a framework of competencies is aligned with the global move towards competency-based approaches to preparing the next generation of health professionals [17, 21]. However, before recommending multi-tool, outcome-based selection as the way forward, it is critical to examine whether this approach does indeed predict performance across competencies. Especially in current times of limited resources and increased accountability demands, it is important to employ an evidence-based selection procedure. Therefore, the aim of this study was to examine whether an outcome-based, holistic selection procedure is predictive of study success in a medical bachelor curriculum. The selection procedure as well as the curriculum and assessment program under study are aligned with the CanMEDS framework of competencies [20], which is used to define the qualifications for medical doctors in the Netherlands [19]. Due to the transition from lottery to selection that occurred during the period of study (see section 2.2), we had the unique opportunity to compare study results of students who were selected (selection-positives) to those of students who were rejected in the same selection procedure, but still got into medical school via the national weighted lottery (selection-negatives). Therefore, our concrete research question was: how does performance in a medical bachelor curriculum differ between students that were selected (selection-positives) or rejected (selection-negatives) in the same outcome-based selection procedure?

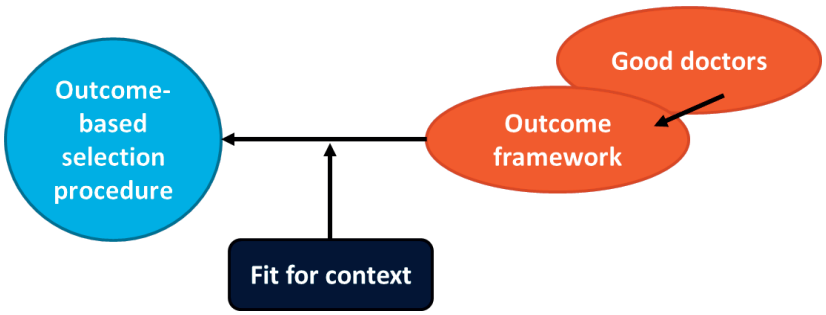


Figure 2.1: Visual representation of the use of backward chaining from the desired end goal ('good doctors') to create an outcome-based selection procedure

2.2 Methods

2.2.1 Context

This study was performed at Maastricht University Medical School (MUMS). As is typical in the Netherlands, MUMS comprises a three-year bachelor and three-year master phase. We focused on the bachelor phase, which encompasses a mix of theoretical and practical educational elements.

This study included three cohorts of students, starting in 2011, 2012 and 2013. In 2011 through 2013, 50 (2011) to 60% (2012 and 2013) of the available study places was assigned through the local, outcome-based selection procedure; this limitation was imposed by the national government. Remaining places were filled via the national weighted lottery, available to applicants who were rejected in the selection procedure or who did not participate in selection at all. This unique situation enabled comparison of selection-positive students' study outcomes with those of selection-negative (i.e. primarily rejected) students. The third group of students, who entered MUMS through lottery only (without participating in the selection procedure), was not included in the present study since their study outcomes could not be related to their performance in the selection procedure. Before 2011, all admissions into MUMS were assigned through the national weighted lottery, while from 2014 onwards MUMS transitioned to full selection of the cohorts. For more information on selection in the Netherlands, the reader is referred to Schripsema et al. [22].

2.2.2 Selection procedure

The selection procedure applied in 2011–13 consisted of two stages, both based on the CanMEDS framework of competencies (Table 2.1; 20, 23).

In the *first round* applicants completed a pre-structured online portfolio, which comprised four parts. The first part (worth 40% of the total score for the portfolio) was pre-university training (including pre-university GPA; pu-GPA). The second part (also 40%) was a description of previous extracurricular activities, requesting skills relevant for a medical student and/or doctor (e.g. communication, collaboration, organization, and professionalism). The last two parts, each worth 10% of the portfolio score, concerned knowledge of and opinion on the medical curriculum and the Problem-Based Learning (PBL) system at MUMS; these parts aimed at establishing the applicants' fit for context (Figure 2.1). Applicants were ranked according to the weighted average of scores for the four parts. A predetermined number of highest ranking applicants in the first round (twice the amount of places to be allotted via selection) were invited to the second round of the selection procedure. The scores for the first-round portfolio were not taken into account in the second round.

The *second round*, a selection day at MUMS, consisted of a Video-based Situational Judgment Test (V-SJT) and a combination of aptitude tests. The derived competencies based on the exemplary framework of competencies (CanMEDS; Table 2.1) formed the blueprint for the assignments in the second round; backward chaining was used to implement these competencies into the assignments. The V-SJT was based on the Computer-based Assessment for Sampling Personal characteristics (CASPer; [24, 25]), and consisted of eight to ten relevant video vignettes accompanied by questions assessing communication, collaboration, social and medical consciousness, ethical awareness, empathy, and reflection. Aptitude tests have shown to be of added value to selection procedures [1, 2, 26]. The aptitude tests used consisted of eight assignments probing talent for transfer (applying knowledge to new information),

textual skills, verbal and inductive reasoning, and organization, as well as the skills assessed by the V-SJT.

For all assignments in the V-SJT and aptitude tests, predetermined answer keys were constructed by a panel of Subject Matter Experts (SMEs; 27). In the first cohort, applicants' answers on each assignment were assessed by two SMEs. Inter- and intra-examiner variation were consistently below 5%. Therefore, in later cohorts, all answers were assessed by a single SME per assignment; intra-examiner variation remained low each year (< 2%). The reliability of the scores (Cronbach's alpha) was 0.71-0.76 per cohort for the V-SJT assignments and 0.54-0.58 for the aptitude tests. At the end of the selection day, candidates rated their satisfaction with the selection procedure and the extent to which the selection procedure assessed characteristics of importance for a medical career as 3.9 ± 0.9 on a scale of 1–5, in which 1 meant strongly disagree and 5 strongly agree.

To determine the final outcome of round two, Z-scores for each assignment were calculated, and applicants were ranked based on their average Z-score for all assignments. A predetermined number of the highest ranking students were admitted to MUMS (selection-positive students). Students who were rejected in either the first or second round of the selection procedure could take part in the national weighted lottery; virtually all primarily rejected students used this opportunity (> 98%). If these primarily rejected students were admitted through the lottery (selection-negative students), they entered the same curriculum as the selection-positive students.

Table 2.1: Translation of the CanMEDS competencies into a blueprint of derived competencies for the selection procedure

CanMEDS	Derived competencies
Medical performance & Knowledge and science[a]	Knowledge shown at pre-university education (pu-GPA[b]) Transfer (knowledge and information integration) Textual comprehension & structuring, verbal & inductive reasoning
Communication	Overall communication skills & strength of arguments
Collaboration	Collaboration skills
Managing	Organizational skills
Health advocating	Social and medical consciousness
Professionalism	Ethical awareness Empathy Reflection skills

[a] combination of two CanMEDS competencies; [b] pu-GPA pre-university Grade Point Average

2.2.3 Outcome variables

The study outcomes available in the bachelor phase varied from cognitively-focused to mainly (inter)personal ones (Table 2.2). Cognitive outcomes included results obtained in theoretical tests at the end of each 4-10 week block (mean Cronbach's α per test: 0.74-0.81), Critical Appraisal of a Topic (CAT) assignments in year 3 (Y3: 28), and progress tests taken four times a year (mean Cronbach's α per test: 0.64-0.76; 29). (Inter)personal outcomes included qualitative evaluations of the students' consulting and reflecting skills (CORE), professional behavior, and first-year portfolio. Evaluation of CORE is based on videotaped simulated patient contacts, peer and expert feedback and self-reflection. Evaluation of professional behavior occurred throughout the whole bachelor in different settings (tutorial groups, group assignments, etc.). In the first-year portfolio, students had to reflect on their own overall performance and progression. Evaluations of these three (inter)personal aspects led to end-of-year assessments with qualifications fail, pass or good.

The OSCE, Objective Structured Clinical Examination, organized in all three bachelor years, was categorized as a 'mixed assessment' in which students had to apply knowledge and skills in (simulated) situations and use interpersonal skills to interact with patients. Multiple CanMEDS competencies are assessed within each OSCE assessment (mean Cronbach's α per test: 0.66-0.76).

Three general outcomes were included in the analysis: drop-out (defined as leaving MUMS without graduating), study delay (graduating from the bachelor in more than three years), and number of credit points obtained within three years (European Credit Transfer System, ECTS; 60 credits per year, accumulating to 180 credits in the three-year bachelor).

The outcome data were stored in the university's electronic administration system, and retrieved with permission (see section 2.2.4) for research purposes.

2.2.4 Ethical approval

During the selection procedure, applicants were asked to give their informed consent for the use of their selection and assessment data for research purposes. It was made clear that not taking part in the study would not adversely influence their progression. All selection applicants agreed to participate. Participant data was anonymized before it was shared with the research team. The study was approved by the Ethical Review Board of the Netherlands Association for Medical Education (NVMO; file number 303).

Table 2.2: Outcome variables based on study results obtained by students during the bachelor phase, with their possible values

	Type of outcome	Measurement level	Possible values
Cognitive	**Block tests**		
	Year 1&2	Continuous	Average of grades at first attempt; 0 - 10
	Year 3	Nominal	Average of grades at first attempt; F/P/G/E
	PT		
		Continuous	Mean Z-score per year, ranging from -2.3 to 4.3
	CAT		
		Nominal	Grade at first attempt; F/P/G
(Inter) personal	**CORE**		
		Nominal	End-of-year grade; F/P/G
	Portfolio		
	Year 1	Nominal	End-of-year grade; F/P
	PB		
	Year 1&2	Nominal	End-of-year grade; F/P
	Year 3	Nominal	End-of-year grade; F/P/G/E
Mixed[1]	**OSCE**		
		Nominal	Once per year; F/P/G
General	**Drop-out**		
	Year 1	Nominal	Yes/No
	Bachelor	Nominal	Yes/No
	Study delay		
	Bachelor	Nominal	Yes/No
	ECTS		
	after 3 years	Continuous	Amount after 3 years medical school; 0-180

Mixed[1] means that the assessment combines cognitive and (inter)personal skills.
F = Fail, P = Pass, G = Good, and E = Excellent.
PT = Progress Test; CAT = Critical Appraisal of a Topic; CORE = Consultation skills and Reflection program; PB = Professional Behavior; OSCE = Objective Structured Clinical Examination; ECTS = European Credit Transfer System.

2.2.5 Statistical analyses

Descriptive statistics were obtained for the demographic variables sex, age and pu-GPA, and for the outcome variables indicated above. Exploratory Chi-Square analyses comparing the selection-positive and selection-negative students on the nominal dependent variables were conducted to obtain a first impression of the results. A repeated measures ANOVA was used to assess the overall progress test difference between groups. A sign test was conducted to investigate the overall difference between the groups taking all outcome measures into account [30].

Confirmatory multiple regression analyses were performed on student level with study performance outcomes as dependent variables, and group membership as independent variable. Group membership was represented by the binary variable groups_SP_SN (0: SN-group: selection-negative students, 1: SP-group: selection-positive students). Cohort and sex (0: male, 1: female) were considered as potential confounders and therefore included as independent variables in the model. The nominal variable cohort corresponds to three categories that are represented in the analysis by two binary (dummy) variables.

Nominal dependent variables were analyzed using logistic regression. Qualitative scores with three or more levels were dummy-coded into fail versus all other scores (i.e. Fail/non-Fail) and the highest possible score versus all other scores (e.g. Good/non-Good). Each of these binary variables was investigated as dependent variable in a logistic regression analysis with independent variables groups_SP_SN, cohort, and sex. For groups_SP_SN, the independent variable of interest, the resulting logistic regression coefficient B, Odds Ratio (OR), Wald statistic and p-value were reported [31]. The OR was used as an indicator of effect size, and Rosenthal's classification values of 1.5, 2.5, and 4 (or equivalent reciprocal values 0.67, 0.40, and 0.25) to indicate small, medium, and large effects, respectively [32].

Continuous dependent variables were similarly analyzed in a linear regression analysis. For each analysis the regression coefficient b, the Standardized Regression Coefficient (SRC), and the corresponding t- and *p*-value (Student's t-test, two-sided) of groups_SP_SN were reported. Here, the SRC was used as an indicator of effect size, using Cohen's classification values 0.1, 0.3, and 0.5 to indicate small, medium, and large effects, respectively [33].

Analyses were conducted using the IBM SPSS Statistics 24.0 software for Windows (SPSS, Inc., Chicago, IL, USA), and results were considered statistically significant if $p < 0.05$.

2.3 Results

Descriptive statistics, categorized by cohort and admission route (selection-positive versus selection-negative), are shown in Table 2.3. The combined cohorts add up to 401 selection-positive and 291 selection-negative students. An independent samples t-test confirms that these groups are significantly different in terms of their performance on the selection assessments in both rounds ($p < 0.001$). Exploratory analyses, performed to obtain a first impression of results, showed significantly better performance of selection-positive compared to selection-negative students, with respect to several cognitive, (inter)personal and mixed outcomes (Figure 2.2). In the following confirmatory analyses, data from the three cohorts (2011–13) were combined while controlling for possible differences between cohort and sex.

Table 2.3: Descriptive statistics of sex, age and pu-GPA per cohort, route of admission and total

	2011 n = 216	2012 n = 238	2013 n = 238	SP n = 401	SN n = 291	Total n = 692
Sex (%)						
Female	63.9	68.9	71.4	70.1	65.6	68.2
Age in years						
Mean (SD)	19.5 (1.4)	18.8 (1.4)	19.3 (1.5)	19.2 (1.5)	19.1 (1.5)	19.2 (1.5)
Pu-GPA*						
Mean (SD)	6.9 (0.6)	6.9 (0.6)	6.9 (0.6)	6.9 (0.6)	6.9 (0.6)	(0.6)

SP: Selected students, SN: primarily Rejected students; *pu-GPA = pre-university Grade Point Average.

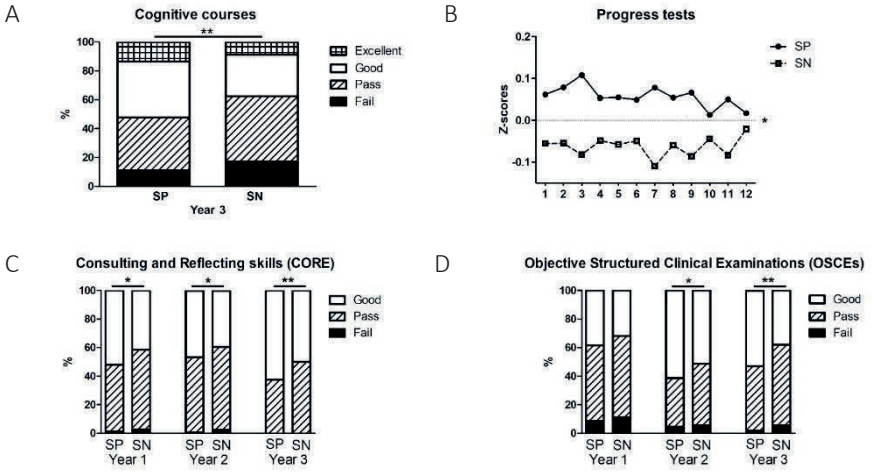


Figure 2.2: Study outcomes of selection-positive (SP) and selection-negative (SN) students on cognitive assignments, i.e. the end-of-course cognitive tests in year 3 (**A**) and the progress tests (**B**), the (inter)personally oriented CORE program (**C**) and the OSCEs (**D**) throughout the three-year bachelor phase.
* $p < 0.05$; ** $p < 0.005$

2.3.1 Cognitive outcomes

During the three-year bachelor program, the selection-positive students outperformed the selection-negative students on several cognitive assignments (Table 2.4). For the cognitive block tests, statistically significant differences were found in year 1 and 3, in favor of the selection-positive students. Furthermore, the mean progress test score was significantly higher for the selection-positive students in the first and second year of the bachelor.

2.3.2 (Inter)personal outcomes

Selection-positive students scored higher than selection-negative students on (inter)personal assessments, although not all differences reached statistical significance (Table 2.4). The selection-positive students performed significantly better on the CORE assessments in the first and last year of the bachelor. Very few students failed professional behavior, yet, selection-positive students appear to be more likely to receive Excellent scores at the end of their bachelor ($p = 0.07$). Lastly, the selection-positive students scored significantly fewer fails on the first-year portfolio.

2.3.3 Mixed outcomes

Notably, selection-positive students significantly outperformed selection-negative students on the OSCEs in all three bachelor years (see Table 2.4).

2.3.4 General study outcomes

The drop-out rate in year 1 was very low and even fewer students dropped out later, without a specific difference between the groups (Table 2.4). The percentage of

delayed students and the amount of ECTS obtained within three years did not significantly differ between the groups.

In summary, controlling for the possible confounders cohort and sex (Table 2.4), the selection-positive students significantly outperformed the selection-negative students on 11 of the 30 outcome variables. In addition, 15 of the remaining 19 non-significant differences were in favor of the selection-positives. These differences occurred across the whole range of variables from cognitive to (inter)personal. The effect sizes of the between-group differences, based on the ORs and SRCs, varied from small to medium/large. Of the four remaining outcome variables, two were equal for both groups; only two outcomes were found to be slightly in favor of the selection-negative students. Applying a sign test to the 30 between-group differences for all outcome variables supports the overall conclusion that study results of selection-positive students are significantly better than those of selection-negative students ($p < 0.001$).

Table 2.4: Comparison of all study performance outcome variables of selection-positive (SP) and selection-negative (SN) students. For all analyses, route of entry was coded SN = 0 and SP = 1, making SN the reference group; cohort and sex were controlled for

Dependent variable	Independent variable: SP versus SN					
Cognitive outcomes						
	SP; M (SD)[a]	SN; M (SD)	B[b]	SRC[c]	t-value	*p*-value
Cognitive courses year 1						
	7.00 (0.88)	6.85 (0.94)	0.151	0.082	2.106	0.036*
Cognitive courses year 2						
	6.82 (0.88)	6.68 (0.89)	0.106	0.059	1.520	0.129
	% of SP	% of SN	B	OR[d]	Wald[e]	*p*-value
Cognitive courses year 3						
Fail/Non-fail	11.0	17.2	-0.507	0.602	4.225	0.040*
Excellent/Non-excellent	13.6	8.8	0.424	1.528	2.369	0.124
CAT[f] year 3						
Fail/Non-fail	10.8	15.9	-0.467	0.627	3.403	0.065
Good/Non-good	5.9	9.3	-0.481	0.618	2.308	0.129
	SP; M (SD)	SN; M (SD)	B	SRC	t-value	*p*-value
Progress tests (Z-scores) year 1						
	0.07 (0.78)	-0.06 (0.82)	0.141	0.087	2.243	0.025*
Progress tests (Z-scores) year 2						
	0.06 (0.83)	-0.07 (0.85)	0.137	0.080	2.013	0.045*
Progress tests (Z-scores) year 3						
	0.05 (0.85)	-0.04 (0.88)	0.090	0.052	1.256	0.210
(Inter)personal outcomes						
	% of SP	% of SN	B	OR	Wald	*p*-value
CORE[g] year 1						
Fail/Non-fail	1.3	2.5	-0.546	0.579	0.830	0.362
Good/Non-good	52.1	41.4	0.464	1.591	8.068	0.005*
CORE[g] year 2						
Fail/Non-fail	0.5	2.3	-1.299	0.273	2.428	0.119
Good/Non-good	46.7	39.5	0.272	1.312	2.630	0.105
CORE[g] year 3						
Fail/Non-fail	0	0	N.A.[i]	N.A.	N.A.	N.A.
Good/Non-good	62.3	49.8	0.494	1.639	8.424	0.004**

(Inter)personal outcomes - Continued						
	% of SP	% of SN	B	OR	Wald	*p*-value
Professional Behavior year 1						
Fail/Non-fail	0.5	0.4	-0.436	0.647	0.124	0.725
Professional Behavior year 2						
Fail/Non-fail	0.0	0.8	N.A.	N.A.	N.A.	N.A.
Professional Behavior year 3						
Fail/Non-fail	0	0	N.A.	N.A.	N.A.	N.A.
Excellent/Non-excellent	12.1	6.8	0.580	1.785	3.343	0.067
Portfolio year 1						
Fail/Non-fail	1.3	4.0	-1.228	0.293	4.931	0.026*
Mixed outcomes						
	% of SP	% of SN	B	OR	Wald	*p*-value
OSCE[h] year 1						
Fail/Non-fail	8.7	11.2	-0.397	0.673	1.961	0.161
Good/Non-good	38.6	32.0	0.433	1.542	5.653	0.017*
OSCE[h] year 2						
Fail/Non-fail	4.6	5.4	-0.176	0.839	0.218	0.641
Good/Non-good	61.3	51.4	0.407	1.502	5.794	0.016*
OSCE[h] year 3						
Fail/Non-fail	2.0	5.4	-1.023	0.359	4.482	0.034*
Good/Non-good	52.9	38.0	0.608	1.837	12.149	0.000**
General outcomes						
	% of SP	% of SN	B	OR	Wald	*p*-value
Drop-out in year 1						
Yes/No	3.0	4.5	-0.366	0.694	0.787	0.375
Drop-out in entire bachelor						
Yes/No	3.5	6.2	-0.566	0.568	2.335	0.127
Study Delay in the bachelor						
Yes/No	19.2	25.5	-0.359	0.698	3.470	0.062
	SP; M (SD)	SN; M (SD)	B	SRC	t-value	*p*-value
ECTS at the end of year 3, including all resits						
	166.5 (35.5)	161.2 (42.6)	4.689	0.060	1.590	0.112

a M (SD) = Mean (Standard Deviation). b B = Regression coefficient. c SRC=Standardized Regression Coefficient. d OR=Odds Ratio. e Wald = Wald statistic. f CAT = Critical Appraisal of a Topic. g CORE = Consulting and Reflecting skills. h OSCE = Objective Structured Clinical Examination. i N.A.= Not Applicable. *p < 0.05, ** p < 0.005.

2.4 Discussion

Backward chaining from the CanMEDS framework was used to develop an outcome-based selection procedure for medical school. This procedure addressed the whole range of competencies, from academic achievement to (inter)personal attributes. We found that the students selected through this procedure significantly outperformed their counterparts who were primarily rejected in the same selection process but were then admitted through an alternative route. Differences in study performance in favor of the selection-positive students were seen across the full range of cognitive, (inter)personal, and mixed outcomes, and throughout the entire three-year bachelor in medicine.

Our finding that selection-positive students performed better than the selection-negative ones on cognitive outcomes was surprising in light of the fact that their pu-

GPA did not differ. This indicates incremental validity of our selection procedure over pu-GPA. The significant differences between the selection-positive and selection-negative students persisted throughout the three-year bachelor. Earlier studies showed that the predictive value of pu-GPA for academic achievement decreases after the first year of medical school [1, 2]. The persisting predictive value is consistent with literature on aptitude tests (e.g. [26, 34]), and therefore likely due to selection. There were only few fails in the end-of-year summative assessments of (inter)personal skills (0-2.4% per outcome measure) and their discriminative value was low. Nevertheless, selection-positive students performed significantly better than selection-negative students, especially with respect to their communication and reflection skills and their portfolio. While almost all students passed the assessment of their professional behavior, selected students were more likely to receive Excellent scores at the end of their bachelor. These findings are in line with previous research on the predictive value of SJTs for (inter)personal performance [25, 35], stating that the predictive value persists over a number of years and predicts performance beyond the cognitively-based pu-GPA.

Interestingly, our combination of tools seems (increasingly) proficient in predicting OSCE performance. So far, OSCE performance has mostly been predicted by MMIs [1, 36], with emerging evidence that SJTs may also be predictive [35]. Aptitude tests, on the other hand, do not appear to predict clinical or pre-clinical OSCE performance [37]. The observed predictive value for the OSCEs in our study inspires confidence with respect to the performance of selected students in the master-phase, where they have to perform in a clinical environment.

General outcomes did not show significant differences between selection-positive and selection-negative students, possibly because of the low frequency of drop-out. Interestingly, other studies from the Netherlands have identified that taking part in a selection process significantly reduces drop-out [22, 38]. This is consistent with our situation; students who entered medical school by lottery only (without participating in the selection procedure) were more likely (about 2.5-2.9 times) to drop-out than selected students [39].

One of the strengths of this study is that the selection procedure somewhat resembled programmatic assessment [40]: combining a number of selection tools with evidence-base [1, 2] as well as the judgments of a variety of examiners (SMEs) to obtain a holistic impression of the applicants. The rater-reliability and internal reliability of the V-SJT and aptitude tests proved acceptable, especially considering the fact that they combined the assessment of multiple competencies. These findings are in line with reviews in this field that have shown good psychometric qualities for SJTs and well-designed aptitude tests [1, 2, 41]. Furthermore, applicants in all cohorts agreed that the selection procedure assessed characteristics of importance for a medical career (supporting face validity). Another strength of this study is the inclusion of three student cohorts that were followed longitudinally throughout their entire three-year bachelor of medicine. This kind of longitudinal research investigating

selection procedures as a whole has been rare, and there have been calls for more of these studies [2, 7]. In addition, the selection-positive students could be compared to selection-negative students within the same cohort, namely the students who were rejected in the same selection procedure but entered medical school through the national weighted lottery.

There are several limitations in the current study that should be kept in mind. Firstly, this was a single-site study, and generalizations to other contexts should be done with caution. However, the use of an internationally known and well-established outcome framework benefits generalizability. It is important to note that the current selection procedure was implemented in a context in which medical schools are considered to be of equal quality. This differs from the situation in other countries, such as the USA and UK, where medical schools are ranked. Secondly, the current study reports on results from the pre-clinical bachelor-phase alone; future research should examine differences across groups in the clinical phase of medical school. Related to the selection procedure itself, there is no way to guarantee that applicants fill in the first-round portfolio themselves. They could receive help from others, or others could even write it for them. However, with the evidence-burden built into this portfolio, this should not affect the applicants' chances of getting into round two. Furthermore, the applicants' score in round one is not taken into account once round two is reached. Lastly, the absence of a face-to-face element in the selection procedure could be seen as a weakness of the selection procedure. On the other hand, including a face-to-face element may also introduce bias [1, 2, 42]. In addition, the chosen approach to selection, having the applicants fill out an online portfolio at home, was found to enable feasible, robust pre-screening at a distance for large numbers of applicants.

2.4.1 Conclusions

All in all, we have shown that an outcome-based, holistic selection procedure is predictive of study success across a variety of cognitive, (inter)personal skills and mixed assessments. Although we did not carry out direct comparisons with other tools, our outcome-based approach seems to address some of the limitations of individual selection tools in relation to predictive validity [7, 10, 13, 15, 43]. We urge others to consider designing and implementing outcome-based selection aligned with curricula and assessment processes, and encourage robust evaluations of the predictive validity of this approach in other contexts, as well as throughout the clinical years and beyond.

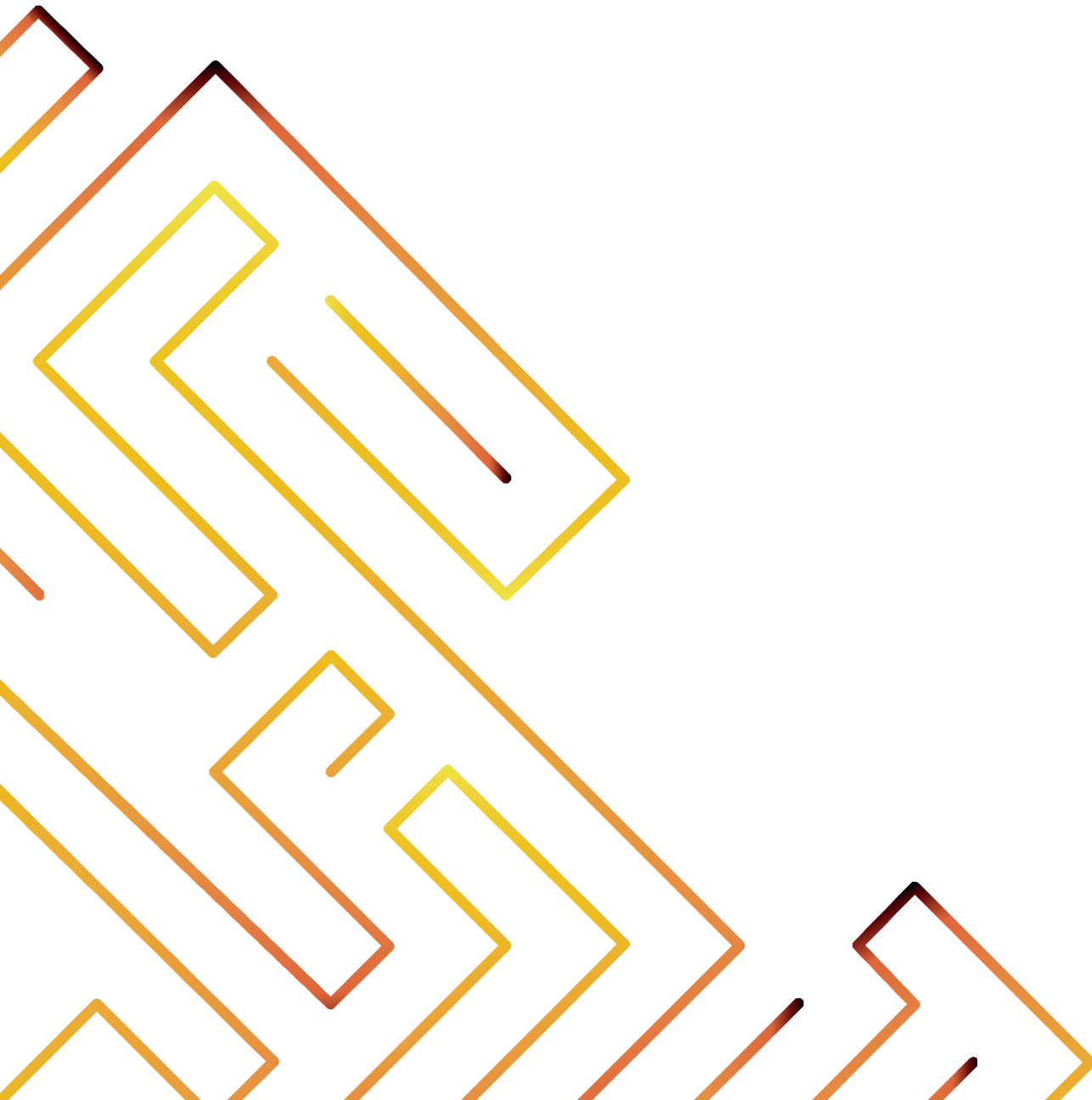
Acknowledgements: The authors would like to thank Dr. Kelly Dore for valuable advice and sharing Computer-based Assessment for Sampling Personal characteristics (CASPer) assignments in order to develop our Situational Judgement Test and Angela Verheyen and Guus Smeets for their essential support in gathering data.

2.5 References

1. Cleland J, Dowell J, McLachlan J, Nicholson S, Patterson F: Identifying best practice in the selection of medical students (literature review and interview survey). 2012.
2. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J: How effective are selection methods in medical education? A systematic review. *Med Educ* 2016, 50(1):36-60.
3. Prideaux D, Roberts C, Eva K, Centeno A, Mccrorie P, Mcmanus C, Patterson F, Powis D, Tekian A, Wilkinson D: Assessment for selection for the health care professions and specialty training: Consensus statement and recommendations from the Ottawa 2010 conference. *Med Teach* 2011, 33(3):215-223.
4. Girotti JA, Park YS, Tekian A: Ensuring a fair and equitable selection of students to serve society's health care needs. *Med Educ* 2015, 49(1):84-92.
5. MacKenzie RK, Dowell J, Ayansina D, Cleland JA: Do personality traits assessed on medical school admission predict exit performance? A UK-wide longitudinal cohort study. *Adv Health Sci Educ* 2016, 22(2):1-21.
6. Patterson F, Lievens F, Kerrin M, Munro N, Irish B: The predictive validity of selection for entry into postgraduate training in general practice: Evidence from three longitudinal studies. *Brit J Gen Pract* 2013, 63(616):E734-E741.
7. Burns CA, Lambros MA, Atkinson HH, Russell G, Fitch MT: Preclinical medical student observations associated with later professionalism concerns. *Med Teach* 2017, 39(1):38-43.
8. Papadakis MA, Teherani A, Banach MA, Knettler TR, Rattner SL, Stern DT, Veloski JJ, Hodgson CS: Disciplinary action by medical boards and prior behavior in medical school. *New Engl J Med* 2005, 353(25):2673-2682.
9. Dore KL, Reiter HI, Kreuger S, Norman GR: CASPer, an online pre-interview screen for personal/professional characteristics: Prediction of national licensure scores. *Adv Health Sci Educ* 2017, 22(2):327–336.
10. Tamblyn R, Abrahamowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, Smee S, Blackmore D, Winslade N, Girard N *et al*: Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA* 2007, 298(9):993-1001.
11. Patterson F, Cleland J, Cousans F: Selection methods in healthcare professions: Where are we now and where next? *Adv Health Sci Educ* 2017, 22(2):229-242.
12. Dore KL, Roberts C, Wright S: Widening perspectives: Reframing the way we research selection. *Adv Health Sci Educ* 2017, 22(2):565-572.
13. Powis D: Selecting medical students: An unresolved challenge. *Med Teach* 2015, 37(3):252-260.
14. Bandiera G, Maniate J, Hanson MD, Woods N, Hodges B: Access and selection: Canadian perspectives on who will be good doctors and how to identify them. *Acad Med* 2015, 90(7):946-952.
15. Hautz SC, Hautz WE, Feufel MA, Spies CD: Comparability of outcome frameworks in medical education: Implications for framework development. *Med Teach* 2015, 37(11):1051-1059.
16. Sklar DP: Who's the Fairest of Them All? Meeting the Challenges of Medical Student and Resident Selection. *Acad Med* 2016, 91(11):1465-1467.
17. Hecker K, Norman G: Have admissions committees considered all the evidence? *Adv Health Sci Educ* 2017, 22(2):573-576.
18. Cleland J, Dowell J, Nicholson S, Patterson F: How can greater consistency in selection between medical schools be encouraged? A project commissioned by the Selecting for Excellence Group (SEEG). 2014.
19. Wilkinson TM, Wilkinson TJ: Selection into medical school: From tools to domains. *BMC Med Educ* 2016, 16(1):258.
20. Ferguson E, James D, Madeley L: Factors associated with success in medical school: Systematic review of the literature. *BMJ* 2002, 324(7343):952–957.
21. Siu E, Reiter HI: Overview: What's worked and what hasn't as a guide towards predictive admissions tool development. *Adv Health Sci Educ* 2009, 14(5):16.
22. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR: Predictive validity of the multiple mini-interview for selecting medical trainees. *Med Educ* 2009, 43(8):767-775.
23. Lievens F: Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests. *Med Educ* 2013, 47(2):182-189.
24. Lievens F, Patterson F, Corstjens J, Martin S, Nicholson S: Widening access in selection using situational judgement tests: Evidence from the UKCAT. *Med Educ* 2016, 50(6):624-636.

25. Wilson IG, Roberts C, Flynn EM, Griffin B: Only the best: Medical student selection in Australia. *MJA* 2012, 196(5):357.
26. Monroe A, Quinn E, Samuelson W, Dunleavy DM, Dowd KW: An overview of the medical school admission process and use of applicant data in decision making: What has changed since the 1980s? *Acad Med* 2013, 88(5):672-681.
27. Frank JR, Snell L, Englander R, Holmboe ES: Implementing competency-based medical education: Moving forward. *Med Teach* 2017, 39(6):568-573.
28. Gruppen L, Frank JR, Lockyer J, Ross S, Bould MD, Harris P, Bhanji F, Hodges BD, Snell L, ten Cate O: Toward a research agenda for competency-based medical education. *Med Teach* 2017, 39(6):623-630.
29. Holmboe ES, Sherbino J, Englander R, Snell L, Frank JR: A call to action: The controversy of and rationale for competency-based medical education. *Med Teach* 2017, 39(6):574-581.
30. Lockyer J, Carraccio C, Chan M, Hart D, Smee S, Touchie C, Holmboe ES, Frank JR: Core principles of assessment in competency-based medical education. *Med Teach* 2017, 39(6):609-616.
31. Frank JR, Danoff D: The CanMEDS initiative: Implementing an outcomes-based framework of physician competencies. *Med Teach* 2007, 29(7):642-647.
32. Holmboe ES, Edgar L, Hamstra S: The Milestones Guidebook. In. Edited by ACGME. Chicago, IL. ; 2016.
33. Medical Council of India: Visions 2015. In. New Delhi; 2011.
34. National Alliance for 39 Physician Competence: A Guide to Good Medical Practice – USA. In. Edited by National Alliance for Physician Competence; 2009.
35. Van Herwaarden CLA, Laan RFJM, Leunissen RRM: Raamplan Artsopleiding 2009. Utrecht; 2009.
36. Frank JR: The CanMEDS 2005 physician competency framework: Better standards, better physicians, better care: Royal College of Physicians and Surgeons of Canada; 2005.
37. Frank JR, Snell LS, Sherbino J: The draft CanMEDS 2015 physician competency framework - series ii. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2014.
38. Dore KL, Reiter HI, Eva KW, Krueger S, Scriven E, Siu E, Hilsden S, Thomas J, Norman GR: Extending the interview to all medical school candidates -- Computer-based Multiple Sample Evaluation of Noncognitive Skills (CMSENS). *Acad Med* 2009, 84(10 Suppl):S9-12.
39. Emery JL, Bell JF: The predictive validity of the BioMedical admissions test for pre-clinical examination performance. *Med Educ* 2009, 43(6):557-564.
40. Patterson F, Zibarras L, Ashworth V: Situational judgement tests in medical education and training: Research, theory and practice: Amee guide no. 100. *Med Teach* 2015, 38(1):3-17.
41. de Brouwer CPM, Kant I, Smits LJM, Voogd AC: Training Critical Appraisal of a Topic. Een onmisbare handleiding in het tijdperk van Evidence Based Medicine: Mediview; 2009.
42. Tio RA, Schutte B, Meiboom AA, Greidanus J, Dubois EA, Bremers AJA, the Dutch Working Group of the Interuniversity Progress Test of M: The progress test of medicine: the Dutch experience. *Perspectives on Medical Education* 2016, 5(1):51-55.
43. Field AP: Discovering statistics using SPSS (and sex, drugs and rock 'n' roll), 3rd edn. Los Angeles: SAGE Publications; 2009.
44. Rosenthal JA: Qualitative Descriptors of Strength of Association and Effect Size. *Journal of Social Service Research* 1996, 21(4):37-59.
45. Cohen J: Statistical Power Analysis for the Behavioral Sciences: Routledge; 1988.
46. Lehmann EL, D'Abrera HJM: Nonparametrics: Statistical methods based on ranks. San Fransisco: Holden-Day; 1975.
47. de Visser M, Fluit C, Fransen J, Latijnhouwers M, Cohen-Schotanus J, Laan R: The effect of curriculum sample selection for medical school. *Adv Health Sci Educ* 2016, 22(1):1-14.
48. Dore KL, Kreuger S, Ladhani M, Rolfson D, Kurtz D, Kulasegaram K, Cullimore AJ, Norman GR, Eva KW, Bates S *et al*: The reliability and acceptability of the Multiple Mini-Interview as a selection instrument for postgraduate admissions. *Acad Med* 2010, 85(10 Suppl):S60-S63.
49. Kelly M, Dowell J, Husbands A, Newell J, O'Flynn S, Kropmans T, Dunne F, Murphy A: The fairness, predictive validity and acceptability of multiple mini interview in an internationally diverse student population- a mixed methods study. *BMC Med Educ* 2014, 14(1):267.
50. Rees EL, Hawarden AW, Dent G, Hays R, Bates J, Hassell AB: Evidence regarding the utility of multiple mini-interview (MMI) for selection to undergraduate health programs: A BEME systematic review: BEME Guide No. 37. *Med Teach* 2016, 38(5):443-455.

51. Husbands A, Mathieson A, Dowell J, Cleland J, MacKenzie R: Predictive validity of the UK clinical aptitude test in the final years of medical school: A prospective cohort study. *BMC Med Educ* 2014, 14(1):88.
52. Bodger O, Byrne A, Evans PA, Rees S, Jones G, Cowell C, Gravenor MB, Williams R: Graduate Entry Medicine: Selection Criteria and Student Performance. *Plos One* 2011, 6(11):e27161.
53. Wouters A, Croiset G, Galindo-Garre F, Kusurkar RA: Motivation of medical students: Selection by motivation or motivation by selection. *BMC Med Educ* 2016, 16(1):1-9.
54. Schripsema NR, van Trigt AM, Borleffs JCC, Cohen-Schotanus J: Selection and study performance: Comparing three admission processes within one medical school. *Med Educ* 2014, 48(12):1201-1210.
55. van der Vleuten CPM, Schuwirth LWT: Assessing professional competence: from methods to programmes. *Med Educ* 2005, 39(3):309-317.
56. De Leng WE, Stegers-Jager KM, Husbands A, Dowell JS, Born MP, Themmen APN: Scoring method of a Situational Judgment Test: influence on internal consistency reliability, adverse impact and correlation with personality? *Adv Health Sci Educ* 2017, 22(2):243–265.
57. Griffin BN, Hu W: The interaction of socio-economic status and gender in widening participation in medicine. *Med Educ* 2015, 49(1):103-113.
58. Griffin BN, Wilson IG: Interviewer bias in medical student selection. *MJA* 2010, 193(6):343-346.
59. Frenk J, Chen L, Bhutta ZA, Cohen J, Crisp N, Evans T, Fineberg H, Garcia P, Ke Y, Kelley P *et al*: Health professionals for a new century: Transforming education to strengthen health systems in an interdependent world. *the Lancet* 2010, 376(9756):1923-1958.



CHAPTER 3

Outcome-based selection can predict performance in the clinical years of medical school: The proof is in the pudding

Schreurs S, Cleutjens K, Cleland J, oude Egbrink MGA. Outcome-based selection can predict performance in the clinical years of medical school: The proof is in the pudding. *Academic Medicine*. 2020; accepted for publication.

Abstract

Purpose

Medical school selection aims to identify the best possible students and, ultimately, the best future doctors from a large, homogeneous pool of applicants. Constructive alignment of medical school selection, curricula and assessment with the ultimate outcomes (e.g. CanMEDS) has been proposed as means to attain this goal. Whether this approach is effective has not yet been established. This gap in the literature was addressed by assessing the relationship between performance in an outcome-based selection procedure and during the clinical years of medical school.

Method

Two groups of students were compared: those selected via an outcome-based (i.e. CanMEDS-based) selection procedure versus those rejected in this procedure who entered the program through a national, GPA-based lottery procedure. Performances of both groups on all seven CanMEDS-roles were compared during clinical rotations, for all three master years. Data were compared for three cohorts (2011-2013).

Results

Selected students significantly outperformed the initially rejected but lottery-admitted students in all years, and the differences between groups increased over time. Differences were apparent in the first year in the roles of Communicator, Collaborator and Professional, in the second year also for the roles of Organizer and Health Advocate, while in the third year in the role of Academic as well.

Conclusions

A constructively aligned selection procedure has increasing predictive value across the clinical years of medical school compared to a GPA-based lottery procedure. This suggests that constructive alignment of selection, curriculum and assessment to ultimate outcomes is effective in creating a selection procedure predictive of clinical performance.

3.1 Introduction

Selection for medical school aims to identify the best possible students and, ultimately, the best future doctors[1-4]. As the number of applicants typically outnumbers the available places, a range of tools has been developed to help medical schools in the selection process. These include cognitive indicators (e.g. pre-university Grade Point Average [pu-GPA], aptitude tests [e.g. the Medical College Admission Test]) and (inter)personal assessments (e.g. Multiple Mini Interviews [MMIs] and Situational Judgement Tests [SJTs])[1,2,5]. Typically, a combination of cognitive and (inter)personal tools within a selection procedure is preferred as both are important elements in educational performance and future career[3,6-8].

While many studies have identified that cognitive selection tools are predictive for study success during the early, mainly pre-clinical years of medical education, their predictive value decreases over time[2,3,7,9]. Conversely, (inter)personal assessments seem more predictive of performance in the later years of medical school[1-3,5,7]. However, there have been few direct investigations of the predictive validity of combinations of cognitive and (inter)personal selection tools for the clinical phase of medical school[7]. The few studies which have looked at this question have used relatively crude overall measures (drop-out, study delay, and/or overall grades per clerkship) and found few differences and small effect sizes[10,11]; only one study showed that MMIs predict clerkship performance at a more granular level[12]. However, there are many more, finer-grained indicators of performance in the clinical years of medical school, including performance on formal and workplace-based assessments using competences or EPAs[13-18]. These finer-grained indicators may be more authentic and more informative indicators of study success[19].

The relative lack of in-depth investigations of the predictive value of selection for the clinical phase of medical school and future performance as a doctor may be due, at least to some extent, to the fact that the relation between the criterion (i.e. outcome: clinical performance) and predictor (selection performance) is distal, much more so than for the pre-clinical phase[1,20,21]. One way to deal with this issue is to blueprint the selection procedure to the outcome criteria at the end of medical school[3,4,7,19]. In this manner, the constructs assessed in selection and in the clinical phase of medical school are more closely related, and therefore, although predictor and criterion are still distal in time, they are more congruent in content. Outcome frameworks that describe the roles and/or competences students should be capable of at graduation and, hence, at the start of their career as a future medical doctor[14,16,22], offer this possibility. This approach overcomes the so-called 'criterion problem', i.e. the impossibility to determine the worth of a selection procedure if it is unclear what outcome should be predicted[1,2,23]. Ultimately, aligning selection with outcomes may decrease the risk of selecting good students instead of good doctors.

Constructive alignment throughout the entirety of medical school - starting with selection and ending at the end of the clinical phase - has been proposed as a way to

improve the predictive validity of selection into the clinical years[5,8,24,25]. A previous study from our team illustrated that a multi-tool selection procedure aligned with the outcome framework used to build the medical curriculum predicted student performance in the pre-clinical phase of that curriculum[4]. Up to now, however, it was not known whether such explicit constructive alignment throughout the entirety of medical school also increases a selection procedure's predictive value in the clinical phase.

We aimed, therefore, to assess the relationship between performance at selection and during the clinical years of a medical program in the context of a medical school where the selection procedure, curriculum and assessments are all aligned with an outcome framework based on the CanMEDS: the 2009 Framework for Undergraduate Medical Education in the Netherlands[16]. Both frameworks expand on the roles a doctor should be able to fulfill, and the competences a medical school graduate should possess to be able to fulfill these roles. The roles are: Medical Expert, Communicator, Collaborator, Organizer (Leader in the 2015 edition), Health Advocate, Scholar and Professional [14-16]. We examined whether students who were selected via the outcome-based selection procedure, focusing on assessing dispositions for the CanMEDS roles, perform better in these roles during the clinical phase of their medical program compared to students who were rejected in this procedure and entered medical school via an alternative admission route, a lottery procedure based purely on pre-university Grade Point Average (pu-GPA).

3.2 Method

3.2.1 Context

We conducted the study at Maastricht University Medical School (MUMS), the Netherlands. In the Netherlands, medical studies consist of a pre-clinical bachelor program followed by a clinical master program in which students complete a predetermined number of rotations. Both are three-year programs aligned with the CanMEDS outcome framework, as is the norm in the Netherlands[16].

We focused on the predictive value of the MUMS selection procedure for student performance during clinical rotations in each of the CanMEDS competences. The MUMS clinical master phase consists of five compulsory rotations (Reflective Medicine [e.g., internal medicine and pulmonology, 12 weeks], Surgical Medicine [12 weeks], Mother and Child [10 weeks], Neurosciences [20 weeks], and Family/Social Medicine [12 weeks]), as well as two elective rotations of 8 and 10 weeks, a scientific research participation (SCIP, 18 weeks), and a senior rotation at a department of choice, the so-called healthcare participation (HELP, 18 weeks). For each rotation, students had to gather quantitative information (e.g. from knowledge tests), but also qualitative and narrative feedback on their performance in all seven CanMEDS roles, on multiple occasions and from different people representing various roles (e.g. supervisors, nurses, peers). The students reflected on this information and feedback as well as on their own progression in their portfolio, guided by a mentor. The role of the mentor

was to advise the students on how to learn from the feedback and progress in their learning.

We also looked at students' results on an inter-university progress test administered four times per year. This progress test focuses on knowledge and is taken into account in the assessment of the CanMEDS role of Medical Expert. This progress test is administered throughout both the bachelor and master phase. The current study focused on performance during the latter.

The combination of all this information, including the reflections on this information by the student, is appraised by the Board of Examiners for Medicine three times during the three-year master program (at T1, T2 and T3). In these appraisals, a final judgement per competence is given on a 3-point scale: 'below expectations', 'as expected' and 'exceeds expectations'. At T1, at least the first two rotations (Reflective and Surgical Medicine) have been completed. At T2, the remainder of the compulsory rotations must be completed (Mother and Child, Neurosciences, and Family/Social Medicine). At T3 the whole program is finished, including the HELP, which is always the last part of the master. The two elective rotations and the SCIP may be completed at any point throughout the three-year program.

We specifically included the student cohorts entering MUMS' pre-clinical bachelor in 2011, 2012 and 2013, as there were two distinct ways to get into MUMS in those years: a local, outcome-based selection procedure and a national pu-GPA-based lottery procedure. Thus, in these years, students who were rejected in the local selection procedure could gain entry to the medical program via the lottery procedure which only took pu-GPA into account, providing a natural control group. Furthermore, the acceptance of rejected applicants through lottery prevents restriction of range: applicants scoring low in the selection procedure could enter through the GPA-based lottery, applicants with a low pu-GPA had a chance of entering through the selection procedure. The lottery procedure was weighted as follows: applicants with a pu-GPA ≥8 (in the Netherlands, students are scores from 0 to 10, 10 being the highest possible grade) were all admitted. Applicants with lower pu-GPAs were divided into groups: GPAs of 7.5 until 8; 7 until 7.5; 6.5 until 7; and 6 until 6.5, and these groups were admitted in the ratio 9:6:4:3[26]. In the current study, we compared the clinical performance of students selected in the MUMS selection procedure (Selection-Positive, SP-group) to that of students who were primarily rejected in the selection procedure but got into medical school through the lottery procedure (Selection-Negative, SN-group). We gathered data for T1, T2 and T3 for the 2011 and 2012 cohorts, and for T1 and T2 for the 2013 cohort, as they had not yet finished T3 assessment at the time of data collection.

3.2.2 Selection procedure

As stated earlier, the selection procedure at MUMS is aligned with the CanMEDS competence framework[16]. In a previous article, we have explained the manner in which we adapted the CanMEDS competences to the level of the applicants,

translating outcome competences into 'derived competences' an 18-year-old applicant may possess. These 'derived competences' are *knowledge shown at pre-university education* (i.e. pu-GPA), *transfer* (knowledge and information integration), *textual comprehension and structuring*, *verbal and inductive reasoning*, o*verall communication and strength of arguments*, *collaboration*, *organization*, *social and medical consciousness*, *ethical awareness*, *empathy*, and *reflection* (for more information, see chapter one or two[4]). The first round consists of a portfolio with four main parts: **pre-university academic performance** (including pu-GPA, additional courses and other academic activities, aimed at measuring knowledge obtained during pre-university education and as indicator for the derived competences *transfer*, *textual comprehension and reasoning*), **extracurricular activities**[27] (a description of these activities with an appropriate explanation of experiences gained in competences relevant for studying and practicing medicine, such as *communication*, *collaboration* and *organization*), **fit with problem-based learning** (to make sure applicants make an informed choice for this educational philosophy) and **fit with MUMS** (to make sure applicants are aware of the Maastricht curriculum and its differences with that of other medical schools in the Netherlands). These parts are weighed 40%, 40%, 10% and 10%, respectively; the first two, heavily-weighted parts are in line with the 'derived competences', whereas the last two parts are mostly aimed at creating awareness about the MUMS program and approach[4].

The second round focuses more specifically on the competences expected from a medical doctor. It consists of two separate tests. First, there is a Situational Judgement Test (SJT), which is a test in which applicants are confronted with situations that relate to the job or study they are applying for, aimed at measuring specific 'Implicit Trait Policies (ITPs)'. ITPs are "beliefs about the cost or benefits of acts expressing compassion, caring and respect for patients, related to candidates' trait expression and values" (Patterson et al., 2016, page 2 [28]). For more information on the theoretical considerations on SJTs, see: Patterson, Zibarras, and Ashworth [28] or Motowidlo, Hooper, and Jackson[29]. In the SJT as it was applied at MUMS, applicants are shown videos of situations (e.g. critical incidents or daily life situations) that are related to medical school, clinical practice or specific CanMEDS-competences. In contrast with 'standard' SJTs, an open-ended format was used in which applicants were asked to respond to, reflect on, think about, or otherwise engage with the situation outlined in the video. The test takes 90 minutes and consists of about 10 assignments with about four sub-questions each. Second is a Written Aptitude Test (WAT) involving a broad range of questions about managing relevant situations, planning skills, fluid intelligence and so on. This test takes 75 minutes and consists of about eight assignments with a varying amount of subitems. Together these two test formats map onto all the CanMEDS competences[14,16], targeted to the level of knowledge, skills and attitudes of about 18 to 19 year old applicants. For more information on the specific selection procedure employed in the current study, please see chapter one or two[4]. The selection procedure has already been shown to be robust and replicable. In another study, we compared the constructs we intended to measure (i.e. the derived competences) to the applicants' results and found that the constructs we intended to

be measured were indeed accurate. For more information on the development, content and psychometric properties of the selection procedure, the reader is referred to chapter four[30].

3.2.3 Outcome variables

The outcome variables we included are assessments of the Board of Examiners (on the three-point scale of 'below expectations', 'as expected' and 'exceeds expectations') of the seven CanMEDS roles in each of the three master years (i.e. at T1, T2 and T3), and the mean progress test result for each year. To compare students within their cohorts, their raw score on each progress test is converted into a z-score within the student's university and cohort. We retrieved these individual z-scores from the university's database and calculated a mean progress test score per year per student, if students had completed at least three out of four progress tests that year (students could pass the progress test requirement with three tests, provided the results were good enough).

We gathered data on the progress test results and on the performance in the seven CanMEDS roles on T1 and T2 in March 2019. In April 2019 we gathered the data for T3.

3.2.4 Ethical approval

During the selection procedure, we asked applicants to give their informed consent for the use of their selection and study assessment data for research purposes. We made it clear that not taking part in this research would not adversely influence their progression. The lead author anonymized participant data before sharing it with the research team. The Ethical Review Board of the Netherlands Association for Medical Education approved this research (NVMO; file number 303).

3.2.5 Statistical analysis

We produced descriptive statistics for all outcomes (the seven roles at all three timepoints and the progress test z-scores per year) as well as the covariates (sex, age, cohort and pu-GPA). These covariates were chosen based on previous research (i.e. sex[31], age[32] and pu-GPA[33]) or because there were some known changes (i.e. cohort; some assessments changed very slightly over the cohorts and therefore cohort was always taken into account as a covariate). We assessed differences between the Selection-Positive and Selection-Negative students on age and pu-GPA using analyses of variance (ANOVAs), and we analyzed differences between Selection-Positive and Selection-Negative students on sex and cohort using a Chi-square analysis. Furthermore, as there were some differences between the cohorts, we added some descriptive statistics per cohort on sex, age and pu-GPA as well (age and pu-GPA using ANOVA and sex using Chi-square analysis).

We considered the performances on the seven roles as our primary outcomes. We explored these first using a Chi-square analysis (or, if applicable, a Fisher's exact test), taking into account all three levels of the outcomes ('below expectations', 'as

expected' and 'exceeds expectations'). The occurrence of 'below expectations' was so rare for all roles except *Medical Expert* (see section 3.3), that we did not take it into account in later analyses. We then applied binary logistic regression as the main analysis method to all outcomes (including Medical Expert). We compared Selection-Positive and Selection-Negative students in how often they achieved 'as expected' versus 'exceeds expectations' for each of the seven roles at each timepoint. We set the predicted outcome for the binary logistic regression to 'as expected'. In each regression analysis, we took all covariates into account.

For the progress test, we compared the mean z-scores per year between the two student groups (Selection-Positive versus Selection-Negative) using an analysis of covariance (ANCOVA), with the same covariates as in the regression analyses. All data were analyzed using SPSS version 24 for Windows (IBM statistics, Armonk, New York).

3.3 Results

3.3.1 Descriptive statistics

The total sample included 692 students, of whom 401 (57.9%) were selected via the outcome-based process and 291 (42.1%) were rejected in this procedure and entered through lottery (i.e. on the basis of their pu-GPA). Gap years and delays in progression (e.g., repeating years) result in missing data. As mentioned earlier, T3-data were not yet available for cohort 2013, leading to a smaller sample size for T3.

The descriptive statistics per covariate are shown in Table 3.1. There were no statistically significant differences between the cohorts on the other covariates. Therefore, the cohorts were combined in all later analyses. Furthermore, we found no significant differences on sex, age, cohort and pu-GPA between Selection-Positive and Selection-Negative students.

Table 3.1: Descriptive statistics of sex, age and pu-GPA per group of medical students (Selection-Positive and Selection-Negative), per cohort (2011, 2012 and 2013, in which Selection-Positive and Selection-Negative were combined), and in total

	SP	SN	2011	2012	2013	Total
N	401	291	216	238	238	692
Sex in %						
Female	69.9	65.9	63.9	68.9	71.4	68.2
Age in years[a]						
Mean (SD)	19.2 (1.5)	19.3 (2.1)	19.6 (2.1)	18.8 (1.6)	19.3 (1.6)	19.2 (1.8)
Pu-GPA						
Mean (SD)	6.94 (0.65)	6.90 (0.67)	6.87 (0.66)	6.95 (0.67)	6.94 (0.65)	6.92 (0.66)
Cohort: n (%)						
2011	111 (51.4)	105 (48.6)				
2012	139 (58.4)	99 (41.6)				
2013	149 (62.2)	89 (37.4)				

Abbreviations: pu-GPA = pre-university Grade Point Average; SD = standard deviation; SP = Selection-Positive; SN = Selection-Negative

[a]at the start of their bachelor in medicine

3.3.2 Performance on the seven CanMEDS roles

The descriptive statistics and exploratory Chi-square analysis are shown in Table 3.2. It can be seen that in the entire sample of students 'below expectations' is rarely ever given, except for the role of Medical Expert, where the overall frequency decreases from 14.4% (n=89) to 8.6% (n=46) to 0% (n=0) for T1, T2 and T3, respectively. In T3, students could not complete their portfolios with a 'below expectations' on their grade list, therefore there are no 'below expectations' to take into account for T3. Table 3.2 also shows significant differences in favor of the Selection-Positive students for the roles of Communicator, Collaborator and Professional at T1. At T2, the Organizer role also shows a significantly better performance of selected students, in addition to the roles of Communicator, Collaborator and Professional. At T3, the Selection-Positive students significantly outperform the Selection-Negative students on all-but-one competences: Communicator, Collaborator, Organizer, Academic, Health Advocate and Professional.

The logistic regressions for the binary outcomes ('as expected' versus 'exceeds expectations'), taking into account all covariates that may affect performance (sex, age, cohort and pu-GPA), are shown in Table 3.3. In line with the findings in the Chi-square analysis, there are significant differences at T1, with the Selection-Positive students being graded more often as 'exceeds expectations' than the Selection-Negative students for the roles of Communicator, Collaborator and Professional. As for the Chi-square analysis, these differences remain in T2, and additionally selected students show significantly favorable performance on the roles of Organizer and Health Advocate. The results at T3 are also in line with the Chi-square analysis: the Selection-Positive students performed significantly better in the roles of Communicator, Collaborator, Organizer, Academic, Health Advocate and Professional. The Odds Ratios of the effects found in the regression analyses indicate small to medium effects[34]. Significant effects of covariates are also shown, and a clear pattern of pu-GPA having an additional positive effect can be seen. Furthermore, cohort and sex occasionally influence performance.

As indicated before, the progress test results are part of the students' performance in the role of Medical Expert. Table 3.4 shows the ANCOVA results, comparing the mean scores per year for the progress test and comparing between the Selection-Positive and Selection-Negative students, controlled for sex, age, cohort and pu-GPA. We found significant differences between the groups for years 2 and 3, with the selected students having a higher mean score on the progress test in these years. Only one of the covariates, pu-GPA, significantly influenced the results of the progress tests, and did so throughout all three years.

Table 3.2: Descriptive statistics and Chi-square analyses of the primary outcome variables: the medical students' (Selection-Positive versus Selection-Negative; the cohorts starting their bachelor programs in 2011, 2012 and 2013; N=692) performance on the seven CanMEDS roles of the medical doctor throughout the clinical master phase

	Group	N	Below exp. (%)	As exp. (%)	Exceeds exp. (%)	X^2 or Fishers[a]	p-value[b]
T1							
Medical Expert	SP	367	47 (12.8)	258 (70.3)	62 (16.9)	1.834	0.399
	SN	253	42 (16.6)	168 (66.4)	43 (17.0)		
Communicator	SP	367	1 (0.3)	181 (49.3)	185 (50.4)	17.048	0.000
	SN	253	2 (0.8)	165 (65.2)	86 (34.0)		
Collaborator	SP	367	2 (0.5)	162 (44.1)	203 (55.3)	16.507	0.000
	SN	253	0 (0.0)	153 (60.5)	100 (39.5)		
Organizer	SP	367	5 (1.4)	271 (73.8)	91 (24.8)	2.239	0.341
	SN	253	4 (1.6)	199 (78.7)	50 (19.8)		
Academic	SP	367	0 (0.0)	305 (83.1)	62 (16.9)	2.495	0.323
	SN	253	2 (0.8)	208 (82.2)	43 (17.0)		
Health advocate	SP	367	1 (0.3)	327 (89.1)	39 (10.6)	4.275	0.099
	SN	253	2 (0.8)	235 (92.9)	16 (6.3)		
Professional	SP	367	0 (0.0)	197 (53.7)	170 (46.3)	11.193	0.002
	SN	253	1 (0.4)	167 (66.0)	85 (33.6)		
T2							
Medical Expert	SP	315	21 (6.7)	212 (67.3)	82 (26.0)	5.669	0.058
	SN	218	25 (11.5)	150 (68.8)	43 (19.7)		
Communicator	SP	315	0 (0.0)	110 (34.9)	205 (65.1)	13.584	0.000
	SN	218	0 (0.0)	111 (50.9)	107 (49.1)		
Collaborator	SP	315	0 (0.0)	96 (30.5)	219 (69.5)	16.316	0.000
	SN	218	0 (0.0)	104 (47.7)	114 (52.3)		
Organizer	SP	315	1 (0.3)	185 (58.)	129 (41.0)	19.985	0.000
	SN	218	0 (0.0)	168 (77.1)	50 (22.9)		
Academic	SP	315	0 (0.0)	224 (71.1)	91 (28.9)	2.255	0.080
	SN	218	0 (0.0)	167 (77.0)	50 (23.0)		
Health advocate	SP	315	0 (0.0)	262 (83.4)	52 (16.6)	2.696	0.064
	SN	218	0 (0.0)	193 (88.5)	25 (11.5)		
Professional	SP	315	0 (0.0)	126 (40.0)	189 (60.0)	7.866	0.003
	SN	218	0 (0.0)	114 (52.3)	104 (47.7)		
T3							
Medical Expert	SP	207	0 (0.0)	163 (78.7)	44 (21.3)	0.006	0.938
	SN	153	0 (0.0)	121 (79.1)	32 (20.9)		
Communicator	SP	208	0 (0.0)	47 (22.6)	161 (77.4)	19.385	0.000
	SN	153	0 (0.0)	68 (44.4)	87 (55.6)		
Collaborator	SP	208	0 (0.0)	44 (21.2)	164 (78.8)	14.970	0.000
	SN	153	0 (0.0)	61 (39.9)	92 (60.1)		
Organizer	SP	208	0 (0.0)	104 (50.0)	104 (50.0)	9.212	0.002
	SN	153	0 (0.0)	101 (66.0)	52 (34.0)		
Academic	SP	208	0 (0.0)	110 (52.9)	98 (47.1)	6.908	0.009
	SN	153	0 (0.0)	102 (66.7)	51 (33.3)		
Health advocate	SP	208	0 (0.0)	160 (76.9)	48 (23.1)	6.626	0.010
	SN	153	0 (0.0)	134 (87.6)	19 (12.4)		
Professional	SP	208	0 (0.0)	66 (31.7)	142 (68.3)	12.744	0.000
	SN	153	0 (0.0)	77 (50.3)	76 (49.7)		

Abbreviations: SP = Selection-Positive; SN = Selection-Negative; exp. = expectations.

[a] X^2 or Fishers means Chi-square analysis was conducted unless one of the expected frequencies was below 5, in which case the Fisher's exact statistics are reported. [b] statistically significant p-values are indicated in bold.

Table 3.3: Binary logistic regressions comparing Selection-Positive medical students to Selection-Negative medical students (starting their bachelor in 2011, 2012 and 2013) on the seven CanMEDS roles of the medical doctor throughout the clinical master phase

	N	B[a]	S.E.	Wald	p-value[b]	OR[c]	Sig. cov.[d] (OR)
T1							
Medical expert	521	-0.051	0.237	0.047	0.828	0.950	GPA (2.565)
Communicator	605	0.673	0.177	14.433	0.000	1.960	GPA (1.661)
Collaborator	606	0.681	0.172	15.623	0.000	1.975	GPA (1.369)
Organizer	599	0.317	0.206	2.356	0.125	1.372	None
Academic	606	0.065	0.231	0.078	0.780	1.067	Cohort GPA (2.128)
Health Advocate	605	0.534	0.313	2.910	0.088	1.705	Cohort
Professional	607	0.534	0.178	9.040	0.003	1.706	Cohort GPA (1.613)
T2							
Medical expert	479	0.443	0.241	3.372	0.066	1.557	GPA (3.603)
Communicator	523	0.676	0.186	13.211	0.000	1.967	GPA (1.476)
Collaborator	523	0.724	0.188	14.767	0.000	2.062	Sex (0.646)
Organizer	522	0.890	0.208	18.260	0.000	2.435	GPA (1.902)
Academic	522	0.311	0.216	2.081	0.149	1.365	GPA (2.207)
Health Advocate	522	0.568	0.281	4.083	0.043	1.765	GPA (2.342)
Professional	523	0.512	0.183	7.828	0.005	1.669	GPA (1.480)
T3							
Medical expert	353	0.166	0.293	0.323	0.570	1.181	GPA (3.906)
Communicator	356	1.015	0.239	18.089	0.000	2.762	Sex (0.549)
Collaborator	356	0.901	0.245	13.466	0.000	2.461	Sex (0.550) GPA (1.766)
Organizer	356	0.756	0.232	10.536	0.001	2.126	GPA (1.841)
Academic	356	0.680	0.235	8.366	0.004	1.973	GPA (2.174)
Health Advocate	356	1.052	0.326	10.396	0.001	2.862	Cohort (2.114)
Professional	356	0.831	0.232	12.869	0.000	2.296	Sex (0.486) GPA (2.032)

Abbreviations: SP = Selection-Positive; SN = Selection-Negative; S.E. = Standard Error of the regression coefficient; GPA = pre-university Grade Point Average.

[a] B, Regression coefficient, where the outcomes are coded 0=as expected and 1=exceeds expectations. [b] statistically significant p-values are indicated in bold. [c] OR, Odds Ratio with SN as reference group (i.e. coded 0=SN and 1=SP). [d] sig. cov. = the covariates that were shown to be significant predictors in the model with all covariates, including (for the continuous and binary variables) their Odds Ratios. Covariates included in the model: Cohort, Sex, Age, and pu-GPA (GPA). Sex was coded Female = 0 and Male = 1, thus the results show the females outperforming the males on four occasions.

Table 3.4: ANCOVA with Selection-Positive versus Selection-Negative medical students compared on their yearly progress test average expressed as z-scores, and controlled for the variables sex, age, cohort (2011, 2012 and 2013) and pu-GPA

	Group	N	Mean	SD	F-value	p-value[a]
Progress test Year 1	SP	356	0.0786	0.8370	0.497	0.481
	SN	243	-0.0062	0.8559		
	Total	599	0.0442	0.8450		
Progress test Year 2	SP	319	0.1482	0.8264	5.457	0.020
	SN	213	-0.0620	0.8341		
	Total	532	0.0640	0.8351		
Progress test Year 3	SP	201	0.2330	0.8028	7.068	0.008
	SN	148	0.0210	0.7582		
	Total	349	0.1431	0.7901		

Abbreviations: pu-GPA = pre-university Grade Point Average; SP = Selection-Positive; SN = Selection-Negative; SD = Standard Deviation.

[a] statistically significant p-values are indicated in bold. pu-GPA was a significant predictor in all three years (Y1, F: 118.438, p=0.000; Y2, F: 102.242, p=0.000; Y3, F: 65.601; p=0.000).

3.4 Discussion

To the best of our knowledge, this is the first study which has looked at the predictive validity of a constructively aligned, outcome-based selection procedure for student performance in the clinical years of medical school. The unique Dutch context of two admissions processes operating in parallel allowed us to compare across two groups: those who were selected using the outcome-based procedure, and their counterparts who were rejected in this procedure and were then admitted via a lottery process which uses only pu-GPA as the basis for selection. We found that students who were selected via the outcome-based procedure outperformed their initially rejected, lottery-entry counterparts on an increasing number of CanMEDS roles and on national progress testing as they progressed through the clinical years of medical school. In other words, the selection procedure not only had predictive value in the clinical phase of medical school, this predictive value increased over time.

One of the most obvious strengths of the current study is that our outcome measures included indicators of actual clinical performance in daily practice rather than focusing solely on the relationship between selection and (a mean of) overall quantitative clerkship grades or rather crude measures like drop-out or study delay[10,11]. Interestingly, in our earlier study where we examined performance on selection with that in the pre-clinical years, we found that the predictive value of the selection procedure was highest for OSCE performance, arguably the pre-clinical outcome measure most closely related to clinical performance[4]. Indeed, drawing on this earlier study, we have identified that it is possible to select, "at the gate", students who perform better at the end of their six years of medical school without losing the predictive value of selection in the early, pre-clinical years.

We were able to use outcome variables (i.e. indicators of progression) which had been carefully designed to align with the CanMEDS-based outcomes framework used in the Netherlands[16]. In one other study, MMIs focusing on non-cognitive skills showed predictive value for clinical performance during clerkships at a more granular level[12].

Other Dutch previous studies have used, for example, extra-curricular activities and found that these are related to clinical achievement[27], whereas other studies found only little[11,35,36], or no value[10,37] of their selection procedures for clinical performance. The comparison of these earlier studies with our own highlights the importance of using the right criterion to which to apply constructive alignment – blueprinting selection, curriculum and assessments to outcomes – when educating students to become competent doctors[7,14,16,19,22,38]. As mentioned earlier, this blueprinting was done in advance of, and independent to the current study, thus avoiding bias.
It is important to note that students of both groups performed well in the clinical phase. This chimes with previous observations that those who apply for medical school are typically relatively homogeneous, and most applicants are capable of completing medical studies, making selection a challenging endeavor[39,40]. However, if the goal of medical schools is to produce the best doctors, our study indicates that constructively aligned selection procedures may identify better-suited students than selecting on pu-GPA only[16]: the Selection-Positive students showed more excellence by the time of graduation (i.e. the grades at T3). We can see that the selected students are much more likely to be excellent at Communicating (21.8% more likely), Collaborating (18.7%), Organizing (16%), Academic activities (13.8%), Advocating health (10.7%) and behaving Professionally (18.6%). These are important results, as they relate closely to later patient care, the ultimate goal of educating a good doctor. The program-specific components of the selection procedure may have identified students who fit better with the educational system and context at MUMS, selecting individuals who may 'blossom' in a PBL environment, compared to those who were rejected in this procedure but entered via lottery[5,7,41]. This suggests that it may be possible to align selection with institutional mission[42], which merits further exploration.

Questions have been raised about the added value of a labor-intensive selection procedure over a simple lottery procedure[37]. The current study clearly shows that there is a positive effect of active selection. Combining this study with previous research[4], we can state that a selection procedure which is aligned with the curriculum, assessment and long-term goals is capable of selecting students who will outperform pu-GPA-selected students throughout medical school. Moreover, this procedure has already been shown to be cost-efficient in the bachelor phase of medical school alone[43], which justifies the effort needed for selection for medical school. Furthermore, the fact that the selection procedure is cost-effective, also means that there is more room to invest in high quality education.
Several additional results are worth pointing out. First of all, the effect of pu-GPA. At the start of their studies, the students started with the exact same starting point: that is, we found no significant difference in pu-GPA between the groups. However, the regression analyses show that throughout the clinical phase there is an additional effect of pu-GPA. Pu-GPA consistently shows an effect for three competences (Medical Expert, Academic, and Professional) but less so for the other competences. Important to note is that the predictive effect of pu-GPA does not take away from the effect of selection: results with and without pu-GPA as covariate are consistent. Furthermore,

selection and pu-GPA seem to be mostly supplemental (i.e. they show incremental value), both show different strengths in terms of predictive value. Another interesting covariate is sex. The effect of sex appears to increase throughout the clinical phase, always favoring the female students. These advantages for women are mostly found in (inter)personal skills or competences: Collaborator at T2 and Communicator, Collaborator and Professional at T3. This is an interesting finding, as we have seen no predisposition in favor of women throughout the selection procedure. Finally, the finding that progress test results significantly differ between the Selection-Positive and Selection-Negative students while no differences for the role of Medical Expert were found may be surprising. However, the role of Medical Expert contains not only the progress test results, but also results and feedback from other knowledge tests and clinical assessments throughout the rotations.

Of course, there are limitations to the current study. We report on one system in one context, where the selection process is a university-specific (local) responsibility. Different selection procedures may be needed for different contexts and goals (e.g. different learning environments at universities, selecting for future demands, or with a specific focus on widening access)[22,42,44]. Nevertheless, where possible, comparisons between institutions as well as explicit comparisons of different holistic selection procedures would be helpful to examine the generalizability of our findings. Furthermore, our own time constraints meant we had to collect data before T3 data were available for the 2013 cohort, causing a smaller sample size for T3. This could be put forward as an explanation on why the effects of the selection procedure increase. However, the lower number of students should have made it more difficult to find significant results, and the significant results actually increase. Furthermore, we found few differences between cohorts generally, and the results that we did find have shown to be very stable. Finally, we lacked information on the demographic characteristics of our students, such as socio-economic status, nationality or non-Western immigration background. In the Netherlands, recent data suggest that there is no overall effect of demographic variables in the current procedures[45]; however, this is contradictory to the situation in other countries[46-51].

In conclusion, this study adds information to the long-lasting debate in the field of selection on what qualities/outcomes can be selected for. Our study showed that careful consideration of intended graduate outcomes, defined using an outcome framework such as the CanMEDS (or another nationally endorsed framework), and clear constructive alignment of selection, curriculum and assessment to these outcomes[5,7,24] was effective in creating a selection procedure predictive of performance in the pre-clinical as well as the clinical phase of medical school.

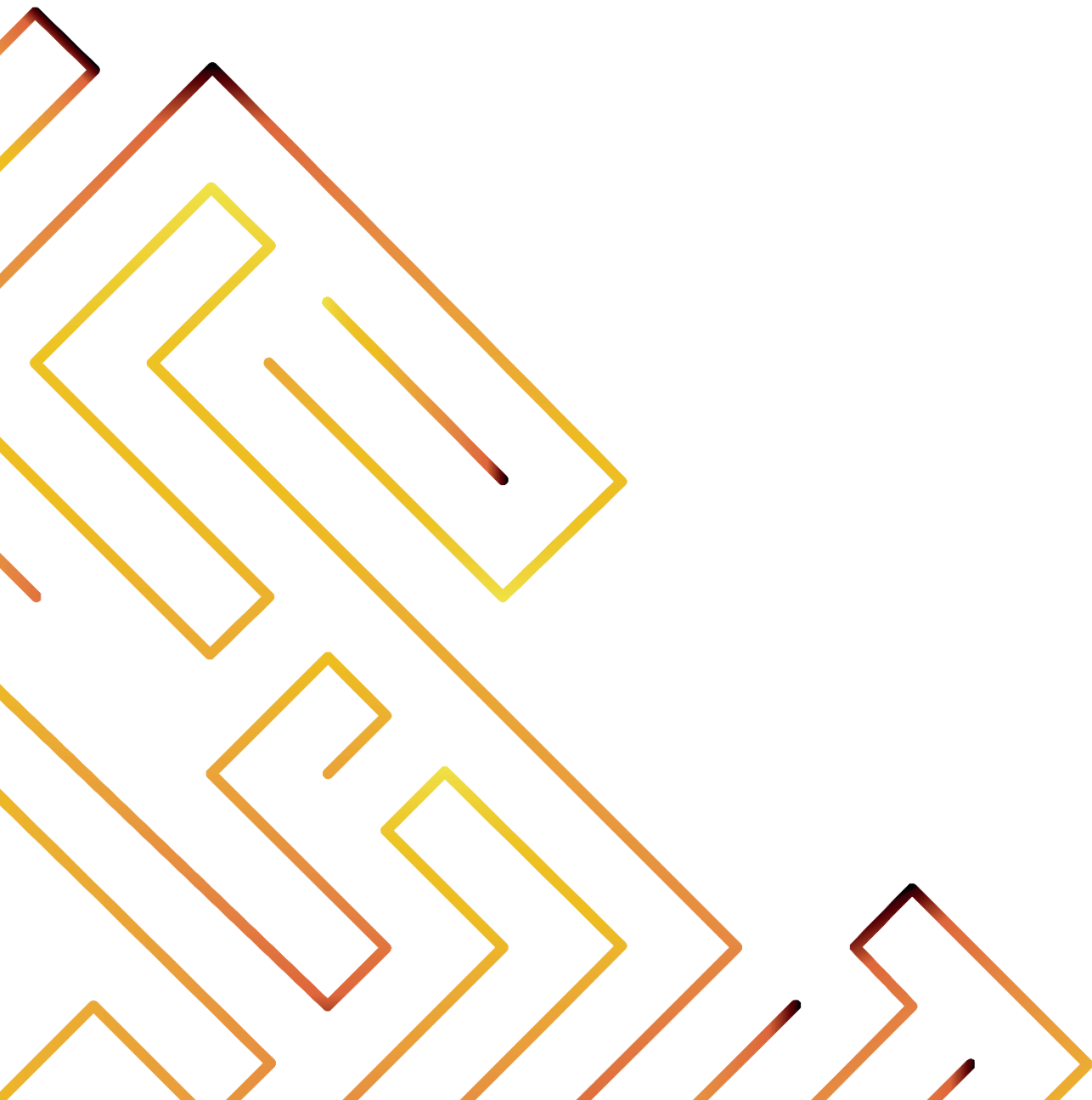
Acknowledgments: The authors wish to thank the late dr. Arno Muijtjens for his indispensable help in the conception of the current study and the statistical support for the previous study that generalized to the current study. The authors also wish to thank Margriet Schoonbrood-Brorens for her support and persistence in the gathering of the data.

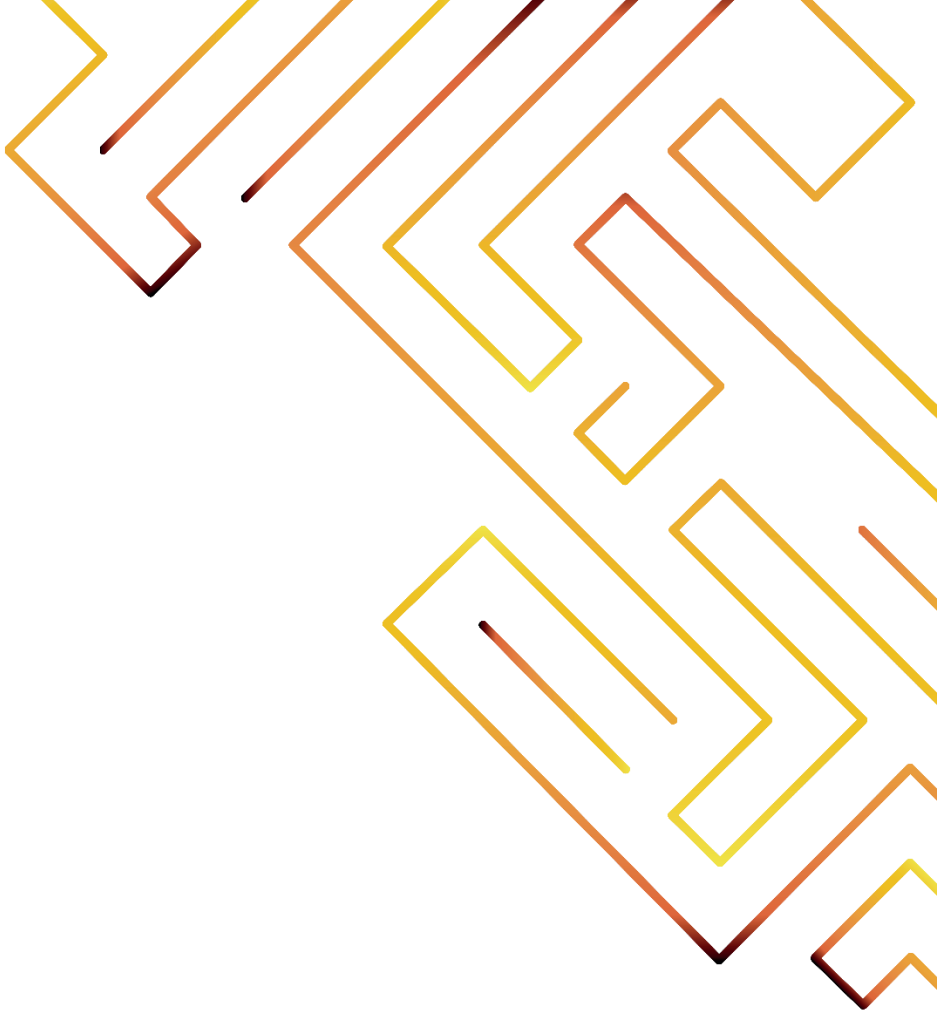
3.5 References

1. Cleland J, Dowell J, McLachlan J, Nicholson S, Patterson F. Identifying best practice in the selection of medical students (literature review and interview survey). General Medical Council.https://www.gmc-uk.org/-/media/about/identifyingbestpracticeintheselectionofmedicalstudentspdf51119804.pdf2012 Published November 2012. Accessed October 2015.
2. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. *Med Educ.* 2016;50:36-60. doi: 10.1111/medu.12817.
3. Prideaux D, Roberts C, Eva K, et al. Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach.* 2011;33:215-223. doi: 10.3109/0142159X.2011.551560.
4. Schreurs S, Cleutjens KB, Muijtjens AMM, Cleland J, Oude Egbrink MGA. Selection into medicine: the predictive validity of an outcome-based procedure. *BMC Med Educ.* 2018;18:214. doi: 10.1186/s12909-018-1316-x.
5. Patterson F, Roberts C, Hanson MD, et al. 2018 Ottawa consensus statement: Selection and recruitment to the healthcare professions. *Med Teach.* 2018;40:1-11. doi: 10.1080/0142159X.2018.1498589.
6. Kreiter CD. A research agenda for establishing the validity of non-academic assessments of medical school applicants. *Adv Health Sci Educ Theory Pract.* 2016;21:1081-1085. doi: 10.1007/s10459-016-9672-y.
7. Patterson F, Zibarras L, eds. Selection and Recruitment in the Healthcare Professions: Research, theory and practice. Cham: Springer Nature Switzerland AG; 2018.
8. Albanese MA, Snow MH, Skochelak SE, Huggett KN, Farrell PM. Assessing personal qualities in medical school admissions. *Acad Med.* 2003;78:313-321.
9. Siu E, Reiter HI. Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. *Adv Health Sci Educ Theory Pract.* 2009;14:759-775. doi: 10.1007/s10459-009-9160-8.
10. Schripsema NR. Effects of medical school admission based on GPA, voluntary multifaceted selection, or lottery on long-term study outcomes. In *Medical student selection: Effects of different admissions processes* [dissertation]. Groningen: Rijksuniversiteit Groningen; 2017.
11. Urlings-Strop LC, Themmen AP, Stijnen T, Splinter TA. Selected medical students achieve better than lottery-admitted students during clerkships. *Med Educ.* 2011;45:1032-1040. doi: 10.1111/j.1365-2923.2011.04031.x.
12. Reiter HI, Eva KW, Rosenfeld J, Norman GR. Multiple mini-interviews predict clerkship and licensing examination performance. *Med Educ.* 2007;41:378-384.
13. Bugaj TJ, Schmid C, Koechel A, et al. Shedding light into the black box: A prospective longitudinal study identifying the CanMEDS roles of final year medical students' on-ward activities. *Med Teach.* 2017;39:883-890. doi: 10.1080/0142159X.2017.1309377.
14. Frank, JR. (Ed). 2005. *The CanMEDS 2005 physician competency framework. Better standards. Better physicians. Better care*. Ottawa: The Royal College of Physicians and Surgeons of Canada. http://www.ub.edu/medicina_unitateducaciomedica/documentos/CanMeds.pdf
15. Frank JR, Snell LS, Sherbino J. *The draft CanMEDS 2015 physician competency framework - series ii.* Ottawa: The Royal College of Physicians and Surgeons of Canada; 2014. http://www.royalcollege.ca/rcsite/documents/canmeds/canmeds-2015-iii-change-record-e.pdf
16. Van Herwaarden CLA, Laan RFJM, Leunissen RRM. *Raamplan Artsopleiding 2009.* Utrecht; 2009. http://www.nvmo.nl/resources/js/tinymce/plugins/imagemanager/files/2009_nvmo_raamplan_artsopleiding_r_laan.pdf
17. Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability Scales: Outlining Their Usefulness for Competency-Based Clinical Assessment. *Acad Med.* 2016;91:186-190. doi: 10.1097/ACM.0000000000001045.
18. Bok HGJ, de Jong LH, O'Neill T, Maxey C, Hecker KG. Validity evidence for programmatic assessment in competency-based education. *Perspect Med Educ.* 2018;7:362-372. doi: 10.1007/s40037-018-0481-2.
19. Wilkinson TM, Wilkinson TJ. Selection into medical school: from tools to domains. *BMC Med Educ.* 2016;16:258.

20. Roberts C, Wilkinson TJ, Norcini J, Patterson F, Hodges BD. The intersection of assessment, selection and professionalism in the service of patient care. *Med Teach.* 2019;41:1-6. doi: 10.1080/0142159X.2018.1554898.
21. Cleland JA, Patterson F, Hanson MD. Thinking of selection and widening access as complex and wicked problems. *Med Educ.* 2018;52:1228-1239. doi: 10.1111/medu.13670.
22. Raffoul M, Bartlett-Esquilant G, Phillips RLJ. Recruiting and Training a Health Professions Workforce to Meet the Needs of Tomorrow's Health Care System. *Acad Med.* 2019;94:651-655. doi: 10.1097/ACM.0000000000002606.
23. Patterson F, Cleland J, Cousans F. Selection methods in healthcare professions: where are we now and where next? *Adv Health Sci Educ Theory Pract.* 2017;22:229-242. doi: 10.1007/s10459-017-9752-7.
24. Stegers-Jager KM. Lessons learned from 15 years of non-grades-based selection for medical school. *Med Educ.* 2018;52:86-95. doi: 10.1111/medu.13462.
25. Conrad SS, Addams AN, Young GH. Holistic Review in Medical School Admissions and Selection: A Strategic, Mission-Driven Response to Shifting Societal Needs. *Acad Med.* 2016;91:1472-1474.
26. Schripsema NR, van Trigt AM, Borleffs JC, Cohen-Schotanus J. Selection and study performance: comparing three admission processes within one medical school. *Med Educ.* 2014;48:1201-1210. doi: 10.1111/medu.12537.
27. Urlings-Strop LC, Themmen APN, Stegers-Jager KM. The relationship between extracurricular activities assessed during selection and during medical school and performance. *Adv Health Sci Educ Theory Pract.* 2017;22:287-298. doi: 10.1007/s10459-016-9729-y.
28. Patterson F, Zibarras L, Ashworth V. Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Med Teach.* 2016;38:3-17. doi: 10.3109/0142159X.2015.1072619.
29. Motowidlo SJ, Hooper AC, Jackson HL. Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *J Appl Psychol.* 2006;91:749-761.
30. Schreurs S, Cleutjens KBJM, Collares CF, Cleland J, Oude Egbrink MGA. Opening the black box in selection. *AHSE.* 2019; Accepted.
31. Cleland JA, Milne A, Sinclair H, Lee AJ. Cohort study on predicting grades: is performance on early MBChB assessments predictive of later undergraduate grades? *Med Educ.* 2008;42:676-683. doi: 10.1111/j.1365-2923.2008.03037.x.
32. Kusurkar R, Kruitwagen C, ten Cate O, Croiset G. Effects of age, gender and educational background on strength of motivation for medical school. *Adv Health Sci Educ Theory Pract.* 2010;15:303-313. doi: 10.1007/s10459-009-9198-7.
33. Cohen-Schotanus J, Muijtjens AM, Reinders JJ, Agsteribbe J, van Rossum HJ, van der Vleuten CP. The predictive validity of grade point average scores in a partial lottery medical school admission system. *Med Educ.* 2006;40:1012-1019.
34. Rosenthal JA. Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research.* 1996;21:37-59.
35. Stegers-Jager KM, Themmen AP, Cohen-Schotanus J, Steyerberg EW. Predicting performance: relative importance of students' background and past performance. *Med Educ.* 2015;49:933-945. doi: 10.1111/medu.12779.
36. Wouters A, Croiset G, Schripsema NR, et al. A multi-site study on medical school selection, performance, motivation and engagement. *Adv Health Sci Educ Theory Pract*. 2017;22:447-462. doi: 10.1007/s10459-016-9745-y.
37. Wouters A. Effects of medical school selection on student motivation: a PhD thesis report. *Perspect Med Educ.* 2018;7:54-57. doi: 10.1007/s40037-017-0398-1.
38. Patterson F, Ferguson E, Thomas S. Using job analysis to identify core and specific competencies: implications for selection and recruitment. *Med Educ.* 2008;42:1195-1204. doi: 10.1111/j.1365-2923.2008.03174.x.
39. Lucieer SM, Stegers-Jager KM, Rikers RM, Themmen AP. Non-cognitive selected students do not outperform lottery-admitted students in the pre-clinical stage of medical school. *Adv Health Sci Educ Theory Pract.* 2016;21:51-61. doi: 10.1007/s10459-015-9610-4.
40. Schripsema NR, van Trigt AM, van der Wal MA, Cohen-Schotanus J. How Different Medical School Selection Processes Call upon Different Personality Characteristics. *Plos One.* 2016;11:e0150645. doi: 10.1371/journal.pone.0150645.

41. Burgess A, Roberts C, Clark T, Mossman K. The social validity of a national assessment centre for selection into general practice training. *BMC Med Edu.* 2014;14:1-11. doi: 10.1186/s12909-014-0261-6.
42. Sklar DP. Who's the Fairest of Them All? Meeting the Challenges of Medical Student and Resident Selection. *Acad Med.* 2016;91:1465-1467.
43. Schreurs S, Cleland J, Muijtjens AMM, Oude Egbrink MGA, Cleutjens K. Does selection pay off? A cost-benefit comparison of medical school selection and lottery systems. *Med Educ.* 2018;52:1240-1248. doi: 10.1111/medu.13698.
44. Conrad SS, Ms, Addams AN, Young GH, PhD. Holistic Review in Medical School Admissions and Selection: A Strategic, Mission-Driven Response to Shifting Societal Needs. *Academic Medicine November.* 2016;91:1472-1474.
45. van den Broek A, Mulder J, de Korte K, Bendig-Jacobs J, van Essen M. *Selectie bij opleidingen met een numerus fixus & de toegankelijkheid van het hoger onderwijs.* Nijmegen: ministerie van OCW; 2018.
46. Griffin B, Hu W. The interaction of socio-economic status and gender in widening participation in medicine. *Med Educ.* 2015;49:103-113. doi: 10.1111/medu.12480.
47. Fielding S, Tiffin PA, Greatrix R, et al. Do changing medical admissions practices in the UK impact on who is admitted? An interrupted time series analysis. *BMJ Open.* 2018;8:e023274. doi: 10.1136/bmjopen-2018-023274.
48. General Medical Council (GMC). *National training survey 2013: socioeconomic status questions.* https://www.gmc-uk.org/-/media/documents/report-nts-socioeconomic-status-questions_pdf-53743451.pdf2013; 2013.
49. Association of American Medical Colleges. *Total enrollment by U.S. medical school and race/ethnicity.* https://www.aamc.org/download/321540/data/factstableb5-1.pdf; 2017.
50. American Medical Association. Diversity in the physician workforce; facts and figures. http://www.aamcdiversityfactsandfigures.org/; 2014.
51. Freeman BK, Landry A, Trevino R, Grande D, Shea JA. Understanding the Leaky Pipeline: Perceived Barriers to Pursuing a Career in Medicine or Dentistry Among Underrepresented-in-Medicine Undergraduate Students. *Acad Med.* 2016;91(7):987-993. doi: 10.1097/ACM.0000000000001020.





CHAPTER 4

Opening the black box of selection

Schreurs S, Cleutjens K, Collares CF, Cleland J, oude Egbrink MGA. Opening the black box of selection. *Advances in Health Sciences Education*. 2019; 1-20 [EPub ahead of print]

Abstract

Medical school selection is currently in the paradoxical situation in which selection tools may predict study outcomes, but which constructs are actually doing the predicting is unknown (the 'black box of selection'). Therefore, our research focused on those constructs, answering the question: do the internal structures of the tests in an outcome-based selection procedure reflect the content that was intended to be measured? Downing's validity framework was applied to organize evidence for construct validity, focusing on evidence related to content and internal structure. The applied selection procedure was a multi-tool, CanMEDS-based procedure comprised of a Video-based Situational Judgement Test (focused on (inter)personal competencies), and a Written Aptitude Test (reflecting a broader array of CanMEDS competencies). First, we examined content-related evidence pertaining to the creation and application of the competency-based selection blueprint and found that the set-up of the selection procedure was a robust, transparent and replicable process. Second, the internal structure of the selection tests was investigated by connecting applicants' performance on the selection tests to the predetermined blueprint using Cognitive Diagnostic Modeling (CDM). The data indicate 89% overlap between the expected and measured constructs. Our results support the notion that the focus placed on creating the right content and following a competency-blueprint was effective in terms of internal structure: most items measured what they were intended to measure. This way of linking a predetermined blueprint to the applicants' results sheds light into the 'black box of selection' and can be used to support the construct validity of selection procedures.

4.1 Introduction

The purpose of medical school selection is to recruit students who will perform well at medical school as well as in their future career as a doctor (Bandiera et al., 2015). To achieve this, many selection procedures are now outcome-based (e.g. Frohlich, Kahmann, & Kadmon, 2017; Patterson et al., 2018; Prideaux et al., 2011; Schreurs, Cleutjens, Muijtjens, Cleland, & Oude Egbrink, 2018; Terregino, McConnell, & Reiter, 2015): 'beginning with the end in mind'. To this purpose, the cognitive and (inter)personal competencies or qualities needed throughout the study program and in future work are integrated as constructs into the selection process (Cleland, Dowell, McLachlan, Nicholson, & Patterson, 2012; Patterson, Knight, et al., 2016). Selection procedures typically consist of multiple tools, with each university individually choosing and combining constructs (i.e. competencies or qualities) and tools (Cleland et al., 2012; Patterson, Knight, et al., 2016; Schreurs, Cleutjens, et al., 2018), often without proper justification. Up to now, most research has focused on the utility of individual selection tools, showing for example that unstructured interviews are neither reliable nor valid, while Multiple Mini Interviews (MMIs) show better psychometric qualities; that previous academic attainment predicts later academic attainment; that there is a plethora of written tests whose psychometric qualities vary with each variation in format and construct; and that the Situational Judgment Test (SJT) may be a useful tool in medical school selection (e.g. Cleland et al., 2012; Patterson, Knight, et al., 2016; Patterson et al., 2018; Prideaux et al., 2011).

Because selection research has typically focused on the qualities of one particular tool or method in its own right, only few studies have looked at combinations of tools as applied by many medical schools. Studies investigating combined tools have typically focused on incremental validity: whether one tool has predictive value above and beyond another tool (McManus, Dewberry, Nicholson, & Dowell, 2013; Patterson, Rowett, et al., 2016; Schreurs, Cleutjens, et al., 2018; Tiffin et al., 2016). Moreover, to date, there has been no systematic consideration of which constructs (e.g. collaboration or empathy) are actually assessed in medical school selection procedures, and whether this is in line with what was intended from their outcome-based focus (Christian, Edwards, & Bradley, 2010; Wilkinson & Wilkinson, 2016). This means that selection may be considered a sort of 'black box' (Kreiter, 2017; Kulasegaram, 2017; Lievens, Peeters, & Schollaert, 2008), a paradoxical situation in which selection tools may predict outcomes but which constructs are actually doing the predicting is uncertain (Cleland, Dowell, Nicholson, & Patterson, 2014; Crossingham et al., 2011; Tiller et al., 2013). It is essential to know more about what is *actually* being measured (i.e. the construct validity of selection; e.g. Christian et al., 2010; Hecker & Norman, 2017; Kreiter, 2017; Kulasegaram, 2017; Patterson, Cleland, & Cousans, 2017) in order to determine whether the intended constructs are measured. Research on this subject is sorely missing (Hecker & Norman, 2017; Kulasegaram, 2017), and would not only greatly benefit the defensibility of selection procedures (Kreiter, 2017), but would also be a first step in the direction of creating more theory-based selection procedures (Patterson et al., 2018; Prideaux et al., 2011).

Moreover, conducting studies on construct validity yields practical implications for selection. For example, if the intended constructs are not measured and the predictive value is insufficient, the selection committee should go back to the drawing board, since the procedure is neither effective nor defensible or fair (Patterson & Zibarras, 2018). Alternatively, if there is predictive value but the intended constructs are not measured, where the predictive value is coming from should be investigated in order to avoid 'being reliably wrong' (Patterson & Ferguson, 2012) and measuring an unrelated construct that, by chance, correlates with study success (e.g. shoe-tying-skills could be predictive of medical school performance, but cannot defensibly be used as a selection tool). All in all, research on construct validity can help the field of selection for medicine move forward in terms of theory and practice.

One way to systematically assess whether we are measuring the constructs we want to measure, and to investigate possibilities for improving validity, is by applying a validity framework. Validity frameworks provide guidelines on how and what information to gather on assessment methods (in this case the medical school selection process, given selection can be considered the first assessment in medicine; Cleland et al., 2012) to investigate whether an assessment is applicable for the proposed use. These frameworks also stimulate researchers to view their assessment from different perspectives and take various sources of information into account. Examples of the frameworks that are used within the field of medical education are Kane (1992; also see Cook, Brydges, Ginsburg, & Hatala, 2015), Messick (1995), and Downing and the Standards for educational and psychological testing (AERA, APA & NCME, 2014; Downing, 2003). Each of these frameworks overlap to a certain degree. Downing explicitly intended his framework to inform research on assessment within medical education, and his framework and the closely related 'Standards' have been used in several studies within (Kelly & O'Flynn, 2017; Mink et al., 2018) and outside (Sorrel et al., 2016) of medical education.

The ultimate aim of the current study was to address the gap in knowledge with respect to the construct validity of medical school selection procedures. This was done by focusing on the content of the procedure on the one hand, and the internal structure of the procedure on the other. The specific question to be answered in the current study was: do the internal structures of the tests in the second round of the selection procedure reflect the content that was intended to be measured? We selected Downing's framework as the means to organize the evidence for construct validity of the tools in the second round of a multi-tool, outcome-based selection procedure (more explanation on the selection procedure itself is provided in the methods section).

4.2 Methods

4.2.1 Context

This study was performed at Maastricht University Medical School (MUMS) in the Netherlands. MUMS administers a multi-tool, outcome-based selection procedure.

The selection procedure consists of two rounds containing three tools (hence, multi-tool). In the first stage, applicants complete a pre-structured online portfolio, focusing on previous academic attainment, extracurricular (distinguishing) abilities, and their fit with problem-based learning and the MUMS medical curriculum. This first stage is used as a broad-brush pre-screening to limit the amount of applicants that proceed to the second part of the procedure. The second stage, a selection day at MUMS, consists of a Video-based Situational Judgment Test (V-SJT) and a Written Aptitude Test (WAT), both of which contain items aimed at measuring predetermined competencies (see below and chapter two). In this second stage, a more fine-grained selection takes place. The current study focused on the second stage, and therewith on two tools within the selection procedure at MUMS: the V-SJT and the WAT.

The MUMS selection procedure is outcome-based, as it is based on a blueprint of competencies derived from the CanMEDS framework, a well-known and internationally accepted outcome framework for medical school (Frank, 2005; van Herwaarden, Laan, & Leunissen, 2009). The CanMEDS describe seven roles: Medical Expert, Communicator, Collaborator, Organizer (Leader in the 2015 edition), Health Advocate, Scholar and Professional. In the second round of the selection procedure, the V-SJT focuses on the (inter)personal competencies in the CanMEDS, while the WAT more broadly assesses aptitude for (inter)personal as well as more cognitively loaded CanMEDS roles. The applicants' results on both tests were converted into z-scores, averaged per test, and the means of the two tool-averages were used to create the rank order of applicants, on which they were selected or rejected.

The MUMS selection procedure as a whole has been studied previously for its predictive value and cost-effectiveness (Schreurs, Cleland, Muijtjens, Oude Egbrink, & Cleutjens, 2018; Schreurs, Cleutjens, et al., 2018). As stated above, the current study focused on what is actually measured during the second stage (i.e. V-SJT and WAT) of the selection procedure. To this purpose, data from all 547 candidates in the second round of the 2016 selection procedure were investigated.

4.2.2 Ethical approval

Applicants were asked to give their informed consent for the use of their selection and assessment data for research purposes. It was made clear that not taking part in the study would not adversely influence their progression. Participant data were anonymized before they were shared with the research team. The study was approved by the Ethical Review Board of the Netherlands Association for Medical Education (NVMO; file number 2018.8.5).

4.2.3 Validity framework

We consulted the literature on contemporary validity frameworks in order to assess the validity of the selection procedure. As stated before, we chose Downing's framework to organize the validity evidence in the current study, since it is applicable to assessment systems such as a selection procedure. In brief, Downing (2003) defines five sources of evidence for construct validity: content (evidence supporting the

content of the assessment, such as the thorough development of its blueprint), response processes (evidence showing that the test-takers do in fact employ the processes that were intended to be employed, for example as measured by eye-tracking or trace data), internal structure (evidence related to the structure of the test, for example item quality and factorial structure), relationship to other variables (evidence relating the performance on the test to performance on another test, which should have the expected relationship), and consequences (evidence related to the impact the score on the test has on the test-taker and in how far these are intended and positive/negative). For more information on the framework, see Downing (2003). The research question set forth for the current study was answered by focusing on two of these sources of evidence: content and internal structure. Content evidence pertains to the competency-based blueprint used to develop the selection procedure, while internal structure evidence relates to the extent to which the blueprint is reflected in the applicants' results. Details on the approaches taken to investigate these two sources of evidence are below.

4.2.4 Content

A qualitative approach based on document analysis, attending the selection committee meetings and checking and confirming the results with the head of the selection committee was used to establish an evidence-base for validity concerning the content of the selection procedure. Document analysis was used to understand the manner in which the blueprint for the selection was established. Furthermore, the head of the selection committee provided additional information on this process, while the lead author of this study attended the selection committee meetings in which the content of the procedure was discussed.

Information was gathered on the development of the blueprint used to design the procedure, the Subject Matter Experts (SMEs) who are members of the selection committee and, hence, in charge of the creation and employment of the blueprint, the relationships between constructs and items, and the representativeness of the items for applicants. The SMEs created the items and paid specific attention to the representativeness of the items for the construct. After reaching consensus on the content and questions in the selection items, the SMEs wrote answer keys to each question: possible answers to those questions and how many points those answer options would result in. If applicants provided an unexpected answer, this was related to the answer key and in doubt, such an answer was discussed in a committee meeting. In addition, to determine the representativeness of the items for applicants, a post-selection questionnaire was employed to investigate whether the applicants found the items representative for what they thought should be assessed in a selection procedure. Participants in the second round of the selection procedure received the post-selection questionnaire after finishing the V-SJT and WAT. It contained 31 questions related to the organization of the selection day, the information that had been provided beforehand, whether the applicants thought the items in the tests were relevant for future medical students and doctors and whether the assessment was complete (i.e. whether they thought there were questions left

unasked that would have been important to include in the selection procedure). Since the focus of the current study was on the selection procedure's representativeness of the items, the reactions to the following statements were taken into account: (1) "The assignments offered me the possibility to present an accurate portrayal of my abilities" (for the V-SJT and the WAT separately), and (2) "In my opinion, the selection procedure as a whole encompasses all aspects needed for the identification of the best suited candidates for the Bachelor of Medicine" (one overarching statement for the entire procedure). Reactions to these statements could be provided on Likert scales of 1 through 5 (1: completely disagree, 2: disagree 3: neutral, 4: agree, and 5: totally agree).

4.2.5 Internal structure

To assess internal structure, data were gathered on the applicants' performance on the V-SJT and WAT. The applicants were first graded according to the answer keys determined by the SMEs. In order to enable comparison of the performances on items for the present study, the raw scores were transformed into z-scores per item (a standardized score taking into account the performance of all other applicants with a mean of zero and a standard deviation of one).

Related to internal structure, internal consistency is a very important characteristic. However, Cronbach's alpha has been criticized for using a tau-equivalent approach to estimate reliability (Peters, 2014) and its inability to cope with multi-dimensionality (Sorrel et al., 2016). Since the data used in the current study are multidimensional in two ways (i.e. there are multiple constructs being measured by different items within each test [multidimensionality *between* items] and the items themselves are measuring multiple constructs at the same time, in different compositions per questions [multidimensionality *within* items]), Cronbach's alpha was considered to be inadequate. Furthermore, the Omega coefficient can account for the multidimensionality *between* items, but not *within* items (Dagnall, Denovan, Parker, Drinkwater, & Walsh, 2018). Thus, in this study, the internal consistency could not be based on Cronbach's alpha or the Omega coefficient. An analysis method that is in fact capable of dealing with multidimensionality between as well as within items is the G-DINA, an analysis in the family of Cognitive Diagnostic Models (CDM). The newest version of G-DINA provides a test level accuracy and attribute level accuracy; these results are provided later. A more detailed description of CDM and G-DINA is given below.

Inter-rater and intra-rater reliability was calculated from historical and current data from round two of the MUMS selection procedure. When the selection procedure was being set up in 2011, both types of reliability were assessed formally. Inter-rater reliability was established by having multiple assessors score the same question for a multitude of applicants, and calculating the correlation between these assessments. Intra-rater reliability was established by having single assessors score all applicants on one question, and then having them go back to the answers two weeks after they had

scored them first, and establishing the correlation between the scores the first and second time these applicants were scored.

An initial exploration of the properties of the items was done using descriptive statistics (i.e. means, standard deviations, item-total correlations) and Cognitive Diagnostic Modeling (CDM, see later in this section), in order to gather information about the overall functioning and fairness of the items. The second-round items were initially analyzed for the group as a whole and later also for subgroups for which no differences were expected (gender, age), in order to investigate possible Differential Item Functioning (DIF). The effect of pre-university Grade Point Average (pu-GPA), which may affect performance, was also investigated. DIF for gender, age and pu-GPA was conducted using independent samples t-tests and linear regression analyses. The critical p-value was set at 0.05. For these analyses, no correction for multiple comparisons was applied as the goal was not to find statistical significance, but to check whether there was possible DIF, i.e. whether there were differences of educational significance (e.g. if gender would determine all scores and therewith who gets selected, this should be changed immediately). However, because of the possibility of finding significant differences by chance, the results should be considered critically.

Next, to answer the main research question, i.e. to identify whether the constructs set forth in the selection blueprint are in fact the constructs measured in the selection process, data on the applicants' performance were linked with blueprint data. Classical Test Theory is commonly used for this task (e.g. Kiessling et al., 2016; Lievens et al., 2008; Patterson et al., 2012). However, as stated before, because of the multidimensionality between as well as within items, inherent to selection tools, Classical Test Theory (e.g. Cronbach's alpha) is inadequate (Sorrel et al., 2016). To overcome this issue, we applied an alternative test theory, Cognitive Diagnostic Modeling (CDM), as this test theory (to the best of our knowledge) is the only one capable of coping with multidimensionality between as well as within items. CDM is comprised of a family of multidimensional categorical-latent trait models that allow the use of latent variables for assessment tools that contain items that measure more than one dimension concurrently (Garcia, Olea, & De la Torre, 2014). In other words, CDM is capable of finding latent variables when there is multidimensionality in the data, both between and within items. CDM is related to Confirmatory Factor Analysis: the structure is provided, and CDM looks at whether that structure is indeed found in the data, or whether alterations to the structure make more sense on the basis of the data provided. Thus, CDM is a confirmatory technique which requires a pre-specified blueprint.

CDM requires two independent sources of input. The first is a so-called Q-matrix, i.e. the abovementioned blueprint, which tells the model which competencies were planned to be assessed in which items. This Q-matrix is tested for accuracy and alterations are proposed; only the constructs already in the blueprint can be 'found' by the analysis (i.e. the analysis does not search for additional constructs). The second

source is the data on applicant performance. Applicant performance is supplied per item, meaning that even if there were four constructs being measured in one item, there was only one score for that item. It is up to the CDM to disentangle the performances in different items measuring different constructs. Importantly, the applicant performance data must be either binary or ordinal.

In summary, in CDM, the expected structure of the latent variables is provided to the CDM, and this structure is tested. Suppose we have a relatively simple test consisting of ten questions measuring three competencies, X, Y and Z. If a test-taker scores relatively high on items in which we intended to measure X and Y but low on items measuring Z, the model can deduce how this test-taker will score on each specific item based on the results on all items. We can assume that the level of competency in the test-taker does not change during the test; hence, if there is an item that is supposed to measure X and Y, but the test-taker scores low, the item may not measure what it is intended to measure. If at the same time another test-taker scores high, although that second test-taker usually only scores high on Z, the model will propose that this specific item may not be measuring X or Y, but instead is measuring Z. The CDM also creates a file with a grid containing all test-takers and constructs, in which it determines which test-takers are capable in which constructs. For more information on CDM, practical guides to usage and syntax, the reader is referred to George (2015; 2016) or Ravand & Robitzsch (2015); for examples of the use of CDM the reader is referred to Garcia et al. (2014) or Sorrel et al. (2016).

The specific model from the CDM family used in this study is the G-DINA model, a generalization of the "deterministic inputs, noisy and gate" (DINA) model (Ravand & Robitzsch, 2015). In G-DINA, each combination of latent variables is called a latent group, which represents one reduced attribute vector and has its own associated probability of success. This allows the G-DINA to paint a more realistic picture of the proportion of variance accounted for each dimension in relation to the original DINA model. This model has been used successfully in competency-based SJTs in areas other than medical education (Garcia et al., 2014). In the current study, a saturated G-DINA model was applied to the V-SJT as well as the WAT. These analyses were conducted separately because the tests are independent from each other. This choice was further supported using model fit indices: AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) model fits. The sample size in the current study (n=547) was too low for more specific models (de la Torre & Lee, 2013) or for analysis of more than two levels (i.e. ordinal data) within the G-DINA; over a thousand test-takers would have been needed for either. Therefore, the applicants' performance in our procedure was converted from z-scores per item (which is how the scores were handled in practice) to 0/1 scores (negative z-scores become 0 and positive z-scores become 1). To our knowledge, this is the first study on medical school selection applying CDM to obtain validity evidence.

The statistical packages used to analyze the descriptive statistics and Differential Item Functioning was SPSS version 24 for Windows (IBM statistics). The package used for

the CDM analyses was the programming environment R (www.r-project.org), specifically the G-DINA package (Ma & De La Torre, 2017; version 2.4.0).

4.3 Results

4.3.1 Content

The team of SMEs (n=8) that formed the selection committee consisted of experts in education and medicine. Most SMEs had multiple roles, including university teacher, medical doctor (hospital or general practice), psychologist, educationalist, study advisor and/or researcher. Furthermore, one bachelor (i.e. pre-clinical) student representative was part of the selection committee. Together, the selection committee comprised all expertise considered necessary, enabling them to create a holistic selection procedure assessing all important competencies. Among the SMEs were also experts on assessment item creation. The team of SMEs was responsible for the representativeness of items and construct-coverage. SME decisions were made during the selection committee meetings on the basis of (discussion until) complete agreement between the members.

The internationally-recognized competency framework CanMEDS (Canadian Medical Education Directives for Specialists; Frank, 2005) and its Dutch derivative (van Herwaarden et al., 2009) were used to define the blueprint of the selection procedure. These outcome frameworks describe the end terms of medical school, a level the applicants have not yet achieved. Therefore, the team of SMEs translated the CanMEDS-competencies into so-called derived competencies applicants may already possess at bachelor entry-level, and which can be measured in a selection procedure. The translation took place by first thoroughly inspecting the CanMEDS competencies. Several meetings were used to gather clinical and medical-school related situations representative for these competencies to inform the content of the blueprint (Motowidlo, Ghosh, Mendoza, Buchanan, & Lerma, 2016; Patterson et al., 2010; Patterson, Ferguson, & Thomas, 2008) and group them into clusters, until unanimous agreement was reached within the group of SMEs. These clusters then formed the basis of the derived competencies. In an iterative process, the selection committee discussed these derived competencies, how they should be defined and to what extent they should be measured. The resulting derived competencies were Transfer (i.e. knowledge and information integration), Textual skills, Reasoning, Communication, Collaboration, Organization, Medical and Societal Consciousness, Ethical awareness, Empathy and Reflection. Importantly, these derived competencies always remain central to discussions within the committee while creating assignments for the selection procedure, consciously mapping the assignments to the blueprint. The derived competency Communication, defined as “related to effectively conveying a message, either in a spoken or written manner”, is measured by all open-ended questions in the selection procedure and was, therefore, not included as a separate, distinguishable competency in the current study. The derived competencies, definitions and explanations are summarized in Table 4.1. The goal of the MUMS selection procedure was to measure these derived competencies.

Next, tools capable of assessing aptitude for the derived competencies were sought, leading to the application of a V-SJT based on the CASPer (Computer-based Assessment for Sampling Personal characteristics, using short video fragments; Dore, Reiter, Kreuger, & Norman, 2017). The content of the V-SJT was adapted to the Dutch context, befitting the problem-based learning applied at MUMS. In contrast to the original CASPer, the V-SJT applied in the MUMS selection procedure applied an open-ended format (which is why the SMEs created answer keys before checking the applicants' answers). In addition, a written aptitude test (WAT) was developed, which follows the V-SJT format as closely as possible (i.e. open-ended, semi-structured questions relating to real-life situations). The number of items in which a construct (i.e. derived competency) was assessed is also shown in Table 4.1. Importantly, all-but-one constructs were assessed in multiple items, and the vast majority of items assessed multiple constructs; the combination of constructs measured per item varied. By doing so, the SME team ascertained the representation of the derived competencies in the assignments they generated. Total test duration and the number of items needed to achieve a reliable picture was based on assessment literature and CASPer and MMI experiences, indicating how many items/stations are needed to achieve a reliable picture (Dore et al., 2017; Knorr & Hissbach, 2014; Thomson, Anderson, Haesler, Barnard, & Glasgow, 2014; van der Vleuten & Schuwirth, 2005): 90 minutes for 11 items in the V-SJT and 75 minutes for 8 items in the WAT.

At the end of the selection day, several students stated that there was a lot of time pressure, and that they could not finish all assessments. Therefore, this effect was assessed and included in our analyses (see later). The time pressure effect was more apparent for the V-SJT than for the WAT, because the V-SJT can only be filled in assignment by assignment; there is no skipping or returning to a question: applicants have to answer the question as they see it. The WAT is a paper test that applicants can browse through, possibly decreasing the effect of time pressure. After the selection process, 458 of the applicants answered the questions in the post-selection questionnaire related to the representativeness of items in the procedure positively. The mean score on the statement whether the V-SJT offered the possibility to present an accurate portrayal of abilities was 3.9 (SD=1.0; 95% CI: [3,80 ; 4,00]) on a five-point Likert scale; for the WAT this result was 3.4 (SD=0.9; 95% CI: [3,31 ; 3,48]). The overarching statement on whether the selection procedure encompassed all aspects needed for the identification of the best suited candidates scored 3.6 (SD=0.9; 95% CI: [3,52 ; 3,68]).

4.3.2 Internal Structure

The test level accuracy (i.e. in how far the test as a whole, including all items, categorizes each applicant into the right category of either possessing the competencies or not) given by the saturated G-DINA analysis was 0.72 for the V-SJT and 0.20 for the WAT. This means that the accuracy was moderate (i.e. between 0.7 and 0.9) for the V-SJT (Swets, 1988), but very low for the WAT. However, the data on the attribute level accuracy (i.e. in how far the specific items related to the singular competencies are capable of classifying applicants into categories of either possessing

a specific competency or not) tells a different story: the results were acceptable for the V-SJT (Ethical awareness, 0.87; Empathy, 0.86; Reflection, 0.90; Medical and Societal Consciousness, 0.84; Collaboration, 0.99) and also for the WAT (Ethical awareness, 0.67; Reflection, 0.67; Medical and Societal Consciousness, 0.67; Transfer, 0.89; Textual Comprehension and Reasoning, 0.78; Organization, 0.82). As stated before, the V-SJT focuses on the more (inter)personal competencies in the blueprint, while the WAT focuses on a broader array of competencies, including the more cognitively loaded competencies. It is likely that this broadness of the WAT caused the low test level accuracy: applicants scoring high on the (inter)personal competencies may have scored low on the more cognitively-loaded ones, or the other way around, while others may have scored high or low on both, diminishing the overall test accuracy.

In the first year the current selection procedure was executed (2011), both the inter- and intra-rater reliability were determined and appeared to be >0.95. Given the low interrater variability, this was not formally assessed in later years. To ensure reliability, intra-rater reliability was assessed across all five subsequent years and has been consistently ≥0.98.

Table 4.2 shows the item functioning results per test. While the applicants' performance on the individual items differed somewhat, scores on the V-SJT items seemed to support the students' suggestion that there was a time pressure effect, especially in the last two items. In the WAT (where applicants could browse through the test), this effect was less obvious. To determine whether there was a real time effect, an Omega analysis was conducted (solely for this purpose) for both tests. A time pressure effect was found for the last three assignments in the V-SJT and for the last two assignments in the WAT. Furthermore, the G-DINA found acceptable accuracies for time pressure (0.99 and 0.69 for the V-SJT and WAT, respectively). Therefore, time pressure was included as a construct in later analyses.

Applicants' chances of getting items right through *guessing* were mostly low (under 0.5). Also, the chance an applicant possesses the competencies that are measured in an item but still got it wrong (i.e. *slipping*) were mostly low, except for the first item in the V-SJT and the fifth one in the WAT. For all items but the first V-SJT item, the item-total correlations were acceptable.

Table 4.2 also shows the results of the DIF analyses: the only factor increasing overall performance in the selection procedure was pu-GPA. Gender and age did not affect the overall performance throughout the selection procedure, as their effects outweigh themselves (two items in favor of men, two in favor of women; one in favor of older applicants, one in favor of younger applicants). The effect of pu-GPA was positive for six of the 19 items in the selection procedure. Mostly, these were items with high cognitive load (e.g. finding appropriate responses and ordering them or combining multiple bits of information to get to the correct answer), or items closely resembling high school content (e.g. textual comprehension or mathematical questions).

Table 4.1: Translation of the CanMEDS competencies into a blueprint of derived competencies for the selection procedure

CanMEDS	Derived competencies	Definition	Relation to items/example	Items
Medical expert & Scholar*	Transfer[1]	Integrating prior knowledge with new information	Text provides new information on a medical subject, must be combined with secondary school knowledge to find an answer	3
	Textual skills & Reasoning	Textual comprehension and structuring skills	Reading comprehension; structuring given information into charts/models	3
		Verbal and inductive reasoning (fluid intelligence)	Task like Raven's matrices (Engle, Tuholski, Laughlin, & Conway, 1999)	
Communicator	Overall communication[2]	Skills related to effectively conveying a message, either in a spoken or written manner	Related to all items, as each item required narrative, written answers: having students actively express themselves towards the assessors	19
Collaborator	Collaboration	Interpreting and responding to (non)verbal communication of others	Related to shared decision-making (collaboration with patients) or PBL small-group sessions (collaboration with peers)	3
Manager	Organization	Planning and time-management skills	An organizational task is presented, students write down which steps to take and how to prioritize; time pressure is induced by length of the test	1
Health advocate	Medical and Societal Consciousness	Awareness of profound developments and whether they can view these from multiple angles	Items concern manners in which to increase well-being, such as advising patients; or communicating given developments to family members	10
Professional	Ethical awareness	Ability to think about and choose a course of action, and provide rationales	Dilemmas are provided, applicants are asked to choose a side, and provide arguments underpinning this choice	9
	Empathy	Degree to which applicants are able to put themselves in someone else's shoes	Confronted with poignant situations (e.g. terminally ill patient) and asked to expand upon how the patients and loved ones are feeling and coping	7
	Reflection	Ability to think about and consider (own) actions and skills	Applicants are asked to remember the last time they received feedback and reflect on how they responded and what they did with the feedback	12

1 knowledge and information integration; related to applying knowledge as in the role of medical expert and "creation, dissemination, application and translation of medical knowledge" as in the role of scholar (Frank, 2005), 2 including strength of written arguments

*combination of two CanMEDS competencies

Table 4.2: Item functioning statistics for the applicant group as a whole (n=547) and differential item functioning assessed for gender, pu-GPA and age

		Mean score % (SD)	Guessing parameter[1]	Slipping parameter[2]	Item-total correlation	Gender[3] t (p-value)	pu-GPA[4] F (p-value)	Age[5] F (p-value)
V-SJT	1	64.06 (17.80)	0.30	0.99	0.04	2.02 (0.04)*	2.26 (0.13)	0.00 (0.98)
	2	62.01 (20.46)	0.28	0.22	0.22	2.66 (0.01)**	2.11 (0.15)	0.19 (0.66)
	3	60.46 (14.52)	0.19	0.00	0.21	-0.68 (0.50)	4.24 (0.04)*	2.41 (0.12)
	4	72.30 (19.80)	0.44	0.00	0.17	1.35 (0.18)	3.03 (0.08)	0.79 (0.38)
	5	68.34 (20.68)	0.00	0.00	0.15	1.96 (0.05)	1.35 (0.25)	0.59 (0.44)
	6	75.88 (18.30)	0.00	0.18	0.23	1.14 (0.25)	10.09 (0.00)**	7.52 (0.01)**
	7	41.97 (19.46)	0.05	0.00	0.36	1.46 (0.15)	0.05 (0.83)	1.32 (0.25)
	8	41.43 (23.48)	0.03	0.00	0.44	1.01 (0.32)	0.11 (0.74)	0.54 (0.46)
	9	42.48 (30.91)	0.00	0.03	0.59	0.16 (0.87)	0.78 (0.38)	0.02 (0.90)
	10	29.43 (29.85)	0.05	0.00	0.43	0.98 (0.33)	0.55 (0.46)	0.41 (0.52)
	11	14.88 (23.08)	0.01	0.00	0.45	0.72 (0.47)	0.43 (0.52)	0.38 (0.54)
Written	1	48.74 (21.23)	0.38	0.44	0.21	-2.81 (0.01)**	1.34 (0.25)	1.05 (0.31)
	2	44.66 (13.77)	0.42	0.36	0.20	-0.33 (0.74)	0.93 (0.34)	10.54 (0.00)**
	3	61.29 (15.42)	0.01	0.04	0.21	1.01 (0.31)	0.95 (0.33)	0.43 (0.51)
	4	35.66 (30.07)	0.07	0.12	0.25	-3.05 (0.00)**	3.97 (0.05)*	0.36 (0.55)
	5	40.35 (30.85)	0.26	0.55	0.17	-1.29 (0.20)	2.31 (0.13)	0.42 (0.52)
	6	43.48 (16.44)	0.20	0.38	0.20	0.62 (0.54)	5.63 (0.02)*	2.60 (0.11)
	7	41.97 (19.46)	0.67	0.00	0.25	0.06 (0.95)	11.04 (0.00)**	0.28 (0.60)
	8	41.43 (23.48)	0.02	0.01	0.21	0.78 (0.44)	5.64 (0.02)*	0.74 (0.39)

* Significant at $p<0.05$, ** Significant at $p<0.01$. [1] Guessing is the probability that a respondent responds correctly to the item although he or she has not mastered all the required attributes; analyzed using the G-DINA model with 0 is low and 1 is high. [2] Slipping is the probability that a respondent responds incorrectly to the item although he or she has mastered all required attributes; analyzed using the G-DINA model with 0 is low and 1 is high. [3] Independent samples t-test with 0=female, 1=male; positive t-values represent higher mean scores for women than for men, negative t-values represent higher mean scores for men than for women.[4] Linear regression analysis with pu-GPA as independent variable and performance on each item as dependent variable. All significant results for pu-GPA are in favor of higher pu-GPAs. [5] Linear regression analysis with Age as independent variable and performance on each item as dependent variable; for item 6 on the V-SJT the older students had an advantage, while they had a disadvantage on item 2 of the written test.

Finally, to gather evidence for the validity of the tools within the selection procedure (i.e. V-SJT and WAT) based on their internal structure, a saturated G-DINA model was applied (Sorrel et al., 2016). The test statistic used to determine which specific model was applied per attribute was based on the Wald test, the decision rule being "simpler model + largest p value rule at 0.05 alpha level; adjusted p values were based on Bonferroni correction" (Ma & De La Torre, 2017). The results of the Q-matrix validation by G-DINA are shown in Table 4.3 (Time pressure was added to the blueprint; see above). It shows which competencies were expected and measured in which items of each test; Collaboration and Empathy were only assessed in the V-SJT, while Transfer, Textual comprehension, Reasoning, and Organization were only assessed in the WAT. The data consisted of only zeroes and ones, with zeroes meaning that a competency is not expected and measured in that item and a one that a competency is expected and measured in that item.

The AIC and BIC model fits were calculated for each test. For the V-SJT, they were 7818.46 and 8675.05, respectively, and for the WAT they were 6198.89 and 6917.73, respectively. By applying the changes to the Q-matrices suggested by the G-DINA, the model fit does not increase significantly; therefore, the original Q-matrix and suggested changes are provided. A G-DINA analysis of both tests together would result in a drastic decrease of the model fit (AIC=14166.52 and BIC=17123.67), which is logical as they are simply different tests. Because of these reasons, the V-SJT and WAT were analyzed separately.

As shown in Table 4.3, SME predictions were overruled by the G-DINA analysis in only 14 of the 122 cases (i.e. 14 of the 122 predictions of which derived competencies were and were not measured by which items were incorrect according to the G-DINA analysis); this is illustrated by the fact that two numbers are shown with an arrow between them. In these cases, items measured other and/or additional competencies than expected by the SMEs. For example, the fifth item in the V-SJT was in fact not measuring Reflection, but did measure Collaboration, the other competency that was intended to be measured. The change was the other way around for the fifth item in the WAT; this item was shown to not only measure Textual comprehension and reasoning, but also Transfer and Organization. All in all, these results show that there is an overlap between expected and measured competencies of 92% for the V-SJT and of 84% for the WAT, adding up to an overlap of 89% between the predetermined, expected Q-matrix for the overall selection procedure and the matrix as validated using G-DINA. Furthermore, the majority of changes that the analysis made to the expected Q-matrix were explicable when the results per item were investigated further and cross-checked with an SME.

Table 4.3: Results of the Q-matrix validation as performed by G-DINA, with 0 meaning that this competency was not expected/measured by an item, and 1 meaning that this competency was expected/measured by an item

		Transfer	Text. & Reasoning[1]	Collaboration	Organization	MSC[2]	Ethical awareness	Empathy	Reflection	Time pressure
V-SJT	1			0		1	1	1	1	0
	2			0 → 1		0	0 → 1	1	1	0 → 1
	3			0		1	1	0	1	0
	4			0		1	1	1	1	0
	5			1		0	0	0	1 → 0	0
	6			0		1	1	1	1	0
	7			1		1	1	0	1	0
	8			0		1	1	1	1	0
	9			0		1	0	1	1	1
	10			1		0	0	1 → 0	1	1
	11			0		1	1	0	0	1
WAT	1	1	0 → 1		0	0 → 1	0 → 1		0 → 1	0
	2	1	1		0 → 1	0	0 → 1		0 → 1	0
	3	0	0		0	1	1		1	0
	4	1	0		0	0	0		0	0
	5	0 → 1	1		0 → 1	0	0		0	0
	6	0	1		0	0	0		0	0
	7	0	0		0	1	1		1	1
	8	0	0		1	0	0		0	1

Empty cell = Not Applicable. V-SJT = Video-based Situational Judgement Test; WAT = Written Aptitude Test. [1]Text. & reasoning = Textual comprehension and reasoning. [2]MSC = Medical and Societal Consciousness. Dark backgrounds and two numbers with an arrow between them indicate that the Q-matrix was changed during the Q-matrix validation analysis; the first number is from the Q-matrix based on the blueprint (expected), the second number is the result of the G-DINA analysis and due to the applicants' scores (measured). All other numbers were expected ánd measured.

4.4 Discussion

The aim of this study was to investigate the evidence related to the construct validity of our selection procedure, in order to open the 'black box of selection'. Our specific focus was on content and internal structure, as these shed the most light into this black box. The set-up of the selection procedure proved to be a multi-step and robust process to determine content, which was transparent and replicable, and translated into representative items according to the applicants. Moreover, the G-DINA Q-matrix validation indicated 89% overlap between the expected and actually measured competencies for the items of the V-SJT and WAT. This shows that focusing on the right content by following the competency blueprint was effective in terms of internal structure, and that we are really measuring what we want to measure.

The majority of the evidence presented in the current study is supportive of the selection procedure's construct validity. Related to the content of the selection

procedure, we found that it was possible for a group of committed SMEs to form a selection committee proficient in carefully creating a blueprint of derived competencies needed for medical school and constructing tests capable of distinguishing between applicants based on these competencies. The applicants agreed with the idea that the selection procedure was fairly representative; they indicated that they could accurately portray their abilities in the selection tests and that the selection procedure as a whole contained all aspects needed to identify suitable candidates. All in all, the process of gathering content for the selection procedure appears to be robust, transparent and replicable. Moreover, from previous research we already know that the current procedure is predictive for pre-clinical (Schreurs, Cleutjens, et al., 2018) and clinical (Schreurs, Cleutjens, Cleland, & Oude Egbrink, 2019) performance during medical school.

Related to the internal structure of both tests, there seems to be a huge overlap (89%) between the expected and actually measured competencies in both the V-SJT and the WAT. This indicates that the internal structure of the tests used in the selection procedure mostly reflects the content that was intended to be measured. Although the overall test accuracy was only acceptable for the V-SJT, both tests showed acceptable attribute level accuracies (≥0.84 for the V-SJT and ≥0.67 for the WAT). As stated before, the difference in test accuracy between both tests is likely caused by the fact that the WAT measured a broader range of competencies (i.e. both (inter)personal and cognitively-loaded ones), while the V-SJT focused specifically on the more (inter)personal competencies. Importantly, the inter-rater as well as the intra-rater reliabilities were very high. In both tests, time pressure was found to influence the applicants' performance in the last few items, which was in line with the applicants' comments. As a result, time pressure was included as a construct in the G-DINA analysis, which confirmed its effect.

Taking time pressure into account, we looked at several other effects. Firstly, the guessing parameter (i.e. the probability that a respondent responds correctly to the item although, based on the scores of the other items, he or she has not mastered all the required attributes) was low (<0.5) for most items. The only item with a higher chance of getting it right through guessing appeared to be item 7 in the WAT. The topic presented in this item was relevant but the text was formulated in a complex manner, which may have introduced a high cognitive load. The latter is supported by the highly significant DIF and relatively large effect size of pu-GPA for this particular item. The relatively high guessing parameter for this item may therefore indicate that performance on this item is related to the applicants' pu-GPA rather than their competencies. The slipping parameter (i.e. the probability that a respondent responds incorrectly to the item although, based on the scores of the other items, he or she has mastered all required attributes) was found to be low for most items as well. The only item with a very high slipping parameter was the first item in the V-SJT. This may mean that this item is actually not measuring what was intended to be measured, which is supported by this item's low item-total correlation. This may be due to the 'first-item effect', caused by the fact that the V-SJT is a new kind of test to most applicants, that

they have to make under high pressure with a lot at stake. This suggests that each first item would suffer from this effect. Further examination in later years has to demonstrate whether this explanation is valid. The only other item with a relatively high slipping parameter was the fifth item in the WAT. This item was a specific test of fluid intelligence (i.e. "defined as reasoning ability, and the ability to generate, transform, and manipulate different types of novel information in real time" (Zaval, Li, Johnson, & Weber, 2015)). However, in hindsight, the competency that was intended to be measured (Textual comprehension and reasoning) was too broad for this specific item.

As for the Differential Item Functioning, age and gender did not affect overall performance: some of the items showed some effects of age and gender, but they outweighed themselves. Pu-GPA, however, was significantly and positively related to performance on two of the V-SJT items and four of the WAT items. Interestingly, the V-SJT scores were less affected by pu-GPA, which probably relates to the fact that the V-SJT primarily assessed the more (inter)personal competencies.

The most important analysis applied in answering the question whether effortful creation of the content of a selection procedure, based on a blueprint, leads to an internal structure in line with that blueprint was the G-DINA. The G-DINA results show that the large majority of the expected competences as reflected in the blueprint were actually measured in both the V-SJT (92% overlap) and the WAT (84% overlap). The changes proposed by the G-DINA were critically assessed by the authors, and in hindsight, most changes make sense, while some do not. These results warrant further investigation.

Some novelties in the current study are worth highlighting. First, the application of Cognitive Diagnostic Modeling (CDM). Although García et al already applied CDM to an SJT used for selection in the financial sector in 2014 (Garcia et al., 2014), its application in medical education research is new. Like García et al., we conclude that this emerging analytical method is appropriate for SJT data as well as for selection data in general; it easily copes with multidimensionality, not only between but also within the items. Furthermore, CDM fits the purpose of the current study perfectly; it indicates whether the items were measuring what they were intended to measure, and whether other competencies unintendedly were measured as well. As a consequence, it is possible to investigate the construct validity of selection processes in addition to their predictive and incremental value. Therefore, applying CDM is the main implication of the current study: it is an extremely versatile test theory that is highly applicable to selection procedures. Importantly, it is easily integrated into validity arguments according to modern validity theories (e.g. AERA, APA & NCME, 2014; Downing, 2003). Therefore, these analyses can be applied at other educational institutes as well, to help them understand their selection procedures more thoroughly and to gather information on the validity of their procedures.

The second important novelty is that, to the best of our knowledge, this is the first time Downing's validity framework has been used to assess evidence related to the construct validity of an outcome-based selection procedure. In the current study, we chose to focus on only two aspects of construct validity in Downing's framework (2003): *content* and *internal structure*. Our previous research provides information on two other aspects of the framework. With regard to *relation to other variables*, a positive relation has been shown between being selected and study success throughout the medical (pre-clinical) bachelor (Schreurs, Cleutjens, et al., 2018) and clinical master (Schreurs et al., 2019). Related to the *consequences* of the selection procedure, the cost-effectiveness of the MUMS selection procedure was investigated as compared with a lottery procedure, and it was found that even though selection requires a significant financial investment, the benefits in the medical bachelor already outweigh the costs of the whole procedure (Schreurs, Cleland, et al., 2018). For evidence related to *response processes*, no thorough empirical research has been performed yet. This is an important future direction for research and a limitation for the current study.

Our findings illustrate that research on selection for medical school can focus on more than predictive validity alone. Investigating construct validity with the help of validity frameworks offers a more general evidence base for the application of selection procedures, making them more defensible and fair. Furthermore, applying newer test theories such as CDM provides information on which constructs are indeed measuring what they were intended to measure, and which should be excluded. In the current study, we have shown that the use of CDM can offer new ways to ameliorate selection procedures. It enables a critical reflection on the value of individual tools and items, and opens ways to make these high-stakes procedures more justifiable and fair. On the basis of CDM, the local selection committee has grown more critical towards the competencies intended to be measured per item.

A limitation of this study was the need to dichotomize the responses of the applicants to enable their use in the G-DINA rather than using a polytomous approach, because of the relatively small sample size. Due to the necessary dichotomization of the data, some of the richness of the data was lost for the G-DINA. Nevertheless, the current results show a huge and convincing overlap with the original blueprint, supporting the construct validity of our selection procedure. Furthermore, all other analyses were conducted using the raw data. The current use of G-DINA can be considered as an initial exploration of its potential in the analysis of medical school selection. More studies applying CDM to selection (and other areas of Health Professions Education) are highly welcome, as are comparisons between dichotomized and polytomous CDM analyses. Another limitation of this study is the use of just one cohort from one institution. Further studies in other contexts are necessary to investigate whether the results obtained in the present investigation are generalizable. Also, the current study focused on the second round of the selection procedure alone, and validity evidence should be gathered for the procedure as a whole. Therefore, in future studies, the entire selection procedure should be taken into account. An important gap to fill in the

general selection literature is also the issue of weighting: how should different constructs and/or tools be weighted in order to achieve the most valid selection procedure?

We conclude that a carefully built blueprint is not only useful to obtain a good level of content validity for medical school selection, but it also proved to have an important positive impact on the quality of the results in terms of internal structure. By linking the blueprint to the applicants' results, we established that we are indeed measuring the constructs we intended to measure, therewith shedding light in the 'black box of selection'. We believe this study shows that it is possible to evaluate the construct validity of medical school selection.

4.5 References

AERA (American Educational Research Association), APA (American Psychological Association) & NCME (National Council on Measurement in Education) (2014). *Standards for educational and psychological testing*. Washington, United States of America: American Educational Research Association.

Bandiera, G., Abrahams, C., Ruetalo, M., Hanson, M. D., Nickell, L., & Spadafora, S. (2015). Identifying and Promoting Best Practices in Residency Application and Selection in a Complex Academic Health Network. *Academic Medicine, 90*(12), 1594-1601. doi:10.1097/ACM.0000000000000954

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational Judgment Tests: Constructs Assessed and a Meta-Analysis of Their Criterion-Related Validities. *Personnel Psychology, 63*(1), 83-117. doi:DOI 10.1111/j.1744-6570.2009.01163.x

Cleland, J., Dowell, J., McLachlan, J., Nicholson, S., & Patterson, F. (2012). *Identifying best practice in the selection of medical students (literature review and interview survey)*. Retrieved from https://www.gmc-uk.org/-/media/about/identifyingbestpracticeintheselectionofmedicalstudentspdf51119804

Cleland, J., Dowell, J., Nicholson, S., & Patterson, F. (2014). How can greater consistency in selection between medical schools be encouraged? A project commissioned by the Selecting for Excellence Group (SEEG). Retrieved from http://www.medschools.ac.uk/SiteCollectionDocuments/Selecting-for-Excellence-research-Professor-Jen-Cleland-et-al.pdf. doi:10.1111/medu.12817

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical Education, 49*(6), 560-575. doi:10.1111/medu.12678

Crossingham, G., Gale, T., Roberts, M., Carr, A., Langton, J., & Anderson, I. (2011). Content validity of a clinical problem solving test for use in recruitment to the acute specialties. *Clinical Medicine, 11*(1), 23-25.

Dagnall, N., Denovan, A., Parker, A., Drinkwater, K., & Walsh, R. S. (2018). Confirmatory Factor Analysis of the Inventory of Personality Organization-Reality Testing Subscale. *Frontiers in Psychology, 9*(1), 1116. doi:10.3389/fpsyg.2018.01116

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald Test for Item-Level Comparison of Saturated and Reduced Models in Cognitive Diagnosis. *Journal of Educational Measurement, 50*(4), 355-373. doi:10.1111/jedm.12022

Dore, K. L., Reiter, H. I., Kreuger, S., & Norman, G. R. (2017). CASPer, an online pre-interview screen for personal/professional characteristics: prediction of national licensure scores. *Advances in Health Sciences Education: Theory and Practice, 22*(2), 327-336. doi:10.1007/s10459-016-9739-9

Downing, S. M. (2003). Validity: on meaningful interpretation of assessment data. *Medical Education, 37*(9), 830-837.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General, 128*(3), 309.

Frank, J. R. (2005). The CanMEDS 2005 physician competency framework: Better standards, better physicians, better care. Retrieved from http://www.ub.edu/medicina_unitateducaciomedica/documentos/CanMeds.pdf

Frohlich, M., Kahmann, J., & Kadmon, M. (2017). Development and psychometric examination of a German video-based situational judgment test for social competencies in medical school applicants. *International Journal of Selection and Assessment, 25*(1), 94-110. doi:10.1111/ijsa.12163

Garcia, P. E., Olea, J., & De la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema, 26*(3), 372-377. doi:10.7334/psicothema2013.322

George, A. C., & Robitzsch, A. (2015). Cognitive Diagnosis Models in R: A Didactic. *Quantitative Methods for Psychology, 11*(3), 189-205. doi:10.20982/tqmp.11.3.p189

George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Unlu, A. (2016). The R Package CDM for Cognitive Diagnosis Models. *Journal of Statistical Software, 74*(2), 1-24. doi:10.18637/jss.v074.i02

Gjalt-Jorn Y Peters. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *The European Health Psychologist, 16*(2), 56-69.

Hecker, K., & Norman, G. (2017). Have admissions committees considered all the evidence? *Advances in Health Sciences Education: Theory and Practice, 22*(2), 573-576. doi:10.1007/s10459-016-9750-1

Kane, M. T. (1992). An Argument-Based Approach to Validity. *Psychological Bulletin, 112*(3), 527-535. doi:Doi 10.1037/0033-2909.112.3.527

Kelly, M. E., & O'Flynn, S. (2017). The construct validity of HPAT-Ireland for the selection of medical students: unresolved issues and future research implications. *Advances in Health Sciences Education: Theory and Practice, 22*(2), 267-286. doi:10.1007/s10459-016-9728-z

Kiessling, C., Bauer, J., Gartmeier, M., Iblher, P., Karsten, G., Kiesewetter, J., . . . Fischer, M. R. (2016). Development and validation of a computer-based situational judgement test to assess medical students' communication skills in the field of shared decision making. *Patient Education and Counseling, 99*(11), 1858-1864. doi:10.1016/j.pec.2016.06.006

Knorr, M., & Hissbach, J. (2014). Multiple mini-interviews: same concept, different approaches. *Medical Education, 48*(12), 1157-1175. doi:10.1111/medu.12535

Kreiter, C. D. (2017). A research agenda for establishing the validity of non-academic assessments of medical school applicants. *Advances in Health Sciences Education, 22*(2), 559-563. doi:10.1007/s10459-017-9758-1

Kulasegaram, K. (2017). Use and ornament: expanding validity evidence in admissions. *Advances in Health Sciences Education: Theory and Practice, 22*(2), 553-557. doi:10.1007/s10459-016-9749-7

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: a review of recent research. *Personnel Review, 37*(4), 426-441. doi:10.1108/00483480810877598

Ma, W., & De La Torre, J. (2017). GDINA [software package in R]. https://cran.r-project.org/web/packages/

McManus, I. C., Dewberry, C., Nicholson, S., & Dowell, J. S. (2013). The UKCAT-12 study: educational attainment, aptitude test performance, demographic and socio-economic contextual factors as predictors of first year outcome in a cross-sectional collaborative study of 12 UK medical schools. *BMC Medicine, 11*, 244. doi:10.1186/1741-7015-11-244

Messick, S. (1995). Validity of Psychological-Assessment - Validation of Inferences from Persons Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist, 50*(9), 741-749. doi:10.1037/0003-066x.50.9.741

Mink, R. B., Schwartz, A., Herman, B. E., Turner, D. A., Curran, M. L., Myers, A., . . . Steering Committee of the Subspecialty Pediatrics Investigator Network (2018). Validity of Level of Supervision Scales for Assessing Pediatric Fellows on the Common Pediatric Subspecialty Entrustable Professional Activities. *Academic Medicine, 93*(2), 283-291. doi:10.1097/ACM.0000000000001820

Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance, 29*(4), 331-346. doi:10.1080/08959285.2016.1165227

Patterson, F., Archer, V., Kerrin, M., Carr, V., Faulkes, L., Coan, P., & Good, D. (2010). FY1 job analysis report: Improving selection to the foundation programme. . Retrieved from https://isfporguk.files.wordpress.com/2017/04/appendix-d-fy1-job-analysis.pdf

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical Education, 46*(9), 850-868. doi:10.1111/j.1365-2923.2012.04336.x

Patterson, F., Cleland, J., & Cousans, F. (2017). Selection methods in healthcare professions: where are we now and where next? *Advances in Health Sciences Education: Theory and Practice, 22*(2), 229-242. doi:10.1007/s10459-017-9752-7

Patterson, F., & Ferguson, E. (2012). Testing non-cognitive attributes in selection centres: how to avoid being reliably wrong. *Medical Education, 46*(3), 240-242. doi:10.1111/j.1365-2923.2011.04193.x

Patterson, F., Ferguson, E., & Thomas, S. (2008). Using job analysis to identify core and specific competencies: implications for selection and recruitment. *Medical Education, 42*(12), 1195-1204. doi:10.1111/j.1365-2923.2008.03174.x

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education, 50*(1), 36-60. doi:10.1111/medu.12817

Patterson, F., Roberts, C., Hanson, M. D., Hampe, W., Eva, K., Ponnamperuma, G., . . . Cleland, J. (2018). 2018 Ottawa consensus statement: Selection and recruitment to the healthcare professions. *Medical Teacher, 40*(11), 1-11. doi:10.1080/0142159X.2018.1498589

Patterson, F., Rowett, E., Hale, R., Grant, M., Roberts, C., Cousans, F., & Martin, S. (2016). The predictive validity of a situational judgement test and multiple-mini interview for entry into postgraduate training in Australia. *BMC Medical Education, 16*(1), 87. doi:10.1186/s12909-016-0606-4

Patterson, F., & Zibarras, L. (Eds.). (2018). *Selection and Recruitment in the Healthcare Professions: Research, theory and practice*. Cham, Switzerland: Springer Nature Switzerland AG.

Prideaux, D., Roberts, C., Eva, K., Centeno, A., McCrorie, P., McManus, C., . . . Wilkinson, D. (2011). Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher, 33*(3), 215-223. doi:10.3109/0142159X.2011.551560

Ravand, H., & Robitzsch, A. (2015). Cognitive Diagnostic Modeling Using R. *Practical Assessment, Research & Evaluation, 20*(11), 1-12.

Schreurs, S., Cleland, J., Muijtjens, A. M. M., Oude Egbrink, M. G. A., & Cleutjens, K. (2018). Does selection pay off? A cost-benefit comparison of medical school selection and lottery systems. *Medical Education, 52*(12), 1240-1248. doi:10.1111/medu.13698

Schreurs, S., Cleutjens, K., Cleland, J., & Oude Egbrink, M. G. A. (2019). Outcome-based selection can predict performance in the clinical years of medical school: The proof is in the pudding. *Academic Medicine*, accepted for publication.

Schreurs, S., Cleutjens, K., Muijtjens, A. M. M., Cleland, J., & Oude Egbrink, M. G. A. (2018). Selection into medicine: the predictive validity of an outcome-based procedure. *BMC Medical Education, 18*(1), 214. doi:10.1186/s12909-018-1316-x

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and Reliability of Situational Judgement Test Scores: A New Approach Based on Cognitive Diagnosis Models. *Organizational Research Methods, 19*(3), 506-532. doi:10.1177/1094428116630065

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293. doi:10.1126/science.3287615

Terregino, C. A., McConnell, M., & Reiter, H. I. (2015). The Effect of Differential Weighting of Academics, Experiences, and Competencies Measured by Multiple Mini Interview (MMI) on Race and Ethnicity of Cohorts Accepted to One Medical School. *Academic Medicine, 90*(12), 1651-1657. doi:10.1097/ACM.0000000000000960

Thomson, J. S., Anderson, K., Haesler, E., Barnard, A., & Glasgow, N. (2014). The learner's perspective in GP teaching practices with multi-level learners: a qualitative study. *BMC Medical Education, 14*(1), 55. doi:10.1186/1472-6920-14-55

Tiffin, P. A., Mwandigha, L. M., Paton, L. W., Hesselgreaves, H., McLachlan, J. C., Finn, G. M., & Kasim, A. S. (2016). Predictive validity of the UKCAT for medical school undergraduate performance: a national prospective cohort study. *Bmc Medicine, 14*(1), 140. doi:10.1186/s12916-016-0682-7

Tiller, D., O'Mara, D., Rothnie, I., Dunn, S., Lee, L., & Roberts, C. (2013). Internet-based multiple mini-interviews for candidate selection for graduate entry programmes. *Medical Education, 47*(8), 801-810. doi:10.1111/medu.12224

van der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: from methods to programmes. *Medical Education, 39*(3), 309-317. doi:10.1111/j.1365-2929.2005.02094.x

van Herwaarden, C. L. A., Laan, R. F. J. M., & Leunissen, R. R. M. (2009). *The 2009 Framework for Undergraduate Medical Education in the Netherlands*. In (pp. 90). Retrieved from https://www.nfu.nl/img/pdf/09.4072_Brochure_Raamplan_artsopleiding_-_Framework_for_Undergraduate_2009.pdf

Wilkinson, T. M., & Wilkinson, T. J. (2016). Selection into medical school: from tools to domains. *BMC Medical Education, 16*(1), 258. doi:10.1186/s12909-016-0779-x

Zaval, L., Li, Y., Johnson, E. J., & Weber, E. U. (2015). Complementary Contributions of Fluid and Crystallized Intelligence to Decision Making Across the Life Span. In T. M. Hess, J. Strough, & C. E. Löckenhoff (Eds.), *Aging and Decision Making* (pp. 149-168). San Diego: Academic Press.

CHAPTER 5

Does selection pay off? A cost–benefit comparison of medical school selection and lottery systems

Schreurs S, Cleland J, Muijtjens AMM, oude Egbrink MGA, Cleutjens K. Does selection pay off? A cost-benefit comparison of medical school selection and lottery systems. *Medical Education*. 2018; 52(12):1240-8

Abstract

Context

Resources for medical education are becoming more constrained, whereas accountability in medical education is increasing. In this constrictive environment, medical schools need to consider and justify their selection procedures in terms of costs and benefits. To date, there have been no studies focusing on this aspect of selection.

Objectives

We aimed to examine and compare the costs and benefits of two different approaches to admission into medical school: a tailored, multimethod selection process versus a lottery procedure. Our goal was to assess the relative effectiveness of each approach and to compare these in terms of benefits and costs from the perspective of the medical school.

Methods

The study was conducted at Maastricht University Medical School, at which the selection process and a weighted lottery procedure ran in parallel for 3 years (2011–2013). The costs and benefits of the selection process were compared with those of the lottery procedure over three student cohorts throughout the Bachelor's program. The extra costs of selection represented the monetary investment of the medical school in conducting the selection procedure; the benefits were derived from the increase in income generated by the prevention of dropout and the reductions in extra costs facilitated by decreases in the repetition of blocks and objective structured clinical examinations.

Results

The tailor-made selection procedure costs about €139,000 when extrapolated to a full cohort of students (n = 286). The lottery procedure came with negligible costs for the medical school. However, the average benefits of selection compared with the lottery system added up to almost €207,000.

Conclusions

This study not only shows that conducting a cost–benefit comparison is feasible in the context of selection for medical school, but also that an 'expensive' selection process can be cost beneficial in comparison with an 'inexpensive' lottery system. We encourage other medical schools to examine the cost-effectiveness of their own selection processes in relation to student outcomes in order to extend knowledge on this important topic.

5.1 Introduction

Medical schools throughout the world are confronted with high applicant numbers for a restricted amount of places (1). They have to select the best candidates from a pool of well-qualified applicants, many of whom also possess the personal qualities considered desirable in a medical student and doctor (2). While there are typically three general domains selected for at the time of admissions (academic achievement, aptitude for medical school/medicine, and non-academic attributes; 1, 2), how these are measured varies widely.

In the current resource-constrained times, with a decrease in public funding and a simultaneously increasing demand for accountability (3), medical schools are under increasing pressure to also justify their selection processes in terms of costs and benefits (4-6). They must ensure efficient and effective use of finances in education organization and delivery (6-9). Up to now, research on selection has mostly focused on the predictive validity of the various selection tools (1, 10).

In reference to admissions procedures, the most inexpensive selection process in terms of costs is probably either using a single selection tool of prior attainment (e.g. secondary school examination results) or a weighted lottery system in which admission chances increase in parallel with increases in applicants' pre-university GPA (for more information on the weighted lottery procedure previously used in the Netherlands, the reader is referred to reference 6, 10, 11 or 12). These are inexpensive approaches because the required data are provided to medical schools by external bodies, at minimum cost to the medical school. At the other end of the spectrum, selection procedures consisting of multiple time-intensive and costly tools are highly expensive. For example, conducting multiple mini-interviews (MMIs) with large numbers of applicants is an expensive endeavor: many assessors and/or actors must be present for long periods of time, rooms must be booked and allocated, and staff and actors provided with refreshments. Furthermore, MMIs require much preparatory work to ensure high levels of reliability and validity across many stations, and the logistics and administration of organizing MMIs can be challenging (4, 5, 7, 8).

So far, cost-benefit analyses of multi-tool selection procedures as a whole are scarce; to the best of our knowledge, this is a gap in the literature. At the same time, this is the focus of much discussion in terms of policy and practice. In the Netherlands, for example, a debate on admissions processes is ongoing, focusing on the question whether the previous, relatively inexpensive weighted lottery system should be reintroduced to replace the current, more expensive selection procedures (13). We add to this debate by focusing on an economic evaluation of medical school selection: what is spent in relation to what value is returned (9, 14).

The aim of the current study is to determine whether the benefits of applying a tailor-made selection process outweigh the costs this process entails when compared to a lottery procedure, from the perspective of the medical school. To safeguard the

quality of the study, the CHEERS statement (Consolidated Health Economic Evaluation Reporting Standards; 15, 16) was followed. We define costs as the extra costs for the medical school in applying the selection procedure over the lottery procedure, while benefit is defined as a combination of preventing loss of income due to student drop-out and preventing additional future costs due to poor performance (see 5.2.5). The ultimate goal of this study is to contribute information for decision-making on whether to continue investing time and money in developing and adapting selection procedures, or to (re)introduce the inexpensive lottery procedure.

5.2 Methods

5.2.1 Setting and population

We were able to examine our research question in a naturalistic setting, specifically that of Maastricht University Medical School (MUMS), where a tailor-made selection process ran in parallel with a lottery system for three years (17). In this context, the costs and benefits of the traditional admissions procedure (i.e. the lottery procedure) could be compared to those of a tailored selection procedure (selection is now common practice in the Netherlands; 10). Up to and including 2010, all students were admitted to MUMS through a national weighted lottery. Hereafter, the admission process gradually changed from the lottery procedure to selection of all students from 2014 onwards. Thus, for three years – 2011 through 2013 – an outcome-based selection procedure ran in parallel to the lottery. In the first year, 111 of 286 of students were admitted through this selection procedure. This amount increased to 141 in 2012 and 149 in 2013. The selection ratios in these years were 6.6, 5.9 and 5.3 applicants per available study place, respectively. The remaining study places were filled through the national weighted lottery procedure. As a result, the cohorts of 2011, 2012 and 2013 consisted of a combination of selected students and students who entered via the lottery. The latter group was composed of students rejected in the selection process who then successfully entered through the lottery procedure, and students who participated in the lottery procedure only. In this study, the costs and benefits related to the selected students (Selection-Positive, SP, n=401 in total) on the one hand and the students that entered through lottery only (LO, n=185) on the other, were determined for all three cohorts. A comparison of both groups (SP and LO) on pu-GPA revealed no difference.

In the Netherlands, medical school is divided into two three-year phases, the bachelor and the master. In the bachelor, education is mostly university-based and pre-clinical. The master phase is clinical and primarily workplace-based. In this study, we focused on the bachelor phase at MUMS, in which a problem-based curriculum was offered.

5.2.2 Perspective and time horizon

The current study was conducted from the perspective of the medical school, all costs and benefits were therefore determined within the context of the medical school. The time span in the current study is the bachelor in medicine (three years) for three

cohorts of students (starting in 2011, 2012 and 2013). All costs and benefits were analyzed in the first quarter of 2018.

5.2.3 Selection procedure

The selection procedure applied at MUMS consists of two rounds, both assessing the CanMEDS competences set forward by Frank in 2005 (18). In the first round, applicants fill out an online portfolio. This contains information on their previous academic attainment (e.g. pre-university GPA), distinguishing abilities gained in extracurricular activities (relating to communicator or collaborator, for example), reasoning behind choosing MUMS, and their knowledge on and self-perceived suitability for problem-based learning. Applicants were ranked based on their scores on the portfolio, and the highest ranking applicants (twice the amount of available study places) were invited to the second round, a selection day at MUMS. During this selection day, two different assessment tools were used: a Video-based Situational Judgment Test (V-SJT, consisting of eight to ten short videoclips with corresponding questions) and a written aptitude test. Both tools focused on talent for the whole set of competences, mostly pertaining to a real-life medical student or doctor setting. In the end, the applicants were ranked using their mean Z-scores on all assignments, and the highest ranking students were offered a place in the curriculum.

Rejected applicants and those who chose not to participate in the selection procedure at all could participate in the national weighted lottery procedure. The weights were based on the students' pre-university GPA.

5.2.4 Costs of admissions procedures

To delineate the costs of both admission procedures from the medical school's perspective, two types of costs were distinguished: fixed costs and variable costs (see Figure 5.1). Fixed costs are the costs that were made to enforce the admissions process independent of the number of applicants, for example staff hours required to develop the assignments. Variable costs are the costs directly related to the number of applicants and to the amount of students to be selected, for example staff hours needed to evaluate assignments and for surveillance throughout the selection day. Both types of costs were expressed as the number of hours allocated to both scientific and support staff (converted to monetary costs), while remaining costs (e.g. external staff for video-editing and test layout) were included as well. Variable costs were extrapolated to costs for a full cohort from the data of the cohorts 2011 through 2013.

5.2.5 Economic benefits due to selection

The expected monetary benefits of selection can roughly be divided into two types: (1) an increased net income because less students left the program without graduating and (2) a decrease of costs for remediation and resits because less students failed exams (see Figure 5.1; 13, 19). With respect to the first type of benefit, it is important to note that education of medical students in the Netherlands is funded by the government; payments are received for each registered year, with a maximum of three years for the bachelor, as well as for graduation. This results in a yearly payment

per student to the medical school. As soon as a student drops out, MUMS no longer receives these yearly payments for this student; at the same time, money is saved because the student no longer participates in educational activities.

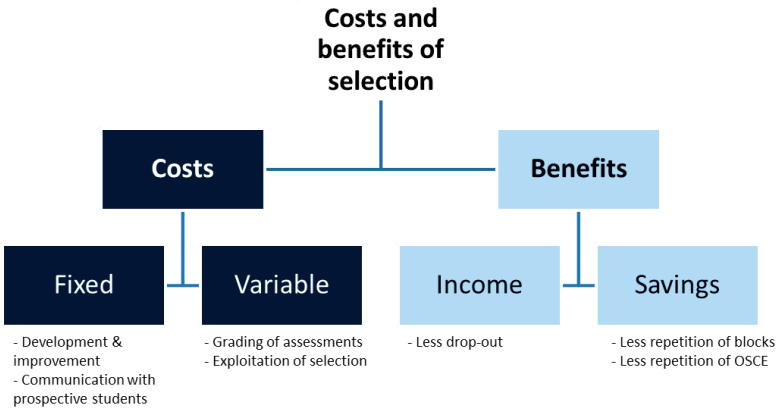


Figure 5.1: The different kinds of costs and benefits of a selection procedure and their sources.

Secondly, poor performance of students results in study delay, with extra costs for remediation and resits of failed blocks and assessments. The problem-based bachelor curriculum at MUMS consists of two 4-week and four 8-week blocks in year 1 and 2, and of four 10-week blocks in year 3. In these thematic blocks, educational activities such as skills trainings, practicals, and simulated and/or authentic patient contacts are organized alongside tutorial group meetings. Assessment of students is either theoretical (block exams and progress tests) or related to attitude and all kinds of skills (e.g., objective structured clinical examinations; OSCEs, one per year).

When a student failed a block (of 4, 8, or 10 weeks) s/he had to either follow the whole block again, which came with considerable costs for the medical school, or redo the exam only, for which the expenses were negligible. The only single assessment causing significant additional costs per student was the OSCE (20). Repetitions of other parts of the assessment program caused negligible expenditures (e.g. doing the progress test again). Other costs related to poor performance, such as the costs of extra guidance to support individual, poorly performing students in completing their studies, were not available due to ethical reasons, and therefore not included in the current analysis.

5.2.6 Cost-Benefit Analysis

In the current article, cost-benefit analysis is defined as investigating whether selection is 'good value' when its costs are compared to its monetary benefits (3, 21). Firstly, the costs of admissions procedures were determined, by sorting out both the fixed and variable costs of lottery and selection, and extrapolating the variable costs to an entire cohort of 286 students. Hereafter, the benefits of selection over lottery were

assessed by determining the average frequency of drop-out and repetition per year in each of the two student groups (SP and LO). The benefit attained was then calculated by extrapolating these frequencies to an entire cohort of either SP- or LO-students.

5.3 Results

5.3.1 Costs of the admissions procedures

The costs of the national weighted lottery for the medical school itself were close to zero. An administration office within the university determined whether applicants fulfilled all legal requirements for all admission procedures, including the lottery. The selection procedure, on the other hand, came with considerable fixed and variable costs. The fixed costs comprised staff hours for setting up the selection procedure, producing content for the selection procedure, and providing information on the selection procedure. The variable costs were costs related to, for example, grading assignments as well as costs for provisions such as equipment, materials, and facilities, all of which increase with the amount of students applying. Both scientific and support staff was involved in the procedure; average salary costs were €90,000 per FTE (fulltime-equivalent, i.e. 1650 hours per year) for scientific staff and €52,000 per FTE for support staff. These costs include on-costs such as employed pension contributions and payroll tax.

On average, 134 students entered MUMS through selection per year in the three cohorts under study. As shown in Table 5.1, the average fixed costs per year were composed of 0.51 FTE scientific staff (€45,900) and 0.5 FTE support staff (€26,000), with €5,800 remaining fixed costs (e.g. video-editing). These fixed costs were unrelated to the number of applicants involved. The average variable quantity amounted to 0.13 FTE for scientific staff and 0.18 FTE for support staff per year. Extrapolation of these variable costs to an entire cohort of 286 students results in sums of 0.28 FTE for scientific staff (€25,200) and 0.38 FTE for support staff (€19,760). The average costs of remaining provisions per year were €7,600, which extrapolates to €16,220 for a full cohort of students. Taken together, this means that -whereas the average costs of selecting 134 of students at MUMS were approximately €106,360 per year- applying the current selection procedure for an entire cohort of 286 students would have resulted in average costs of €138,880.

Table 5.1: Average yearly costs of the selection procedure at Maastricht University Medical School in the years 2011 through 2013, extrapolated to a full cohort

	Scientific staff	Support staff	Remaining costs	Total
Fixed				
FTE*	0.51	0.50		1.01
Costs	€45,900	€26,000	€5,800	€77,700
Variable				
Extrap. FTEs full cohort	0.28	0.38		0.65
Extrap. costs full cohort	€25,200	€19,760	€16,220	€61,180
Total costs full cohort	€71,100	€45,760	€22,020	€138,880

*FTE = fulltime-equivalent (1650 hours per year; €90,000 for scientific staff, €52,000 for support staff).
Extrap. = Extrapolated

5.3.2 Economic benefits because of selection

The medical school received an average payment from the hosting faculty of about €10,000 per registered student per year; this payment represents a compensation for the educational activities provided by the medical school staff. The total costs of education of a medical student exceeds this amount by far; the additional overhead and more generic educational costs were paid to and covered by the faculty and the university (e.g. infrastructure, IT, library, service center and management costs, and clinical workplace-based costs). The payment of a fixed amount of money per year means that if a student dropped out in year 1, the payments for the second and third year were missed (i.e. €20,000). For drop-out during year 2, MUMS missed out on €10,000. If students dropped out in year 3 or later, this no longer affected MUMS' budget for educational activities during the bachelor. However, one should also take into account that when a student drops out, the medical school no longer has to provide this student's education. The majority of the costs for education do not change with a slight decrease of the amount of students (e.g. course development, lectures). Nevertheless, a small portion of the costs for education will decline with a slight reduction of the amount of students; this aggregates to an estimated €2,260 per student per year (four 8- & two 4-week blocks and an OSCE in year two, and four 10-week blocks and an OSCE in year three; for specification of these costs: see later on in this section). Therefore, preventing drop-out in year 1 or 2 increased the medical school's net income by about €15,480 and €7,740 per student, respectively. As shown in Table 5.2, a higher percentage of lottery-admitted than selected students dropped out in year 1; when extrapolated to full cohorts, 20 lottery-admitted students would have dropped out versus 7 selected students. In year 2, it was estimated that 2 selected students would have dropped out, versus 1 lottery-admitted student. In monetary terms, a lottery-admitted cohort would result in missing out on €317,340, while a selection-admitted cohort would result in €123,840 loss of income. Therefore, the drop-out reduction caused by selection would increase the average income for a cohort over the entire bachelor by €193,500.

When students failed a block exam and had to redo the entire block, the amount of staff hours needed for teaching increased. This led to additional costs of €220, €445 and €555 per student for a 4-, 8-, and 10-week block, respectively. The data in Table 5.2 show that lottery-admitted students had to redo all three kinds of blocks more often that selected students. Hence, when extrapolated to full cohorts, a lottery-admitted cohort would have to repeat these blocks more often than a selection-admitted cohort. As a result, the average total costs of repetition of blocks over the entire bachelor for one cohort would be €11,645 lower in a full cohort of selected students.

The OSCE is an individual test in which all costs, i.e. assessors (scientific staff), simulated patients, and provisions, increase with each student. The mean total cost of an OSCE was €40 per student. As shown in Table 5.2, an entire lottery-admitted cohort would have to complete 122 OSCE resits throughout the bachelor (€4,880), while a full

cohort of selected students would need 82 resits (€3,280). This results in a difference in costs of €1,600.

Combining all abovementioned data shows that an average full cohort of selected students would be less expensive in terms of drop-out and repeating blocks and OSCEs than an average full cohort of lottery-admitted students (see Table 5.2). This benefit adds up to a total of €206,745. Because the yearly costs of the selection procedure applied for these cohorts were €138,880, the applied selection procedure appears to be cost-beneficial.

Table 5.2: Average yearly benefits of the selection procedure versus the national weighted lottery procedure at Maastricht University Medical School

	SP	LO	SP	LO	Difference (LO - SP)	Gains[2]	Total gains of selection
	Average % per cohort		Extrapolated number of students per full cohort[1]			Gains per student not dropping out	
Drop-out year 1	2.5	7.0	7	20	13	€ 15,480	€ 201,240
Drop-out year 2	0.7	0.5	2	1	-1	€ 7,740	€ -7,740
	Average % per cohort		Extrapolated number of repetitions per full cohort			Gains per repetition avoided	
Repetitions 4-wk blocks	2.4	3.5	27	40	13	€ 220	€ 2,860
Repetitions 8-wk blocks	3.2	3.9	73	89	16	€ 445	€ 7,120
Repetitions 10-wk blocks	0.9	1.1	10	13	3	€ 555	€ 1,665
Resits OSCE	9.6	14.2	82	122	40	€ 40	€ 1,600
Total							**€ 206,745**

[1] A full cohort consists of 286 students; data from cohorts 2011-2013 (401 selected students and 185 students admitted through lottery only) were extrapolated to a full cohort of selected (SP) or lottery (LO) admitted students during the entire bachelor. Without repetitions, a full cohort of students (n= 286) represents 4 x 286 = 1144 4-week blocks, 8 x 286 = 2288 8-week blocks, 4 x 286 = 1144 10-week blocks and 3 x 286 = 858 OSCE-participations during the bachelor of medicine.
[2] Gains can be (1) an increase in income because the amount of drop-out is decreased or (2) savings for the medical school because blocks or assessments (OSCEs) are repeated less often.

5.4 Discussion

We compared the costs and benefits of a tailor-made selection procedure at one medical school with those of a weighted lottery system. The cost-benefit analysis indicates that -from the perspective of the medical school- the benefits of this selection procedure outweigh its costs, compared to the benefits and costs of the lottery system. This result is due mostly to lower rates of drop-out and/or failure of selected students versus lottery students.

The selection procedure under study was more expensive than the lottery procedure. In the three years under investigation, the average costs of selecting almost half of the students at MUMS was approximately €106,000 per year. Extrapolating this to a

situation where all students in a cohort are admitted through selection (i.e. without a lottery entry stream), the total costs per year would be approximately €139,000.

While calculation of costs is relatively simple, estimating the monetary benefits of selection over lottery is more complex. We divided the benefits of selection into two types of gain: (1) a decrease in the number of students dropping out, therewith increasing the net income for the medical school, and (2) a decrease in the amount of repeated courses and exams, therewith decreasing extra costs for the medical school (which are not reimbursed by government funding). To do so, the average number of selected and lottery-admitted students dropping out and repeating blocks and OSCEs in the years under study was calculated. These average numbers were then extrapolated to a complete selection-admitted cohort and referenced against the outcomes of an entire cohort of lottery-admitted students. This comparison indicated that shifting completely to selection would amount to a total benefit of almost €207,000. It is important to note that this is a conservative estimate of the benefits as we did not calculate the costs of extra support for underperforming students or additional administrative costs; both are difficult to estimate and data on underperforming students are not available for ethical reasons.

When combining the costs (~€139,000) and benefits (~€207,000) of the selection procedure under study, we can conclude that the selection procedure is cost-effective to the tune of about €68,000 per cohort compared to the lottery. This implies that even a relatively complex and time-consuming selection procedure can be cost-beneficial if it has predictive value in terms of performance throughout the bachelor phase of medical school (17).

It is important, however, to keep in mind that an extrapolation to a full cohort of 286 students was conducted based on data from a total number of 401 selected and 185 lottery-admitted students in three cohorts. Our assumption was that the performance of hypothetical students who would be added to these actual groups to get to a full cohort of students would be equal to the ones in the respective actual groups. We do not know if this would be the case. For the selected students, there is support that this sample is representative for an entire cohort, as drop-out remained as low as in our extrapolation in the years MUMS proceeded to selection of the entire cohort (cohorts 2014 and later). To support our extrapolation further, we determined that there was no significant difference in any of the current outcomes between students who performed in the top 10 percent versus those who performed in the bottom 10 percent of the selection procedure. This suggests that an entire cohort of selected students would likely perform as well as the selected students in the current study. Whether the lottery-admitted students in the current study are representative for a full cohort of lottery-admitted students is questionable. However, we could not carry out a retrospective comparison because the bachelor curriculum was completely revised in 2011. We looked to see what would happen if the full lottery-admitted cohort would consist of the 185 actual lottery-admitted students used in the current study, supplemented with a hypothetical 50/50 mixture of students who would have

been admitted through either selection or lottery. In this rather conservative hypothesis, the selection procedure would still be cost-effective (the benefits would still outweigh the costs by more than €42,000). In reality, the true gain is likely to be much closer to the extrapolation based on a full extrapolation of both groups, since students are recruited from a large pool.

Though the current study provides important insights, many questions relating to cost-effectiveness remain unanswered. These include the effect on cost-effectiveness of: (1) other perspectives, (2) different combinations of selections tools, and (3) different weighting of selection tools. We focused on the medical school perspective. However, the perspectives of other stakeholders like students, patients and society are at least equally important (19), and merit study. For example, from a societal perspective, the biggest gain from selection would be increasing the quality of future doctors, which also increases cost-effectiveness. Secondly, different universities use different ways of selecting their students (1). This makes it nearly impossible to conduct a study that would be easily generalizable to other contexts. However, the broad combination of pu-GPA with an aptitude test and a tool focused on (inter)personal skills is relatively common (4). It is also important to look with more granularity, to examine which specific features of a selection procedure make it cost-effective (13): what do the different tools cost, and what is their contribution to the predictive value of the procedure as a whole? This may help to create a selection procedure that is as 'lean' as it possibly can be. Lastly, cost-effectiveness analyses can be used as an outcome to optimize the weighting of different tools and/or assessed features within the selection procedure. Weighting of different tools and/or features has been proposed as a new field of research before (2, 8, 10, 22), but may be well-combined with the more granular understanding of costs and effectiveness within selection. Finding an optimal weighting of the tools and/or content within the selection procedure may result in better prediction and, in turn, a more cost-efficient selection procedure.

Like all research, the current study has its limitations. We conducted the study in a single medical school. As discussed above, medical school selection processes vary and so it is difficult to compare across contexts. Second, we focused on the bachelor phase of medical school only and limited the analysis to outcomes that were predicted to have a high monetary impact (for an in-depth comparison on specific educational outcomes, the reader is referred to Schreurs et al.; 17). We may have underestimated the cost benefits of the selection procedure given that drop-out anywhere in the bachelor also affects the incomes in the master phase. Following up our three cohorts as they progress through the later part of their medical degree is underway. Lastly, not all points on the CHEERS checklist were relevant to the current study e.g., health outcomes, discount rates (15, 16). A strength in the current study is the use of three year groups, which controls for possible cohort effects.

In conclusion, we responded to Patterson et al.'s conclusion (1) that very little research has explored the relative cost-effectiveness of medical selection methods by carrying out a cost-benefit comparison of a tailor-made medical school selection

procedure with a lottery system. The knowledge from this kind of cost-benefit analyses can help relevant stakeholders determine the optimal use of resources when planning selection, and can help inform decision making about resource allocation. Furthermore, in contexts where there is debate as to whether or not anything more than GPA or a lottery are needed at all in medical selection, our study provides important intelligence.

Acknowledgments: The authors would like to thank Marielle Heckmann and Raymond Bastin for their help in relation to the financial data.

5.5 References

1. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ. 2016;50(1):36-60.
2. Hecker K, Norman G. Have admissions committees considered all the evidence? Adv Health Sci Educ. 2017;22(2):573-6.
3. Maloney S, Cook D, Foo J, Rivers G, Golub R, Tolsgaard M, et al. AMEE Guide no. *: An Educational Decision-Makers Guide to Evaluating and Applying Studies of Educational Costs. Med Teach. In press.
4. Cleland J, Dowell J, McLachlan J, Nicholson S, Patterson F. Identifying best practice in the selection of medical students (literature review and interview survey). http://www.gmc-uk.org/Identifying_best_practice_in_the_selection_of_medical_students.pdf_51119804.pdf; 2012.
5. Hissbach JC, Sehner S, Harendza S, Hampe W. Cutting costs of multiple mini-interviews - changes in reliability and efficiency of the Hamburg medical school admission test between two applications. BMC Med Educ. 2014;14(1):54-63.
6. ten Cate TJ, Hendrix HL, de Fockert Koefoed KJJ, Rietveld WJ. Studieresultaten van toegelatenen binnen en buiten de loting. Tijdschrift voor Medisch Onderwijs. 2002;21(6):253-60.
7. Dore KL, Kreuger S, Ladhani M, Rolfson D, Kurtz D, Kulasegaram K, et al. The reliability and acceptability of the Multiple Mini-Interview as a selection instrument for postgraduate admissions. Acad Med. 2010;85(10 Suppl):S60-3.
8. Patterson F, Cleland J, Cousans F. Selection methods in healthcare professions: Where are we now and where next? Adv Health Sci Educ. 2017;22(2):229-42.
9. Maloney S. When I say... cost and value. Med Educ. 2017;51(3):246-7.
10. Stegers-Jager KM. Lessons learned from 15 years of non-grades-based selection for medical school. Med Educ. 2018;52(1):86-95.
11. Schripsema NR, van Trigt AM, Borleffs JCC, Cohen-Schotanus J. Selection and study performance: Comparing three admission processes within one medical school. Med Educ. 2014;48(12):1201-10.
12. Schripsema NR, van Trigt AM, Lucieer SM, Wouters A, Croiset G, Themmen APN, et al. Participation and selection effects of a voluntary selection process. Adv Health Sci Educ. 2017;22(2):463-76.
13. Foo J, Ilic D, Rivers G, Evans DJR, Walsh K, Haines TP, et al. Using cost-analyses to inform health professions education - The economic cost of pre-clinical failure. Med Teach [Internet]. 2017 Dec 7:[1-10 pp.]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29216780.
14. Nestel D, Brazil V, Hay M. You can't put a value on that… Or can you? Economic evaluation in simulation-based medical education. Med Educ. 2018;52(2):139-41.
15. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. BMJ. 2013;346:f1049.
16. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)--explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. Value Health. 2013;16(2):231-50.
17. Schreurs S, Cleutjens KB, Muijtjens AMM, Cleland J, Oude Egbrink MGA. Selection into medicine: the predictive validity of an outcome-based procedure. BMC Med Educ. 2018;18(1):214.
18. Frank JR. The CanMEDS 2005 physician competency framework: Better standards, better physicians, better care. 2005. Available from: http://www.ub.edu/medicina_unitateducaciomedica/documentos/CanMeds.pdf.
19. Foo J, Rivers G, Ilic D, Evans DJR, Walsh K, Haines T, et al. The economic cost of failure in clinical education: a multi-perspective analysis. Med Educ. 2017;51(7):740-54.
20. Brown C, Ross S, Cleland J, Walsh K. Money makes the (medical assessment) world go round: The cost of components of a summative final year Objective Structured Clinical Examination (OSCE). Med Teach. 2015;37(7):653-9.
21. Maloney S, Haines T. Issues of cost-benefit and cost-effectiveness for simulation in health professions education. Advances in Simulation. 2016;1(1):13-8.
22. Kreiter CD. A research agenda for establishing the validity of non-academic assessments of medical school applicants. Adv Health Sci Educ. 2016;21(5):1081-5.

CHAPTER 6

Increasing value in research: Cost evaluations in health professions education

Schreurs S, Cleutjens K, & oude Egbrink MGA. Increasing value in research: cost evaluations in health professions education. *Medical Education*. 2019; 53(12), 1171-1173.

6.1 Commentary

Nowadays, higher education institutes have to deal with a combination of limited financial resources and increasing demands for accountability. As a result, considerations of costs and value have become paramount in stakeholders' choices. This also holds for health professions education (HPE): imagine a department proposing to buy a high-quality spinning bike because they find it important and interesting to use this in practical sessions on energy expenditure during exercise for medical students. Ideally, the medical school's management should make a well-informed decision between spending the required amount of money to buy this device, ór saving it for other educational activities. To enable such a decision, they should have information on whether this financial investment is balanced by increased student learning. However, this information is often not available: studies on costs are extremely scarce in HPE, and studies including value are even more rare. Furthermore, the systematic review of Foo and colleagues in this issue of Medical Education (1) concluded that the quality of most of the available research is substandard and has not increased since the beginning of this century. They found that shortcomings in the cost-evaluation literature are mostly related to the methodology and reporting specific to these analyses. This is, at least in part, caused by a lack of expertise in performing economic evaluations within the field of HPE. In addition, Foo and colleagues show that very few studies reporting cost data in HPE represent full economic evaluations of educational activities, in which costs and value of two or more approaches are evaluated and contrasted (1). These findings mirror the practice of higher education institutes: decisions on which educational activities to invest in are often not fueled by full empirical economic evaluations.

One educational area in which cost evaluations are particularly interesting is that of selection for medical school. Over the past few decades, many different selection tools have been developed, employed and combined into different selection procedures at various medical schools (2). It is as if every school reinvented the wheel, because 'track and weather' conditions are slightly different everywhere. Medical schools are under increasing pressure to justify their unique and often expensive selection procedures in terms of costs and value. However, research on these costs and especially the value of selection procedures, or even the tools within these procedures, is sorely lacking (3). Therefore, not only is each medical school using a different set of wheels, but importantly, the price tag on these wheels is blank. Furthermore, because research in the field of selection has focused almost solely on the predictive and incremental value of different tools, there is barely any information on the construct validity of selection procedures, as defined in the modern validity theories (4, 5). Research has focused on supporting the 'relation to other variables' (e.g. predictive validity) of selection tools, severely neglecting the other sources of evidence for validity (i.e. content, response processes, internal structure and consequences; 6, 7). Thus, selection wagons are running on different wheels, of which we do not know what they cost nor how crooked they are. This was already observed in 1998 by Tekian: "To begin to understand the nuts and bolts of the admission machine, we opened the gear box, but only to find a jumble of parts that prevents the

machine from running smoothly" (8). In an attempt to get the wagons running smoothly, research in the field of selection for medical school should move towards a more general understanding of validity. Importantly, the 'consequences' pillar of validity should be taken into account explicitly, and must include costs as well as value related to selection tools and procedures, preferably also in comparison with other tools and/or procedures as in cost-benefit evaluations (1).

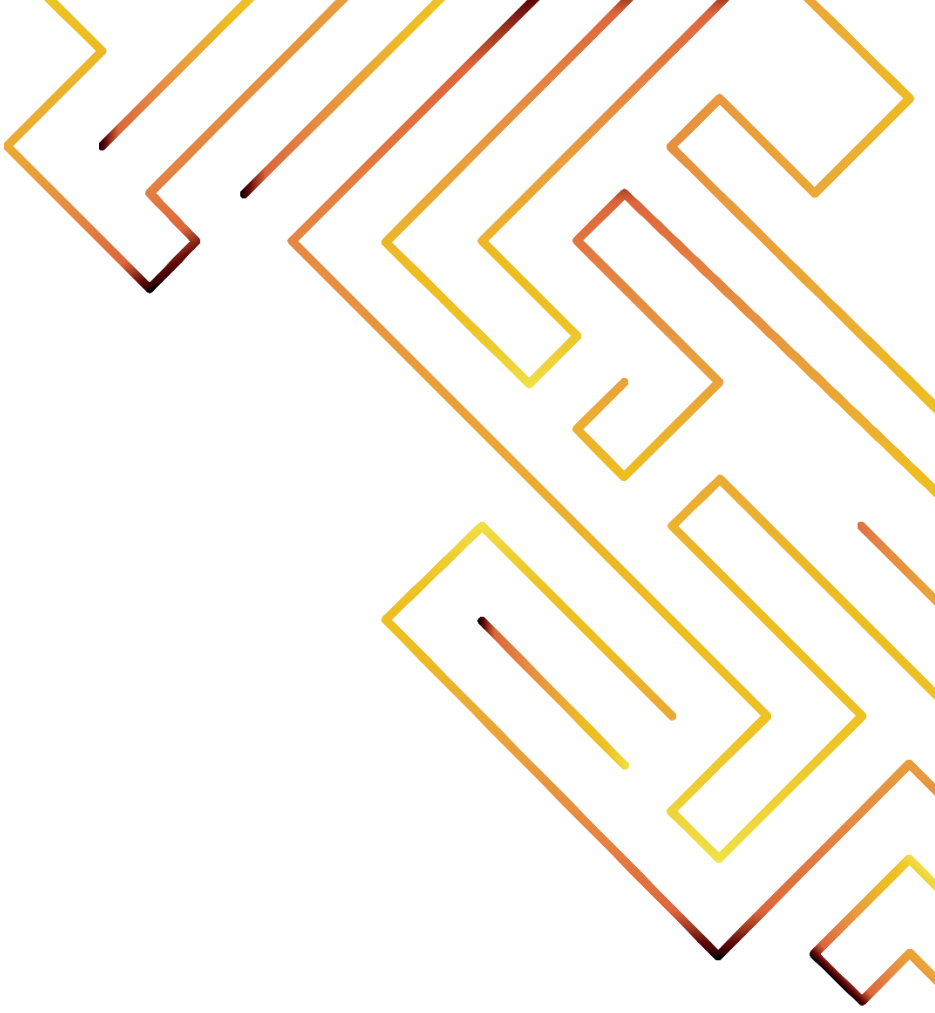
Some rare examples of cost evaluations in the field of selection do exist, for example on how many stations are needed in a Multiple Mini Interview to achieve a certain reliability and acceptability (9). In an effort to start the cost-and-value discussion on selection procedures as a whole, we have recently investigated the costs of the selection procedure at Maastricht University Medical School, and contrasted this with an inexpensive, weighted lottery procedure. We also looked into the value these procedures returned, and concluded that although our tailor-made selection procedure was much more expensive to conduct than the lottery procedure, it already paid itself back in terms of value during the three pre-clinical years of medical school alone (10). In our study, we used the CHEERS statement (11) to safeguard the quality of the study. Nevertheless, in line with Foo et al.'s conclusions (1), we experienced challenges like using the right nomenclature, choosing the correct methodology and putting a price on the value-side of selection. The recently published AMEE guide on how to read studies on educational costs (12) and some exemplary studies (e.g. 13) are highly welcomed aids to support the process of setting up, conducting and reporting future economic evaluations in HPE. Moreover, Foo et al.'s suggestion to involve experts in health or educational economics at the earliest possible stage is essential to improve the quality of cost-evaluation research (1).

Importantly, high quality cost-evaluation research is not only relevant at the level of individual educational institutes, but also on a national level. As an example, we refer to the situation in the Netherlands: a few years ago, medical schools were obliged to switch from a national weighted lottery to decentralized selection procedures which differ from school to school. Because several of these procedures are insufficiently predictive to warrant the monetary investments (see e.g. 14), the cost-effectiveness of selection in general is questioned and reintroduction of a national lottery system is asked for. This dispute has even sparked discussion at the government level. In our opinion, research into the general construct validity of the selection procedures, including high-quality full economic evaluations, are crucial to support an evidence-based governmental decision on this issue.

All in all, there is a lack of high-quality economic evaluations in HPE in general and selection in particular. This puts the stakeholders in a precarious position in which they have to remain accountable regarding their limited financial resources while choosing between educational activities without price tags. Therefore, educational researchers should join efforts to conduct more economic evaluations to provide relevant stakeholders with economically valid arguments to make well-informed decisions that benefit the quality of education.

6.2 References

1. Foo J, Cook DA, Walsh K, Golub R, Abdalla ME, Ilic D, et al. Cost evaluations in health professions education: a systematic review of methods and reporting quality. Med Educ. 2019.
2. Cleland J, Dowell J, McLachlan J, Nicholson S, Patterson F. Identifying best practice in the selection of medical students (literature review and interview survey). https://www.gmc-uk.org/-/media/about/identifyingbestpracticeintheselectionofmedicalstudentspdf51119804.pdf; 2012.
3. Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical education*, *50*(1), 36-60.
4. Royal KD. Four tenets of modern validity theory for medical education assessment and evaluation. Adv Med Educ Pract. 2017;8:567-70.
5. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, United States of America: American Educational Research Association; 2014.
6. Kulasegaram K. Use and ornament: expanding validity evidence in admissions. Adv Health Sci Educ Theory Pract. 2017;22(2):553-7.
7. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-7.
8. Tekian A. Minority students, affirmative action, and the admission process: a survey of 15 medical schools. Acad Med. 1998;73(9):986-92.
9. Dore KL, Kreuger S, Ladhani M, Rolfson D, Kurtz D, Kulasegaram K, et al. The reliability and acceptability of the Multiple Mini-Interview as a selection instrument for postgraduate admissions. Acad Med. 2010;85(10 Suppl):S60-3.
10. Schreurs S, Cleland J, Muijtjens AMM, Oude Egbrink MGA, Cleutjens K. Does selection pay off? A cost-benefit comparison of medical school selection and lottery systems. Med Educ. 2018;52(12):1240-8.
11. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. BMJ. 2013;346:f1049.
12. Maloney S, Cook DA, Golub R, Foo J, Cleland J, Rivers G, et al. AMEE guide no. 123 - How to read studies of educational costs. Med Teach. 2019;41(5):497-504.
13. Foo J, Ilic D, Rivers G, Evans DJR, Walsh K, Haines TP, et al. Using cost-analyses to inform health professions education - The economic cost of pre-clinical failure. Med Teach. 2018;40(12):1221-30
14. Wouters A. Effects of medical school selection on student motivation: a PhD thesis report. Perspect Med Educ. 2018;7(1):54-7.



CHAPTER 7

General discussion

In this chapter, the main findings of the empirical chapters (two through five) and the commentary in chapter six are summarized and the results are reflected upon using the validity framework set forward in the General Introduction (Chapter one). Moreover, implications and suggestions for the practice of and research on selection are debated.

7.1 Context and aims

The research conducted for this thesis was conducted in the medical curriculum at Maastricht University, where a multi-tool, outcome-based selection procedure was put in place in 2011. The context in the Netherlands was especially interesting for selection research since 'decentralized' (i.e. university-specific) selection procedures and national Grade Point Average (GPA) weighted lottery ran in parallel for several years. For the medical curriculum at Maastricht University, this was the case for three cohorts: 2011, 2012 and 2013. This means that in the years 2011, 2012 and 2013, students could either be admitted through the decentral selection procedure or through the national GPA-weighted lottery. In total, there were three routes of entry, as shown in Figure 7.1: being selected in the selection procedure (the Selection-Positive, SP, group), being rejected in the selection procedure, but getting into the curriculum through the national weighted lottery (the Selection-Negative, SN, group) or not participating in the selection procedure at all and being admitted to the curriculum through the national weighted lottery (the Lottery-Only, LO, group). Which groups of applicants and/or students were researched in which study of this thesis is also shown in Figure 7.1.

The overall aim of this thesis was to investigate whether the relatively new manner of devising an outcome-based selection procedure, as applied in Maastricht, would be supported in terms of evidence for validity. Therefore, this dissertation describes a 'quest for validity'. To investigate the different sources of evidence for validity, each empirical study included in this thesis looked at (at least one of) the pillars of validity, as defined in the general introduction. The specific aims per chapter were as follows:

7.1.1 Chapter two

The main research question of this first empirical chapter was: how does performance in a medical bachelor curriculum differ between students that were selected (SP) versus those who were rejected (SN) in the same outcome-based selection procedure? In answering this question, we aimed to examine whether our outcome-based, holistic selection procedure is predictive of study success in a pre-clinical medical bachelor.

7.1.2 Chapter three

In this chapter we aimed to assess the relationship between performance at selection and during the clinical years (i.e. master program) of medical school, in the context of a medical school where the selection procedure, curriculum and assessments are all aligned with the Canadian Medical Education Directions for Specialists (CanMEDS)

competency framework. We examined whether students who were selected via the outcome-based (i.e. CanMEDS-based) selection procedure perform better in the CanMEDS roles during the clinical phase of their medical program, compared to students who were rejected in the selection procedure and entered medical school via a national lottery weighted on pre-university GPA (pu-GPA).

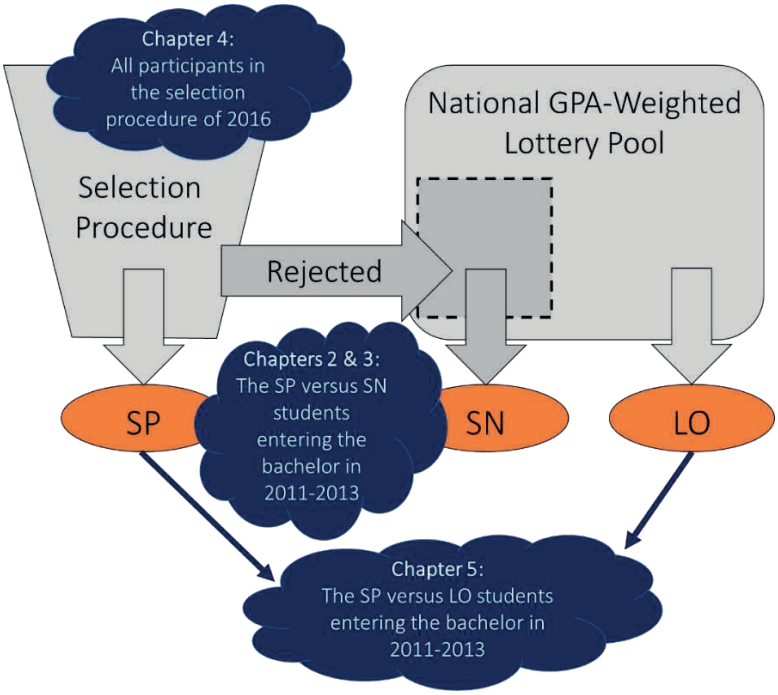


Figure 7.1: Schematic of the three routes of entry into the medical curriculum at Maastricht University in the years 2011-2013; as of 2014 all students entered through selection. Included in the dark clouds is an overview of participants in the empirical studies corresponding to the different chapters of this thesis, with SP = Selection Positive (i.e. selected in the selection procedure), SN = Selection Negative (i.e. rejected in the selection procedure but admitted through lottery) and LO = Lottery Only (i.e. did not participate in the selection procedure and admitted through lottery)

7.1.3 Chapter four

The aim of the fourth chapter was to address an important gap in knowledge with respect to the construct validity of medical school selection procedures. This was done by focusing on the content of the procedure on the one hand, and the internal structure of the procedure on the other. The specific question to be answered in this study was: do the internal structures of the tests in the selection procedure reflect the content that was intended to be measured? To be able to answer this question, a modern validity theory was applied (i.e. Downing's validity framework; 1), as well as a novel test theory: Cognitive Diagnostic Modeling (2).

7.1.4 Chapter five

The aim of Chapter five was to determine whether the benefits of applying a tailor-made outcome-based selection procedure outweigh the costs this process entails in comparison with a lottery procedure, from the perspective of the medical school. The ultimate goal of this study was to contribute information for decision making on whether to continue investing time and money in developing and adapting selection procedures, or to (re)introduce the inexpensive lottery procedure.

7.1.5 Chapter six

The sixth chapter in this dissertation was a commentary to an article by Jonathan Foo and colleagues (3), who concluded that cost evaluation studies conducted in the field of Health Professions Education (HPE) are rare and often of substandard quality. In order for stakeholders to make well-founded decisions, information on costs as well as value is paramount. However, this information is sorely missing in the literature. In the context of selection for medical school, the situation is equally dire. We generalized this situation to validity: research in selection has focused on predictive value, ignoring among others the 'consequences' pillar of validity. Additional research into costs as well as value of selection tools and procedures is crucial to support evidence-based decisions of stakeholders on admission procedures for medical school.

7.2 Main findings

Overall, the empirical studies included in this dissertation all add information to a validity argument for a selection procedure based on the principles of constructive alignment (backward chaining). Chapters two and three provided insight into our selection procedure's *relation to other variables*, chapter four looked into the *content* and *internal structure* of the selection procedure's second round, and the fifth chapter looked at the *consequences* of our selection procedure in terms of costs and benefits. All of these pieces of information can be integrated into the validity framework established in the General Introduction (chapter one) of this dissertation, based on Downing's framework of validity (1) but incorporating more of the modern validity theories as well as specific Health Professions Education literature (4-8). In the following sections, the pieces of information that were gathered throughout this dissertation are integrated into the validity framework below, based on the toolbox in the Appendix. An overview of the main findings in the context of the validity framework introduced in the General Introduction is shown in Figure 7.2.

7.2.1 Proposed use

The specific assessment conducted in the current dissertation was a selection 'at the gate' for medical school. This selection is necessary, as there are many more applicants than places in the curriculum, and the students must be carefully chosen. In fact, it is mandated by Dutch law to select students based on at least two qualitative criteria.

Results from the selection procedure's assessments are interpreted not independently, but solely in relation to the results of other applicants completing the same assessments. No clear classification into categories of competent or incompetent is defined, a decision of admission or rejection is solely based on how well applicants perform in comparison to each other, within their cohort. This admission-decision is very high-stakes (9), and should not be taken lightly. Therefore, we have looked at the evidence gathered on the five sources of evidence for validity.

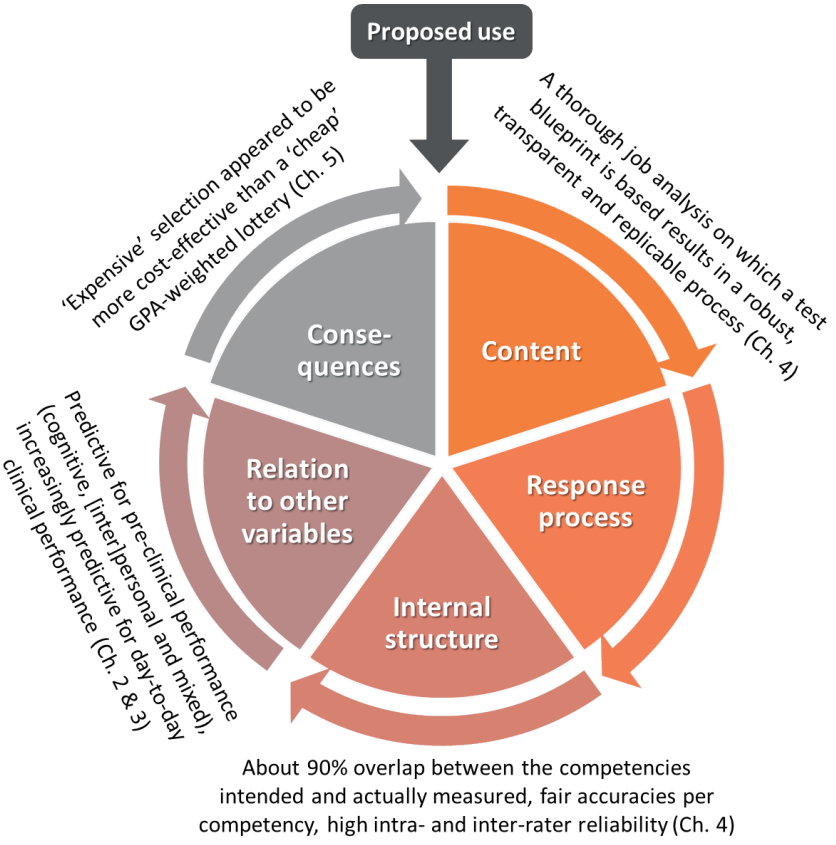


Figure 7.2: An overview of the main results derived from the different chapters of the current thesis, provided within the context of the validity theory set forth in the General Introduction (Ch. = Chapter)

7.2.2 Content

The evidence we gathered in relation to the content of the selection procedure was based mainly on the blueprint that was derived from an outcome framework which is relatively universally applied: the CanMEDS (10, 11). This competency framework is seen as an extensive and thorough job analysis of the future clinical performance needed from a graduating Medical Doctor. The strength of blueprinting the selection to this outcome framework is in the fact that the entire medical curriculum at

Maastricht University is also blueprinted to these same outcomes. Therefore, constructive alignment was established from selection, through pre-clinical and into clinical practice (see also chapter three).

As explained shortly in chapter one and two and extensively in chapter four, the CanMEDS framework was modified for the selection procedure: the level of the competencies was adapted to the level of the applicants (instead of competencies for students graduating medical school). Also, the applicants' knowledge of, and fit with, the Problem-Based Learning (PBL) environment at Maastricht University was taken into account. The translation of the outcome framework into the so-called 'derived competencies' (the CanMEDS competencies translated to entry-level competencies) was done by Subject Matter Experts (SMEs), a committee consisting of eight experts in education and medical practice. Together, this selection committee comprised all expertise necessary to perform that translating to entry-level, as well as to care for the representativeness and quality of the items, which they also wrote. Coverage of the domains was ensured by the head of the committee, who mapped each item to the blueprint in agreement with the SMEs. More specific information on the evidence based on content can be found in chapter four. For now, the evidence gathered throughout this dissertation gives us a fair base of evidence based on content.

7.2.3 Response processes

During the course of the PhD trajectory, no designated empirical study to gather validity evidence related to response processes has been performed. However, some preliminary results have been found that warrant discussion. Firstly, we had already invested effort in an attempt to find an optimal weighting of the different assessments and/or competencies within the assessments. A plethora of possibilities of weights and especially combinations of weights combined with the fact that no modelling technique was known to us made our attempt to find the optimal weighting a lot like trying to find a very specific needle in a massive stack of slightly different needles. To make it manageable, we focused on the portfolio, i.e. the first round of the selection procedure. The outcomes we looked at were mostly 'widening access': finding out whether changes in the composition of applicants in the second round would occur in terms of gender and age if the weighting of the different items within the first-round portfolio (academic performance, extracurricular activities, fit with PBL, fit with the medical curriculum at Maastricht University) were to be changed. This proved to be the case: with every alternative weighting we studied, it would be harder for male and older applicants to get into round two. These are already the minority groups, as they simply apply to the medical curriculum at Maastricht University less. Information on other minority groups (e.g. low Socio-Economic Status, first generation university students, etcetera) was unavailable to us, and therefore could not be taken into account. In the weighting as it is currently applied, there is no bias against males or older students. Therefore, based on these preliminary analyses, we decided not to change the weighting in the portfolio yet. In contrast, Fielding and colleagues found no obvious changes in the proportions of subgroups in their admissions in a UK-wide, longitudinal study, although large changes in selection policies occurred (12). For the

future, it remains an intriguing avenue of research we remain highly interested in, especially if modeling techniques were to be found.

Another important item in the toolbox under response processes is the correspondence between the item and the response options. About 80% of the items throughout the selection procedure were open-ended questions, because this provided the applicants with the opportunity to provide their own insights, opinions, approaches, and so on. In order to make our Situational Judgement Test (SJT, part of the second round of our selection procedure) more 'situational' (i.e. to make the situation more real and hopefully therewith the results more valid), the SJT was video-based. Reviews of literature found that video-based SJTs are generally perceived as more acceptable and have higher 'operational validity' and criterion-related validity, as well as less adverse impact (13, 14). To get the most variable and personal answers from candidates, open-ended items were seen as the best fit with this specific tool. The same goes for the Written Aptitude Test: the more situational the questions were, the more open-ended the items were. The first-round portfolio consisted of fill-in exercises (e.g. for pu-GPA) as well as structured open-ended questions in which interpersonal differences could be emphasized (i.e. for distinguishing skills and knowledge of and fit with PBL).

Furthermore, the Cognitive Diagnostic Modelling (CDM) approach explicated in chapter four provided us with information on which competencies were needed to answer which items, and to which extent. This way, CDM gave us information on whether the processing in the applicants' mind indeed comprised the intended competencies. This option opens up new avenues of research and has practical implications as well (see 7.4). For the current dissertation, however, the abovementioned limited pieces of evidence show that there is some, but not yet a strong foundation of evidence in terms of response processes for our selection procedure.

7.2.4 Internal structure

Internal structure mostly concerns the psychometric qualities of assessment, and whether these fit with what was expected, based on the content. Several aspects of internal structure were assessed over the course of this dissertation, mainly in chapter four. The test reliability was shown to be acceptable for the subscales (i.e. competencies) in round two. Inter- and intra-rater reliabilities were high (chapters two and four). Item functioning was good, as assessed qualitatively using expert opinions. Standard procedures for quantitatively assessing item functioning were inadequate for the data of our selection procedure (multidimensional between as well as within items). Therefore, the psychometric model was assessed using CDM, which showed a huge overlap between the expected blueprint and the actually measured constructs: 92% for the SJT and 84% for the Written Aptitude Test. Fairness was assessed in terms of Differential Item Functioning, and it was shown that there was no bias against gender or age. However, having a higher pu-GPA consistently correlated with higher performance on the selection assessments. Overall, there is fair evidence supporting

the internal structure of our selection procedure, with a specific focus on the second round.

7.2.5 Relation to other variables

For the relation to other variables, and in line with international research on selection, we focused on convergent predictive measures: outcomes throughout the students' medical studies (predictive) that are supposed to be highly related to what we assessed in our selection procedure (convergent). In chapter two we focused on the relatively short-term outcomes during the medical school bachelor (i.e. pre-clinical phase of medical school). We found that our selection procedure as a whole was predictive of performance throughout the medical bachelor, and especially so for the objective structured clinical examinations (OSCEs). This was encouraging, as OSCEs are known to be highly related to medical practice (15), and therefore we hypothesized that our selection procedure would also be predictive of clerkship performance in the master phase (i.e. clinical phase) of medical school. This was assessed in chapter three. We looked at the end-of-year assessments per CanMEDS competency, which take into account all clerkships of about a year (more specific information can be found in chapter three). The main advantage in this study was the fact that the clerkships are graded per CanMEDS competency, taking into account a multitude of feedback in different situations and by different assessors. As we also selected our students based on the CanMEDS competencies, this study could help us see whether we 'translated' the CanMEDS competencies correctly and whether this resulted in a predictive effect even throughout the clinical phase. Not only did we find the selection procedure to be predictive of performance in the clinical phase of medical school, the predictive value actually increased throughout the clinical phase, even though the predictor and criterion were very distal. Over 90% of the students who do not drop out in the first year of the medical curriculum at Maastricht University achieve the required level expected from them at the end of medical school (i.e. as defined in the 'Raamplan'; 16). However, the selected students exceeded expectations significantly more often than the originally rejected students who got in through lottery. All in all, we can conclude that there is good evidence for the relationship of our selection procedure to other variables.

7.2.6 Consequences

There are many possible consequences related to any high-stakes assessment, intended as well as unintended. From the plethora of possible consequences we could investigate, we chose to look at the cost-effectiveness of our procedure, as this is a much called-for area of research in general (3, 17, 18), but it is also a highly debated topic in selection practice (see chapter six). In the Netherlands, this discussion especially focuses on whether to reintroduce a national (weighted) lottery (more on that in section 7.3.4, 'The selection versus lottery discussion'). Therefore, we compared the costs and benefits of our selection procedure to those of the lottery procedure. This study was explicitly conducted from the perspective of the medical school, which is absolutely not the only and likely not even the most important perspective, but it was the only calculable perspective. We found that the selection

procedure we execute every year is (as was to be expected) much more expensive than the lottery procedure, which (from the medical school's perspective) does not cost anything. However, the selection procedure was found to pay itself back in terms of value during the bachelor (i.e. pre-clinical) phase of medical school alone. There is even some 'profit' gained by the selection procedure, which can be used for further improvement of quality of education. This study provides strong evidence for validity on the basis of *consequences*. However, it should be kept in mind that the area of consequences is huge, and a lot remains to be done in cost and value research, but also on (un)intended consequences, false positives and/or false negatives, and so on (see Appendix).

In the next sections, the results from the empirical studies in this thesis will be discussed in relation to (inter)national discussions and literature.

7.3 (Inter)national discussions

In the General Introduction of this thesis (chapter one), some discussions were touched upon, such as the cognitive versus (inter)personal competencies divide, the discussion on which tools are best, or which content should be assessed (could constructive alignment be a way forward?). Below, these relevant topics are elaborated, using our own findings as well as literature. In addition, the discussion specific to the Netherlands, concerning whether or not it is time to go back to the national lottery, will be addressed.

7.3.1 Cognitive versus (inter)personal

As explained in the General Introduction (chapter one), cognitive ability has been used as a tool for selecting medical students for over a century now (19). A lot of research has already shown that cognitive, or academic, ability is highly predictive of the first phase of medical school (13, 19-22). The first phase is pre-clinical and mostly involves relatively standard learning of new information, combined with skills needed for medical practice. This predictive value has previously been shown to decrease from the pre-clinical to the clinical phase (20). In contrast to this evidence, we found that previous academic attainment, and therewith academic ability, remains highly predictive of performance even throughout the clinical phase (chapter three). The next development was set forward in the Edinburgh declaration (23) and embraced after the study by Papadakis and colleagues, who showed that unprofessional behaviors in medical practice were often preceded by unprofessional behaviors during medical studies (24): the idea to select on (inter)personal competencies (25). This has led to an explosion of research looking at the predictive value of assessments of (inter)personal skills, often without much consideration of what specific constructs were actually being measured (26-28). Because of this lack of specificity, mixed results were found all over the world, with no way to compare them or draw conclusions. One way of making sense of all these different results, was by categorizing them in terms of the tools used. This is expanded upon in the next section.

7.3.2 Which tools should be used?

The categorization of research into tools is very popular, and much research has focused on finding out which tools are most reliable, acceptable and valid. In doing so, it is often ignored that, for example, every school applying a Multiple Mini Interview (MMI) has a different number and duration of stations, with assessors who differ in many ways including experience, level of training, attitudes towards selection, etc. These are the simple things to be streamlined: research might converge and be able to tell us how much of each of these is needed (e.g. the methodologies used in 29, 30). However, the next problem in reaching consensus on (for example) MMIs is to talk about which constructs should be assessed. Each medical school measures slightly (or entirely) different constructs, in slightly (or entirely) different ways, even if they give their tools the same names. Of course, meta-analyses on the tools are useful in general, showing us that traditional interviews, autobiographical submissions and references are simply never useful: not in terms of predictive value, not in terms of reliability and validity, and not in terms of fairness (13, 19-22). These results are so stable that as a field, we can throw them in the bin with complete peace of mind. However, tools such as the MMI, but also the SJTs, structured interviews and aptitude tests often find varying results due to the abovementioned reasons. Therefore, the real challenge is in determining which of those mixed-evidence tools to use or combine.

Instead of focusing on this mixed evidence and attempting to make sense of the overwhelming amount of research on the tools' predictive values, why do we not look at the content of these tools and use thorough methods of introducing content into those tools to create reliable, valid and predictive selection procedures (31)? MMIs, structured interviews, academic records, SJTs and aptitude tests have all shown merit, although maybe not at every single use. Is the same not true for curricular assessments? If we would choose our tools based on what we want to measure and place these tools in service of the content, we might just get a little closer to that holy grail we have been searching for: a predictive, reliable, valid, fair, acceptable and cost-effective selection procedure.

7.3.3 What content should be assessed?

One of the things the field of selection seems to agree on is that selection should be based on the outcomes of medical school. Best practice is starting from a job analysis and basing the selection procedure on that (32). However, a tool is often chosen first, and then relatively random-appearing outcomes may be chosen to base the content of the tool on. In cost evaluations, researchers often skip over how they calculated costs or values (3); in the same manner, selection researchers often skip over how they chose which constructs to measure. In contrast to cost evaluations, however, this has gotten much better over the years, as Patterson, Kerrin and colleagues have been spreading the word on the best practice mentioned above (25, 32-34). Once again, we echo their call in using content as the absolute starting point of devising a selection procedure, and basing the choice of tools on the constructs you want to assess and the manner in which you want to assess them. Furthermore, like Foo et al. for cost

evaluations (3), we ask for clearer explanations of constructs and methodologies for selection research (35). In this way, we may be able to create a more generalizable base of research, which also provides more opportunities for (inter)national collaborations.

7.3.4 The selection versus lottery discussion

In the Netherlands, there currently is a discussion on whether selection is worth the hassle. Some researchers find no added value of selection, while questioning whether selection hampers student diversity (36). At another university, researchers showed that cognitive and non-cognitive procedures seemed to predict different parts of the curriculum, leading them to recommend 'curriculum sample procedures' (37): the parts of the curriculum sampled in the selection procedure should be the most relevant ones, the courses/exams/activities intended to be predicted in actual medical school. At yet another university, three admissions processes were compared, showing that pu-GPA was the best predictor for pre-clinical academic performance, while non-academic performance within the bachelor was best predicted by a multifaceted selection procedure (a first-round portfolio containing pre-university education, extracurricular activities and reflection, and a second-round containing a writing assignment, a patient lecture with subsequent assignments, a scientific reasoning block, and a series of short interviews and role-plays; (38). The same author later compared the predictive value of selection procedures at three universities for first-year results, concluding again that top pu-GPA was most predictive, and distinguishing between a participation and selection effect (39). The participation effect is the effect that the applicants who participate in a selection procedure often appear more motivated and therefore attain some beneficial results (e.g. less drop-out) when compared to those entering through the national weighted lottery only. The selection effect would be the effect where the students who were accepted in the selection procedure outperform the students who were rejected in the selection procedure but entered medical school through the national weighted lottery. The study comparing three universities found a participation effect for most outcomes at most universities, but only rarely found a selection effect (39). Because of these differences in results, and a relative focus on the trouble in finding clear selection effects, the consensus in practice seems to be tilting towards a preference for a lottery procedure.

With this dissertation we hope to stop that tilting and direct more attention towards the careful planning of constructs and thorough evaluation of selection procedures, so that we can achieve fair and predictive selection procedures. Our results are in line with some of the results already presented in the Dutch literature. For example, throughout the medical curriculum at Maastricht University too, there is a participation effect (lottery-only students are especially prone to drop-out, as shown in chapter five). However, there are also opposing results: a clear selection effect is found throughout the medical curriculum at Maastricht University (over the course of the bachelor as well as the master, and especially in selecting for excellence). Furthermore, in line with the literature, there indeed seems to be an independent and

persistent positive effect of students' pu-GPA throughout the preclinical bachelor, but in contrast to the majority of research, we also found pu-GPA to have a persistent positive effect throughout the clinical master program. Important to note is that the predictive effect of pu-GPA is independent from the effect of selection (i.e. they show incremental value).

7.4 Implications

There are several implications worth expanding on, for research and theory on selection as well as for selection practice. These are expanded on below.

7.4.1 Research and theory

For research and theory, the implications of this dissertation are manifold. On the one hand, there are implications for the selection procedure in itself and the way it was set up. On the other, there are some methodological and statistical procedures that would benefit research on selection and hopefully bring a deeper understanding to the researchers in the field of selection, as with this dissertation, we have only scratched the surface.

Related to our selection procedure, we endorse Patterson's (25, 40, 41) and Kerrin's (32) calls to continue, or start, using a job analysis as the starting point for a selection procedure, reducing the importance of the tools the competencies stemming from the job analysis are assessed with (31). Furthermore, it may be helpful to view our selection procedure as an example of a translation from outcomes to an outcome-based selection procedure using a constructive alignment approach. It is likely that many medical schools are in fact already applying this best practice, but do not explicitly state this in their research papers. We hope that in the future, this reporting will be improved.

Related to our methodological and statistical procedures, we hope to have shown some of the numerous possible manners in which research on selection can be conducted. Of course, our Dutch situation with the natural control group is a methodological luxury that does not occur often. However, a cost evaluation study or hopefully even cost-benefit analyses can be conducted at any university, but there is a distressing shortage of such studies. For statistical procedures, one of the main implications of this dissertation is the application of Cognitive Diagnostic Modelling (CDM) in selection research. Applying this test theory has helped us enormously, not only in finding out whether we were measuring what we wanted to measure. It also stimulated us to think more critically about what we were really trying to measure at the item-writing stage and to have clearer definitions of the constructs in our heads. Moreover, this technique is extremely sophisticated and has countless more options, each one providing more information than the last. The one empirical study we wrote on that subject could have been expanded into a book, had we taken full advantage of the CDM test theory and had we had ample time.

From a theoretical point of view, diving deep into the modern validity theories has been about as transformative as applying CDM. It has been said that "validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few" (Ebel, 1961; 42). Validity is seen as an important feature, but the stamp 'valid' is given with relative ease. The modern validity theories, although each in their own manner, go against this easiness (1, 4, 5, 7, 8). They make you critical towards your own research and others' research as well as your own selection procedure, assessments, and curriculum. In the field of selection, especially, the application of the modern validity theories would help push the research forward immensely.

7.4.2 Practice

The implications for practice were touched upon slightly in the paragraphs above, but warrant some more attention. The studies conducted throughout the course of this dissertation have greatly helped the practice of selection at our medical school. One unexpected but positive consequence of our studies was the recognition and appreciation for our devoted selection committee members. Their job as selection committee is important and (rightfully) scrutinized, but they worked hard to create, refine, assess and evaluate all the items, every year again. This research showed them that they were doing a great job, and that all of their effort was really making a difference. Furthermore, it clearly showed the university-stakeholders why the selection procedure was in place and that is was worth keeping in this form (in terms of predictive value as well as costs).

As alluded to before, the application of CDM had important implications, especially for theory and research. However, CDM also has important practical inferences. For example, CDM produces a graph for each item stating the chances of getting the item right depending on which competencies each applicant possesses, helping the selection committee understand what they are actually asking of their applicants. Furthermore, CDM can export a table containing all applicants and providing an overview on which competencies each applicant is above or below average (in a dichotomized analysis). This could be used, for example, as feedback to students. If this data were to be provided to the students, this could help them use the selection procedure as a learning moment. The possibilities with CDM seem endless, and they appear worthwhile for practice.

The application of modern validity frameworks may also be useful in practice, because it not only induces more critical reasoning about selection, but it almost provides a roadmap on how to create procedures for which validity evidence will actually be supportive. For example, it is a reminder not to skip to tools before thinking about the content, to always keep the goal in mind, to put ourselves in the applicants' shoes (i.e. to take into account the response process), and to think about what the psychometrics are actually saying instead of being overwhelmed by all the numbers.

7.5 Strengths and limitations

The strengths of our studies have been set forward previously, and are repeated very shortly: the multi-tool outcome-based selection procedure, the natural control group in the Netherlands, the new methodologies as well as statistical techniques, the application of modern validity theories, the research into procedures as a whole as well as studying the entirety of medical school (bachelor as well as master phase), and the use of multiple cohorts in most studies.

One limitation of the studies in the current dissertation was the fact that all research was conducted at a single institution: Maastricht University. This makes it difficult to generalize our findings to other institutions, especially since all institutions use a different selection procedure. Therefore, in the implications section, we have placed focus on the methods and analyses that are in fact generalizable to other institutes. These other institutes may be able to apply the same methods, creating a basis for comparisons.

Another limitation, standing in the way of using more thorough or elegant analyses, is the sample size. Even though we combined up to three cohorts in our empirical studies, some analyses simply need more participants in order to achieve enough power. For the CDM, for example, we could only use the dichotomized data, because otherwise the model simply would not converge. The fact that the SJT and Written Aptitude Test items change every year rendered the combination of several cohorts impossible for this study.

A limitation related to the validity framework set out in this thesis is the fact that we did not conduct a study on evidence related to response processes. This was due to time limitations within the PhD trajectory. We look forward to studying this subject further. However, for the current dissertation and for the support gathered for the selection procedure this is a clear gap.

7.6 Further research

There are a few studies that we think are very important to conduct in the near future. The first is a study on the optimal weighting of the different constructs. As stated before, up to now we have only performed a first attempt that had to be stopped due to time restrictions. A well-executed weighting study could combine striving for cost-effectiveness with the endeavor of fair selection as well as making sure that the right outcomes are predicted using the right constructs.

Another important avenue for research is the collaboration with other universities. For example, where other universities in the Netherlands have found results different from ours, it would be interesting to try and find out the origin of those differences. Furthermore, international collaborations would be interesting, to look at the cultural generalizability of selection constructs and procedures.

Furthermore, as stated above, CDM provides a competency-mastery score (above average or below average) for each competency per applicant. It would be interesting to link these with the related competencies in the clinical master phase, to see whether these constructs are stable and to further support that these constructs are indeed the constructs we intended to measure.

7.7 Conclusions

In conclusion, based on the current best practice in creating content (choosing constructs) to assess at selection, and based on the positive results we found throughout this dissertation, we endorse explicit constructive alignment as a way forward for selection. Furthermore, as Kulasegaram states: “creating fair, responsible, and robust admissions processes will require us to continually develop our theories as well as our tools” (26). Thus, we would recommend evaluating selection procedures using modern validity frameworks to achieve a more general understanding of the validity of our procedures than just prediction and sometimes incremental value. A promising test theory to help in applying those validity theories is CDM. In combination, they could really help the field of selection forward and towards a ‘holy grail’: a predictive, reliable, valid, fair, acceptable and cost-effective selection procedure.

7.8 References

1. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-7.
2. Sorrel MA, Olea J, Abad FJ, de la Torre J, Aguado D, Lievens F. Validity and Reliability of Situational Judgement Test Scores: A New Approach Based on Cognitive Diagnosis Models. Organizational Research Methods. 2016;19(3):506-32.
3. Foo J, Cook DA, Walsh K, Golub R, Abdalla ME, Ilic D, et al. Cost evaluations in health professions education: a systematic review of methods and reporting quality. Med Educ. 2019.
4. Messick S. Validity of Psychological-Assessment - Validation of Inferences from Persons Responses and Performances as Scientific Inquiry into Score Meaning. American Psychologist. 1995;50(9):741-9.
5. Royal KD. Four tenets of modern validity theory for medical education assessment and evaluation. Adv Med Educ Pract. 2017;8:567-70.
6. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 2015;49(6):560-75.
7. Kane MT. An Argument-Based Approach to Validity. Psychological Bulletin. 1992;112(3):527-35.
8. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, United States of America: American Educational Research Association; 2014.
9. Mercer A, Hay M, Hodgson WC, Canny BJ, Puddey IB. The relative predictive value of undergraduate versus graduate selection tools in two Australian medical schools. Med Teach. 2018:1-8.
10. Frank JR. The CanMEDS 2005 physician competency framework: Better standards, better physicians, better care. 2005. Available from: http://www.ub.edu/medicina_unitateducaciomedica/documentos/CanMeds.pdf.
11. Frank JR, Snell LS, Sherbino J. The draft CanMEDS 2015 physician competency framework - series ii. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2014.
12. Fielding S, Tiffin PA, Greatrix R, Lee AJ, Patterson F, Nicholson S, et al. Do changing medical admissions practices in the UK impact on who is admitted? An interrupted time series analysis. Bmj Open. 2018;8(10):e023274.
13. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ. 2016;50(1):36-60.
14. Lievens F, Peeters H, Schollaert E. Situational judgment tests: a review of recent research. Pers Rev. 2008;37(4):426-41.
15. Eva KW, Reiter HI. Where judgement fails: pitfalls in the selection process for medical personnel. Adv Health Sci Educ Theory Pract. 2004;9(2):161-74.
16. Van Herwaarden CLA, Laan RFJM, Leunissen RRM. Raamplan Artsopleiding 2009. NFU, editor. Utrecht2009.
17. Maloney S, Foo J, Cook D, Walsh K. Cost, Value, and the Sustainability of Our Choices Concerning Simulation. Acad Med. 2018;93(3):342-3.
18. Cook DA, Beckman TJ. High-value, cost-conscious medical education. JAMA Pediatr. 2015;169(2):109-11.
19. Siu E, Reiter HI. Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. Adv Health Sci Educ Theory Pract. 2009;14(5):759-75.
20. Cleland J, Dowell J, McLachlan J, Nicholson S, Patterson F. Identifying best practice in the selection of medical students (literature review and interview survey). https://www.gmc-uk.org/-/media/about/identifyingbestpracticeintheselectionofmedicalstudentspdf51119804.pdf; 2012.
21. Patterson F, Roberts C, Hanson MD, Hampe W, Eva K, Ponnamperuma G, et al. 2018 Ottawa consensus statement: Selection and recruitment to the healthcare professions. Med Teach. 2018;40(11):1-11.
22. Prideaux D, Roberts C, Eva K, Centeno A, McCrorie P, McManus C, et al. Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):215-23.
23. Roddie IC. The Edinburgh Declaration. The Lancet. 1988;332(8616):908.
24. Papadakis MA, Teherani A, Banach MA, Knettler TR, Rattner SL, Stern DT, et al. Disciplinary action by medical boards and prior behavior in medical school. N Engl J Med. 2005;353(25):2673-82.

25. Patterson F, Zibarras L, editors. Selection and Recruitment in the Healthcare Professions: Research, theory and practice. Cham, Switzerland: Springer Nature Switzerland AG; 2018.
26. Kulasegaram K. Use and ornament: expanding validity evidence in admissions. Adv Health Sci Educ Theory Pract. 2017;22(2):553-7.
27. Hecker K, Norman G. Have admissions committees considered all the evidence? Adv Health Sci Educ Theory Pract. 2017;22(2):573-6.
28. Kreiter CD. A research agenda for establishing the validity of non-academic assessments of medical school applicants. Adv Health Sci Educ Theory Pract. 2016;21(5):1081-5.
29. Hissbach JC, Sehner S, Harendza S, Hampe W. Cutting costs of multiple mini-interviews - changes in reliability and efficiency of the Hamburg medical school admission test between two applications. BMC Med Educ. 2014;14:54.
30. Castanelli DJ, Moonen-van Loon JMW, Jolly B, Weller JM. The reliability of a portfolio of workplace-based assessments in anesthesia training. Can J Anaesth. 2019;66(2):193-200.
31. Pearce J, Jackel B. SJT, MCQ, ETC… The worrying conflation of format and content. Med Educ. 2018;52(9):993-.
32. Kerrin M, Mossop L, Morley E, Fleming G., Flaxman C. Role Analysis: The Foundation for Selection Systems. In: Patterson F, Zibarras L, editors. Selection and Recruitment in the Healthcare Professions: Research, theory and practice. Cham, Switzerland: Springer Nature Switzerland AG; 2018.
33. Patterson F, Tavabie A, Denney M, Kerrin M, Ashworth V, Koczwara A, et al. A new competency model for general practice: implications for selection, training, and careers. Br J Gen Pract. 2013;63(610):e331-8.
34. Patterson F, Zibarras L, Ashworth V. Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. Med Teach. 2016;38(1):3-17.
35. Schreurs S, Cleutjens K, Oude Egbrink MGA. Increasing value in research: cost evaluations in health professions education. Med Educ. 2019;53(12):1171-3.
36. Wouters A. Effects of medical school selection on student motivation: a PhD thesis report. Perspect Med Educ. 2018;7(1):54-7.
37. de Visser M, Fluit C, Cohen-Schotanus J, Laan R. The effects of a non-cognitive versus cognitive admission procedure within cohorts in one medical school. Adv Health Sci Educ Theory Pract. 2018;23(1):187-200.
38. Schripsema NR, van Trigt AM, Borleffs JC, Cohen-Schotanus J. Selection and study performance: comparing three admission processes within one medical school. Med Educ. 2014;48(12):1201-10.
39. Schripsema NR, van Trigt AM, Lucieer SM, Wouters A, Croiset G, Themmen APN, et al. Participation and selection effects of a voluntary selection process. Adv Health Sci Educ Theory Pract. 2017;22(2):463-76.
40. Patterson F, Archer V, Kerrin M, Carr V, Faulkes L, Coan P, et al. FY1 job analysis report: Improving selection to the foundation programme. 2010 [125-240]. Available from: https://isfporguk.files.wordpress.com/2017/04/appendix-d-fy1-job-analysis.pdf.
41. Patterson F, Ferguson E, Thomas S. Using job analysis to identify core and specific competencies: implications for selection and recruitment. Med Educ. 2008;42(12):1195-204.
42. Ebel RL. Must all tests be valid? American Psychologist. 1961;16(10):640.

CHAPTER 8

Summary

Selection for medical school is an important and somewhat controversial topic. In Chapter 1, we explained that especially in the Netherlands, discussions about whether we should be selecting our medical students are flaring up. The biggest argument used to banish selection procedures is that they are expensive and unpredictive. However, every university does it in their own manner, because of which each university has different expenses and predictive values. Currently, it is hard to see the wood for the trees in the forest that is selection: what tools should be used, should they combine cognitive and (inter)personal competencies, what specific competencies or skills should be measured, should we focus on selecting students or doctors?

Current best practice for selection is starting with a job analysis and using the competencies emerging from this job analysis to base your selection procedure on. At Maastricht University, we did this by applying '*constructive alignment*': the outcome frameworks used to assess whether a student is ready to be a doctor by the end of medical school, are also the focus of the clinical and pre-clinical phase of medical school, as well as our selection procedure. In other words, we used backward chaining to reason back from the end goals of medical school to input for our selection procedure. The competencies described in the outcome framework were translated into derived competencies adapted to the level of medical school applicants.

In the empirical studies in this thesis, we took on the quest for validity of this new setup of selection for medical school. To be able to assess the usefulness of this 'constructive alignment' approach to selection, we applied a 'modern validity framework'. Modern validity frameworks explain all validity as construct validity (i.e. are you measuring what you intend to measure), and state that validity is a continuum and an ongoing process related to inferences, not tools or assessments in themselves. The validity framework we derived from multiple modern validity frameworks focused on supporting construct validity using five sources of evidence: content, response processes, internal structure, relation to other variables, and consequences. This validity framework was applied to the Maastricht selection procedure over the course of four empirical studies.

In the empirical study described in Chapter 2, the focus was on evidence for validity based on the 'relation to other variables' pillar of the validity framework: the students' performance throughout the medical bachelor. The aim of this study was to examine if the outcome-based selection procedure developed at Maastricht University was predictive of study success in the bachelor phase. To do so, the relationship between performance at selection and later study success across the three-year bachelor program was examined in three student cohorts (2011, 2012 and 2013). These cohorts were chosen because in these years, there were two parallel admission routes providing entry to the medical bachelor curriculum: the decentralized (i.e. individual to each university) selection procedure or the national weighted lottery. In this study, two groups were compared. Firstly, the students who were selected in the selection

procedure and therefore got into the medical curriculum through selection (selection-positive students). Secondly, the students who entered the medical bachelor curriculum through the national weighted lottery procedure after having been rejected in the selection procedure (selection-negative students). Study results were compared between these selection-positive and selection-negative students.

The student results taken into account in this study were all graded assessments in the medical bachelor curriculum, divided into four categories: cognitive, (inter)personal, mixed and general. It was found that selection-positive students outperformed their selection-negative counterparts throughout the entire bachelor program on assessments measuring cognitive (e.g. written exams), (inter)personal (e.g. communication and reflection skills) and combined outcomes (i.e. objective structured clinical examinations). No significant effects on general parameters (e.g., drop-out and delay) were found. Of the 30 outcome variables, selection-positive students scored significantly higher in 11 cases. Fifteen other, non-significant group differences were also in favor of the selection-positives. An overall comparison using a sign test indicated a significant difference between both groups ($p < 0.001$), despite equal pre-university Grade Point Averages (GPAs).

All in all, it seemed that the use of a constructively aligned selection approach addresses some of the predictive validity limitations of commonly-used selection tools: a constructively aligned selection procedure appears to be capable of predicting multiple categories of outcomes in the medical bachelor. Selection-positive students significantly outperformed their selection-negative counterparts across a range of cognitive, (inter)personal, and mixed outcomes throughout the entire three-year bachelor curriculum.

In Chapter 3, a follow-up study is presented. In this chapter, we investigated whether the selection procedure explained above was predictive of the students' performance in the clinical master phase of the medical curriculum as well. Thus, in terms of the validity framework, it also studied the selection procedure's relation to other variables. The aim of this study was to assess the relationship between performance at selection and during the clinical years of a medical program. Again, the context was a medical curriculum in which the selection procedure, curriculum content and assessments are all aligned with the same outcome framework (i.e. the Canadian Medical Education Directives for Specialists; CanMEDS). The set-up of this study was similar to the study described in Chapter 2. Two groups of students were compared: those selected via an outcome-based selection procedure (selection-positive students) versus those rejected in this procedure who entered the program through a national, GPA-based lottery procedure (selection-negative students).

Performances of both groups on all seven CanMEDS-roles were compared during clinical rotations, for each of the three master years. Students' performances on all competences were recorded in an electronic portfolio, in a programmatic manner: students gathered quantitative information (e.g., on knowledge tests) as well as qualitative feedback from different assessors (e.g., medical specialists, peers, nurses,

patients) at different times for different assignments. At the end of a year, the exam committee appraised all information in the portfolio and provided a final judgement for each competency: below expectations, as expected, or exceeds expectations. Data were compared for the three cohorts starting their medical bachelor curriculum in 2011, 2012 or 2013, since these are the only cohorts in which both selection-positive and selection-negative students were admitted to the medical curriculum. It was found that selection-positive students significantly outperformed the selection-negative students in all three master years, and that the differences between the two groups increased over time. In the first year, differences were apparent in the roles of Communicator, Collaborator and Professional, in the second year also for the roles of Organizer and Health Advocate, and in the third year in the role of Academic as well.

The conclusion of this study was that the constructively aligned selection procedure had an increasing predictive value across the clinical years of a medical curriculum. This suggests that constructive alignment of selection, curriculum and assessment to ultimate outcomes is effective in creating a selection procedure predictive of clinical performance.

In Chapter 4, we moved away from evidence for validity based on relation to other variables, and went on to investigate the content and internal structure of our selection procedure more closely. For some time now, medical school selection finds itself in the paradoxical situation in which selection tools may predict study outcomes, but which constructs are actually doing the predicting is unknown (the 'black box of selection'). Therefore, this study focused on those constructs, answering the question: do the internal structures of the tests in a constructively aligned selection procedure reflect the content that was intended to be measured? To do so, we had to first explicate what the content of the selection procedure was supposed to be, and then delve deep into the psychometric qualities of the tests in the procedure to see whether the psychometric qualities were a fit with the content we were intending to measure.

First, we examined content-related evidence pertaining to the creation and application of the competency-based selection blueprint. Constructive alignment was achieved by finding a well-established job analysis (i.e. the CanMEDS) and translating the competencies in the CanMEDS to the level of an applicant in the selection procedure. To do so, 'backward chaining' was applied. The resulting selection procedure was a multi-tool, CanMEDS-based procedure consisting of two rounds. The focus of this study was on the second round, which comprised a Video-based Situational Judgement Test (focused on (inter)personal competencies) and a Written Aptitude Test (reflecting a broader array of CanMEDS competencies). We found that the set-up of the selection procedure was a robust, transparent and replicable process.

Second, the internal structure of the selection tests was investigated by connecting applicants' performance on these tests to the predetermined blueprint using Cognitive Diagnostic Modeling (CDM). CDM is capable of finding latent variables when there is

multidimensionality in the data, both between and within items, as was the case in our data: different items measured different competencies, but most items also measured multiple competencies. CDM is related to Confirmatory Factor Analysis: the structure (in this case the selection procedure's blueprint) is provided, and CDM looks at whether that structure is indeed found in the data, or whether alterations to the structure make more sense on the basis of the data provided. The data indicated 89% overlap between the expected and measured constructs.

All in all, our results supported the notion that the focus placed on creating the right content and following a competency-blueprint was effective in terms of internal structure: most items measured what they were intended to measure. This way of linking a predetermined blueprint to the applicants' results sheds light into the 'black box of selection' and can be used to support the construct validity of selection procedures.

An underexposed pillar of the validity framework named in the beginning of this summary chapter is the 'consequences' one. The focus of the empirical study in Chapter 5 was on one specific part of consequences: cost-efficiency, specifically from the perspective of the medical school. This research is especially important, as resources for medical education are becoming more constrained, whereas accountability in medical education is increasing. In this constrictive environment, medical schools need to consider and justify their selection procedures in terms of costs and benefits. To date, however, there have been no studies focusing on this aspect of selection. Therefore, in this study, we aimed to examine and compare the costs and benefits of two different approaches to admission into medical school: a constructively aligned, multimethod selection process versus a GPA-weighted lottery procedure. Our goal was to assess the relative effectiveness of each approach and to compare these in terms of benefits and costs from the perspective of the medical school.

The study was conducted in the medical curriculum at Maastricht University, where a decentralized selection process and a weighted lottery procedure ran in parallel for three years (2011–2013). The costs and benefits of the selection process were compared with those of the lottery procedure over three student cohorts throughout the bachelor program. The extra costs of selection represented the monetary investment of the Institute for Education that organized the medical curriculum in conducting the selection procedure; the benefits were derived from the increase in income generated by the prevention of dropout and the reductions in extra costs due to decreases in the repetition of blocks and objective structured clinical examinations. We found that the constructively aligned selection procedure cost about €139,000 for a full cohort of students (n = 286). The lottery procedure came with negligible costs. However, the average benefits of selection compared with the lottery system added up to almost €207,000 per cohort of students. This resulted in an overall benefit of selection over lottery to the tune of €68,000.

In conclusion, this study not only showed that conducting a cost–benefit comparison is feasible in the context of selection for medical school, but also that an 'expensive' selection process can be cost beneficial in comparison with an 'inexpensive' lottery system.

Chapter 6 is a commentary to a meta-analysis by Jon Foo and colleagues, in which they showed that studies on costs are extremely scarce in Health Professions Education (HPE), and studies including value are even more rare. Furthermore, the quality of most of the available research appeared to be substandard and had not increased since the beginning of this century. These findings mirror the practice of higher education institutes: decisions on which educational activities to invest in are often not fueled by full empirical economic evaluations. We indicated that one educational area in which cost evaluations are particularly interesting is that of selection. While medical schools are under increasing pressure to justify their unique and often expensive selection procedures in terms of costs and value, research on the costs and especially the value of selection procedures, or even the tools within these procedures, is sorely lacking.

Importantly, high quality cost-evaluation research is not only relevant at the level of individual educational institutes, but also on a national level. As an example, we referred to the situation in the Netherlands: in 2017, medical schools were obliged to switch from a national weighted lottery to decentralized selection procedures which differ from school to school. Because several of these procedures were shown to be insufficiently predictive to warrant the monetary investments, the cost-effectiveness of selection in general was questioned and reintroduction of a national lottery system is proposed. This dispute has even sparked discussion at the government level. In our opinion, research into the general construct validity of the selection procedures, including high-quality full economic evaluations, are crucial to support an evidence-based governmental decision on this issue.

All in all, there is a lack of high-quality economic evaluations in HPE in general and selection in particular. This puts the stakeholders in a precarious position in which they have to remain accountable regarding their limited financial resources while choosing between educational activities for which the costs are unknown. Therefore, educational researchers should join efforts to conduct more economic evaluations to provide relevant stakeholders with economically valid arguments to make well-informed decisions that benefit the quality of education.

In Chapter 7, the General Discussion, the main findings of all previous chapters were summarized and implications and suggestions for the practice of and research on selection were debated. Overall, the empirical studies included in this dissertation all add information to a validity argument for a selection procedure based on the principles of constructive alignment. All of the studies can be integrated into the validity framework established in Chapter 1 of this dissertation:

- ✓ *Content*: a thorough job analysis on which a test blueprint is based results in a robust, transparent and replicable process

- ✓ *Response processes*: first, preliminary data suggest no effect of changes in weighting of portfolio subparts on age and gender of selected students
- ✓ *Internal structure*: about 90% overlap between the intended and actually measured competencies, fair accuracies per competency, high intra- and inter-rater reliability
- ✓ *Relation to other variables*: predictive for pre-clinical performance during a medical bachelor curriculum (cognitive, [inter]personal and mixed results), increasingly predictive for day-to-day clinical performance during the master phase
- ✓ *Consequences*: 'expensive' selection appears to be more cost-effective than a 'cheap' GPA-weighted lottery

These findings can have important repercussions for several (inter)national discussions: the cognitive versus (inter)personal discussion should not so much be a black-and-white choice, but more of a continuum, focusing on measuring relevant competencies along the entirety of this continuum. Furthermore, blindly relying on evidence of tools and disregard for the content causes mixed results in research as well as practice. Also, applying a job analysis to determine what to measure in a selection procedure is paramount. Importantly, there currently is a discussion on whether selection is worth the hassle. Results differ considerably between universities, but the focus in the discussion is on the negative results.

With this dissertation we hope to stop that tilting and direct more attention towards the careful planning of constructs and thorough evaluation of selection procedures, to achieve fair and predictive selection procedures. The implications of this thesis for theory, research and practice are manifold: the use job analyses as the starting point for a selection procedure; the use of innovative methods for selection research (e.g. Cognitive Diagnostic Modelling and cost-effectiveness); and the application of modern validity theories.

The main conclusion from the current thesis is that explicit constructive alignment can be seen as a way forward for selection. Furthermore, it is recommended to evaluate selection procedures using modern validity frameworks to achieve a more general understanding of validity. A promising test theory to help in applying those validity theories is Cognitive Diagnostic Modelling. In combination, constructive alignment, modern validity frameworks and innovative research methods (e.g. CDM and cost-effectiveness) could really help the field of selection forward and towards a 'holy grail': a predictive, reliable, valid, fair, acceptable and cost-effective selection procedure.

Samenvatting

Selectie voor de geneeskundeopleiding is een belangrijk en enigszins controversieel onderwerp. Hoofdstuk 1 beschrijft de discussies, voornamelijk in Nederland, over of geneeskunde studenten wel of niet geselecteerd zouden moeten worden. Het belangrijkste argument om van selectie af te stappen is dat selectieprocedures duur zijn, maar niet voorspellen. De kanttekening hierbij is dat de selectieprocedures verschillen per universiteit, waardoor iedere universiteit te maken heeft met andere kosten en verschillende voorspellende waardes. Op dit moment is het moeilijk om door de bomen het bos nog te zien op het gebied van selectie: welke methodes/toetsen moeten we gebruiken, moeten we daarin cognitieve en (inter)persoonlijke competenties combineren, welke specifieke competenties of vaardigheden moeten we meten en moeten we ons richten op het selecteren van de beste studenten of juist van de beste toekomstige artsen?

De huidige '*best practice*' binnen selectie is om te beginnen met een analyse van de taken van een toekomstige arts en de competenties die daarvoor nodig zijn. Die competenties worden vervolgens gebruikt als de basis voor het ontwerp van de selectieprocedure. Aan de Universiteit van Maastricht (UM) hebben we dit gedaan door '*constructive alignment*' toe te passen, met het raamplan voor de geneeskundeopleiding als basis. Dit raamplan beschrijft de uitkomsten van de geneeskundeopleiding: de competenties en de niveaus binnen die competenties die de studenten aan het einde van hun opleiding behaald moeten hebben. Het raamplan wordt dus gebruikt om te toetsen of een student klaar is om arts te zijn tegen het einde van de geneeskundeopleiding, maar ook het curriculum en de toetsing in de klinische (master)fase en pre-klinische (bachelor)fase van de UM zijn gebaseerd op deze competenties. Door middel van '*backward chaining*' werden deze competenties ook de focus van de selectieprocedure: de einddoelen van de geneeskundeopleiding zijn terug geredeneerd en aangepast aan het niveau van een kandidaat in de selectieprocedure. De resulterende 'afgeleide competenties' vormden de blauwdruk voor de selectieprocedure.

De empirische studies in dit proefschrift vormen een 'quest' voor validiteit van deze nieuwe opzet van selectie voor de geneeskundeopleiding. Om te onderzoeken hoe nuttig *constructive alignment* is als aanpak voor het opzetten van een selectieprocedure hebben we een zogenaamd modern validiteitsraamwerk toegepast. Moderne validiteitsraamwerken leggen alle validiteit uit als constructvaliditeit (meet dit instrument wat het zou moeten meten?). Binnen deze raamwerken wordt validiteit gezien als een continuüm en is validering een aanhoudend proces dat gerelateerd is aan gevolgtrekkingen (specifiek gebruik van het middel/de toets) en niet aan de middelen of toetsen zelf. Het validiteitsraamwerk dat we gebruiken in dit proefschrift is een afgeleide en combinatie van meerdere moderne validiteitsraamwerken, en is gericht op het ondersteunen van constructvaliditeit door middel van vijf bronnen van bewijs: inhoud, reactieprocessen, interne structuur, relatie met andere variabelen en

consequenties. Dit validiteitsraamwerk werd toegepast op de selectieprocedure voor de geneeskundeopleiding in Maastricht, in vier empirische studies.

Het empirisch onderzoek dat beschreven wordt in Hoofdstuk 2 was gericht op het verzamelen van validiteitsbewijs op basis van de relatie tussen selectie en andere variabelen: de prestaties van de studenten gedurende de geneeskunde bachelor. Het doel van de studie was om te onderzoeken of de selectieprocedure die ontwikkeld is in Maastricht ook daadwerkelijk voorspellend is voor het studiesucces in de geneeskundebachelor. Om dit doel te kunnen bereiken, hebben we de relatie tussen de prestatie in de selectieprocedure en het latere studiesucces in de drie jaar durende bachelor onderzocht in drie studentencohorten (2011, 2012 en 2013). Deze cohorten werden gekozen, omdat er in deze jaren twee parallelle routes waren om het geneeskundecurriculum binnen te komen: de decentrale (oftewel voor iedere individuele universiteit verschillende) selectieprocedure, of de centrale (nationale) gewogen loting. In dit onderzoek zijn twee groepen vergeleken: de studenten die geselecteerd werden in de lokale selectieprocedure (*selectie-positieve studenten*) en de studenten die afgewezen werden in diezelfde selectieprocedure maar alsnog het geneeskundecurriculum binnenkwamen via de nationale gewogen loting (*selectie-negatieve studenten*). Alle beoordelingen die studenten gedurende het bachelor curriculum kregen, werden meegenomen. Deze beoordelingen werden onderverdeeld in vier categorieën: cognitief, (inter)persoonlijk, een combinatie van deze twee, en algemene gegevens (uitval en studievertraging).

De bevindingen waren als volgt: de selectie-positieve studenten presteerden beter dan hun selectie-negatieve medestudenten gedurende het hele bachelor programma, op zowel cognitieve (bijvoorbeeld geschreven bloktoetsen), (inter)persoonlijke (bijvoorbeeld communicatie- en reflectie-vaardigheden) als gecombineerde toetsen (stationstoetsen). Voor de algemene uitkomstvariabelen werden geen significante effecten gevonden. Van de in totaal 30 uitkomstvariabelen deden de selectie-positieve studenten het significant beter op 11 uitkomsten. Vijftien andere, niet-significante groepsverschillen waren ook in het voordeel van de selectie-positieve studenten. Een algehele vergelijking van alle uitkomsten door middel van een '*sign test*' liet dan ook een significant verschil tussen de beide groepen zien (p<0.001), ondanks de vergelijkbare middelbare school gemiddelden van de beide groepen.

Al met al lijkt een selectieaanpak die gebruik maakt van '*constructive alignment*' het gebrek aan voorspellende waarde dat gevonden wordt bij andere, veelgebruikte selectietoetsen te kunnen opheffen: een selectieprocedure die helemaal in lijn is gebracht met de inhoud, toetsing en de eindtermen van de studie kan uitkomsten in de geneeskundebachelor voorspellen. Selectie-positieve studenten presteren significant beter dan hun selectie-negatieve medestudenten over een reeks van cognitieve, (inter)persoonlijke en gecombineerde uitkomsten gedurende de hele drie jaar durende geneeskunde bachelor.

In Hoofdstuk 3 wordt een vervolgstudie beschreven. In dit hoofdstuk werd onderzocht of de hierboven beschreven selectieprocedure ook voorspellend was voor de

prestaties van de studenten in de klinische masterfase van het geneeskunde curriculum. Op deze manier draagt ook deze studie bij aan het bewijs voor validiteit op basis van de relatie van de selectieprocedure met andere variabelen. Het doel van dit onderzoek was om de relatie tussen de prestaties in de selectieprocedure en tijdens de klinische jaren van het geneeskunde programma vast te stellen. Belangrijk hierbij is weer de context: de selectieprocedure, de inhoud van het curriculum en de toetsen tijdens de studie zijn allemaal in lijn met hetzelfde uitkomstenraamwerk, namelijk het Raamplan Artsopleiding 2009. De opzet van deze studie lijkt erg op de studie beschreven in hoofdstuk 2: twee groepen studenten werden vergeleken, namelijk de studenten die geselecteerd werden via de selectieprocedure (selectie-positieve studenten) versus de studenten die afgewezen werden in de selectieprocedure en het geneeskunde curriculum alsnog binnenkwamen via een nationale, op basis van middelbare school cijfers gewogen, loting procedure (de selectie-negatieve studenten).

Prestaties van beide groepen studenten op alle zeven in het raamplan beschreven competenties (Medisch deskundige, Communicator, Samenwerker, Organisator, Gezondheidsbevorderaar, Academicus en Beroepsbeoefenaar) werden vergeleken gedurende de klinische rotaties (co-schappen), voor ieder masterjaar. Deze prestaties werden bijgehouden in een elektronisch portfolio. In het portfolio verzamelen studenten zowel kwantitatieve informatie (bijvoorbeeld resultaten van kennistoetsen) als kwalitatieve feedback van verschillende beoordelaars (bijvoorbeeld medisch specialisten, medestudenten, verpleging, patiënten) op verschillende momenten voor verschillende taken. Op deze manier wordt een grote bron aan informatie opgebouwd. Aan het einde van het jaar checkt de examencommissie alle, door de student verzamelde, informatie in het portfolio en wordt een definitief, programmatisch oordeel voor iedere competentie gegeven: beneden verwachting, naar verwachting, of boven verwachting. Data van de drie cohorten die begonnen aan hun geneeskunde bachelor curriculum in 2011, 2012 of 2013 werden vergeleken, aangezien dit de enige cohorten zijn waarin zowel selectie-positieve als selectie-negatieve studenten toegelaten werden. De selectie-positieve studenten bleken significant beter te presteren dan de selectie-negatieve studenten in alle drie masterjaren, en de verschillen tussen deze twee groepen bleken toe te nemen in de tijd. In het eerste jaar deden de selectie-positieve studenten het significant beter in de rollen Communicator, Samenwerker en Beroepsbeoefenaar, in het tweede en derde jaar hiernaast ook nog in de rol van Organisator en Gezondheidsbevorderaar, en in het derde jaar bleek dat hier bovenop ook te gelden voor de rol van Academicus.

De conclusie van deze studie was dat een selectieprocedure die '*constructively aligned*' is met het geneeskundecurriculum en het einddoel van de studie, een toenemende voorspellende waarde heeft in de klinische fase van het geneeskundecurriculum. Dit suggereert dat *constructive alignment* van selectie, curriculum en toetsing met de einddoelen van de opleiding een effectieve manier kan zijn om een selectie 'aan de poort' op te zetten die voorspellend is voor de klinische prestaties van studenten aan het einde van hun geneeskunde studie.

Hoofdstuk 4 richtte zich in meer detail op de inhoud en interne structuur van de selectieprocedure. Al een hele tijd bevindt het veld van selectie voor geneeskunde zich in de paradoxale situatie waarin we weten dat een aantal selectietoetsen bepaalde studie-uitkomsten voorspellen, maar welke constructen (onderwerpen/competenties) binnen de selectieprocedure nou eigenlijk zorgen voor die voorspellende waarde weten we niet (dit noemen we ook wel de '*black box*' in selectie). Daarom was deze studie specifiek gericht op het onderzoeken van welke constructen gemeten moesten worden, hoe die geïntegreerd konden worden in een procedure en in hoeverre de resulterende selectieprocedure erin slaagde om deze constructen daadwerkelijk te meten. Daarmee werd volgende vraag beantwoord: passen de psychometrische kwaliteiten van de toetsen in een '*constructively aligned*' selectieprocedure bij de inhoud die gemeten zou moeten worden? Om dit te kunnen onderzoeken werd de beoogde inhoud van de selectieprocedure uiteengezet, om vervolgens de psychometrische kwaliteiten te achterhalen en te onderzoeken of deze passen bij die beoogde inhoud.

Als eerste hebben we naar het inhoud-gerelateerde bewijs gekeken, met specifieke focus op de manier van opzetten en toepassen van de competentie-gebaseerde selectieblauwdruk. Zoals eerder benoemd werd *constructive alignment* bereikt door de eindtermen van de geneeskundeopleiding te raadplegen en te vertalen naar het niveau van een deelnemer aan de selectieprocedure. Hiervoor werd *backward chaining* toegepast. De resulterende selectieprocedure omvatte meerdere toetsen in twee rondes, waarvan de inhoud steeds gebaseerd was op de eindtermen. De focus van dit onderzoek was de tweede ronde, waarin een video-gebaseerde situationele beoordelingstoets (een Situational Judgement Test, gericht op het meten ven (inter)persoonlijke competenties) en een schriftelijke aanlegtoets (een Aptitude Test, waarin een breder spectrum van de eindtermen competenties aan bod komt) werden afgenomen. De opzet van de selectieprocedure bleek robuust en transparant, en het proces van het ontwikkelen van de inhoud was reproduceerbaar.

Als tweede werden de interne structuren van de selectietoetsen onderzocht door de prestaties van de deelnemers op de selectietoetsen te relateren aan de vooraf opgestelde blauwdruk van de toetsen (deze beschrijft welke items in de toetsen bedoeld zijn om welke competenties te meten) door middel van '*Cognitive Diagnostic Modelling*' (CDM). CDM is ertoe in staat latente variabelen te vinden, zelfs als er sprake is van multidimensionaliteit in de data, zowel binnen als tussen items. Dit was in onze data het geval: verschillende items meten verschillende competenties, maar de meeste items meten ook meerdere items. CDM is gerelateerd aan '*Confirmatory Factor Analysis*': de structuur (in dit geval de blauwdruk van de selectietoetsen) wordt ingevoerd in de analyse, en CDM bepaalt of diezelfde structuur inderdaad terug te vinden is in de data, of dat er aanpassingen nodig zijn in de structuur zodat deze beter past bij de data. Uit deze analyse bleek dat er sprake was van een overlap van 89% tussen te beoogde en daadwerkelijk gemeten constructen (competenties).

Alles bij elkaar genomen ondersteunen deze resultaten het idee dat de focus die we geplaatst hebben op het creëren van de juiste inhoud en het volgen van een

competentie-blauwdruk effectief was in termen van de interne structuur van de selectietoetsen: het overgrote deel van de items meet inderdaad wat ze zouden moeten meten. Deze manier van linken van een vooraf opgestelde blauwdruk aan de daadwerkelijke resultaten van de deelnemers werpt licht in de '*black box*' van selectie, en ondersteunt tevens de constructvaliditeit van de selectieprocedures.

Een andere onderbelichte pilaar binnen het validiteitsraamwerk dat in het begin van deze samenvatting genoemd werd, zijn de consequenties. Daarom lag de focus van het empirische onderzoek in Hoofdstuk 5 op een specifiek onderdeel van deze consequenties: de kosteneffectiviteit van selectie vanuit het oogpunt van de geneeskundeopleiding. Dit onderzoek is belangrijk omdat de middelen voor medisch onderwijs meer en meer beperkt worden, terwijl er tegelijkertijd steeds meer verantwoording over de besteding van deze middelen afgelegd moet worden. In deze beperkende omgeving moeten de geneeskundeopleidingen nadenken over hun selectieprocedures en moeten deze verdedigbaar zijn in termen van kosten en baten. Tot op heden zijn er echter nog geen studies die zich op dit aspect van selectie richten. Daarom was het doel van deze studie om de kosten en opbrengsten van twee verschillende benaderingen van toelating tot het geneeskundecurriculum te vergelijken: een '*contructively aligned*' selectieprocedure met meerdere toetsen versus een lotingsprocedure die gewogen is naar middelbare school gemiddelden. Ons doel was om de relatieve effectiviteit van de beide aanpakken te onderzoeken en te vergelijken in termen van kosten en opbrengsten vanuit het perspectief van de geneeskundeopleiding.

Deze studie is gedaan binnen het geneeskundecurriculum aan de UM, waar in de periode van 2011 tot 2013 een decentrale selectieprocedure en gewogen loting procedure parallel toegepast werden. De kosten en opbrengsten (over de gehele bachelor-periode) van de selectieprocedure werden vergeleken met die van de lotingprocedure voor drie studentencohorten (2011, 2012, en 2013).

De extra kosten voor selectie bestonden uit de financiële investering van de geneeskundeopleiding in het opzetten en uitvoeren van de selectie; de 'opbrengsten' werden afgeleid van de toename in inkomsten door uitval te voorkomen (voor een student die uitvalt in jaar 1 krijgt de geneeskundeopleiding 2 jaar geen inkomsten, voor uitval in jaar 2 krijgt de opleiding 1 jaar geen inkomsten) en de extra kosten door herhaling van blokken en stationstoetsen te verminderen. De organisatie van een '*contructively aligned*' selectieprocedure kostte ongeveer €139.000 voor een compleet cohort van 286 studenten. De kosten van de lotingsprocedure waren voor de geneeskundeopleiding verwaarloosbaar (deze kosten werden nationaal gedragen). De gemiddelde 'opbrengsten' van de selectieprocedure ten opzichte van de lotingprocedure bedroegen bijna €207.000 per cohort studenten. Dit resulteert in een voordelige balans van selectie ten opzichte van loting van ongeveer €68,000.

Concluderend laat deze studie niet alleen zien dat het mogelijk is om kosteneffectiviteitsstudies te doen binnen de context van selectie voor de

geneeskundeopleiding, maar ook dat een 'dure' selectieprocedure kosteneffectief kan zijn ten opzichte van een 'goedkoop' lotingssysteem.

Hoofdstuk 6 betreft een commentaar op een meta-analyse van Jonathan Foo en collega's. Zij tonen aan dat studies gericht op kosten heel zeldzaam zijn binnen het 'Health Professions Education' veld, en dat studies waarin ook de waarde (baten, opbrengsten) worden meegenomen nog minder voorkomen. Verder bleek de kwaliteit van de meerderheid van de beschikbare literatuur niet aan de maat te zijn, en bleek die kwaliteit ook niet toegenomen te zijn in de afgelopen decennia. Deze bevindingen spiegelen de situatie in de praktijk van het hoger onderwijs: beslissingen over investeringen in onderwijsactiviteiten worden vaak niet gedreven door volledige empirische en economische evaluaties. Een specifiek veld binnen het onderwijs waarin kostenevaluaties bijzonder interessant zijn is de selectie: geneeskundeopleidingen moeten hun unieke en vaak dure selectieprocedures verdedigen in termen van kosten en baten, terwijl onderzoek naar de kosten en vooral de baten van de selectieprocedures, of zelfs maar de toetsen binnen deze procedures, uitermate zeldzaam is.

Belangrijk is ook dat goede kosten-baten evaluaties niet alleen belangrijk zijn voor individuele instituten, maar ook op nationaal niveau. In Nederland, bijvoorbeeld, zijn alle geneeskundeopleidingen sinds 2017 verplicht om hun studenten te selecteren. De nationale gewogen loting werd afgeschaft, en iedere universiteit moest een eigen selectieprocedure ontwikkelen, die per universiteit verschillend is. Aangezien de voorspellende waarde van deze procedures wisselend en vaak beperkt bleek te zijn, werd de kosteneffectiviteit van selectie in het algemeen in twijfel getrokken en is er discussie ontstaan over eventuele herintroductie van de gewogen loting. Deze discussie bereikt zelfs de regering. Vanuit ons oogpunt zijn onderzoeken naar de algemene constructvaliditeit van selectieprocedures (in plaats van alleen onderzoek naar de voorspellende waarde), inclusief volledige economische evaluaties van hoge kwaliteit, cruciaal voor het ondersteunen van een parlementaire beslissing op dit punt.

Al met al is er dus een groot tekort aan economische evaluaties van hoge kwaliteit in het veld van het hoger onderwijs in het algemeen, en binnen het gebied van selectie in het bijzonder. Dit is problematisch voor de belanghebbenden, omdat zij aansprakelijk zijn voor het gebruik van de gelimiteerde financiële middelen, terwijl ze moeten kiezen tussen onderwijsactiviteiten zonder prijskaartjes. Meer samenwerking tussen onderwijskundige onderzoekers op het gebied van economische evaluaties is van belang om belangrijke beslissingen te kunnen baseren op economisch valide argumenten, hetgeen de kwaliteit van onderwijs ten goede zal komen.

In Hoofdstuk 7, de Algemene Discussie, zijn alle hoofdbevindingen uit de eerdere hoofdstukken samengevat en werden implicaties en suggesties voor de praktijk en het onderzoek rondom selectie bediscussieerd. Overkoepelend kunnen we zeggen dat de empirische studies in dit proefschrift bijdragen aan een validiteitsargument voor een selectieprocedure die gebaseerd is op het principe van '*constructive alignment*'. Alle

studies kunnen geïntegreerd worden in het validiteitsraamwerk dat in Hoofdstuk 1 uiteen werd gezet:

- ✓ *Inhoud*: een grondige analyse van welke competenties er nodig zijn, waarop een blauwdruk voor de toetsen gebaseerd wordt, resulteert in een robuust, transparant en herhaalbaar proces.
- ✓ *Reactieprocessen*: eerste, preliminaire data laat zien dat het variëren met wegingen in het portfolio geen effect heeft op leeftijd en geslacht van de geselecteerde studenten
- ✓ *Interne structuur*: er is ongeveer 90% overlap tussen de beoogde en daadwerkelijk gemeten competenties, met redelijke betrouwbaarheden per competentie en hoge intra- en inter-beoordelaar betrouwbaarheid
- ✓ *Relatie met andere variabelen*: de selectieprocedure is voorspellend voor preklinische prestaties tijdens de geneeskundebachelor (voor cognitieve, (inter)persoonlijke en gecombineerde uitkomsten) en is in toenemende mate voorspellend voor de dagelijkse klinische prestaties tijdens de masterfase
- ✓ *Consequenties*: een 'dure' selectieprocedure is uiteindelijk kosten-effectiever dan een 'goedkope' lotingprocedure op basis van middelbare school gemiddelden

Deze bevindingen hebben belangrijke repercussies voor een aantal (inter)nationale discussies: de cognitieve versus (inter)persoonlijke competenties-discussie zou geen zwart-wit keuze moeten zijn, maar selectie moet gefocust zijn op het meten van relevante competenties over het hele continuüm. Verder zorgen het blind vertrouwen op bewijs voor individuele toetsen/middelen en een gebrek aan focus op en validering van de inhoud voor inconsistente resultaten in onderzoek en in de praktijk. Het uitvoeren van een analyse welke competenties er nodig zijn als geneeskundestudent en dokter moet voorop staan bij het vaststellen van de inhoud die in de selectieprocedure gemeten moet worden. Ook de discussie over of de opbrengst van selectie eigenlijk wel de moeite die het kost waard is, is van belang: de resultaten verschillen behoorlijk tussen universiteiten, maar de focus in de discussie wordt vaak gelegd op de negatieve resultaten. Met dit proefschrift hopen we die neiging om vooral op negatieve resultaten te focussen te stoppen, zodat we niet het kind met het badwater weggooien. Er zou meer aandacht gevestigd moeten worden op het zorgvuldig plannen van de constructen en het grondig evalueren van de procedures, om zo te komen tot goed onderbouwde, kosteneffectieve en voorspellende selectieprocedures.

De implicaties van dit proefschrift voor theorie, onderzoek en praktijk zijn veelvoudig: het beginnen met het opzetten van een procedure om te analyseren welke competenties er nodig zijn in de studie en latere klinische praktijk; het gebruik van innovatieve methoden van onderzoek (bijvoorbeeld *Cognitive Diagnostic Modelling* en kostenevaluaties); en de toepassing van moderne validiteitstheorieën.

De voornaamste conclusie uit dit proefschrift is dat expliciete '*constructive alignment*' een goede weg vooruit is voor selectie. Daarnaast is het aanbevelenswaardig om

selectieprocedures te evalueren door middel van moderne validiteitsraamwerken, om zo een breder begrip van de validiteit van de procedures te verkrijgen. Een veelbelovende testtheorie die hieraan zou kunnen bijdragen is *Cognitive Diagnostic Modelling*. Een combinatie van deze drie aanbevelingen, '*constructive alignment*', moderne validiteitsraamwerken en innovatieve methoden (waaronder CDM), zou het veld van selectie kunnen helpen dichter bij die 'heilige graal' te komen: een voorspellende, betrouwbare, valide, eerlijke, acceptabele en kosteneffectieve selectieprocedure.

Valorization / Valorisatie

The importance and implications of the research described in this dissertation have been stated before, especially in the General Discussion. Selection is a controversial subject with heated discussions flaring up at many levels: personal, institutional, and governmental. To enable informed discussions and decisions, the stakeholders require high-quality research, especially on the "most fundamental consideration in developing tests and evaluating tests": validity. However, up to now, research on selection in terms of validity has been incomplete. Therefore, this dissertation focused on a more inclusive view on validity, taking into account the importance of the content, response processes, internal structure, relation to other variables and consequences of our selection procedure. In doing so, we procured information vital to stakeholder decisions related to selection.

The most important implications have been described in detail in Chapter 7, the General Discussion. These implications are manifold. Related to our selection procedure, we endorse previous calls to use job analyses as the starting point for a selection procedure. Related to our methodological and statistical procedures, we hope to have shown several innovative manners in which research on selection can be conducted, for example applying Cognitive Diagnostic Modelling, and conducting cost-effectiveness research. From a theoretical point of view, diving deep into the modern validity theories has been transformative in looking at selection: they make you more critical towards your own research and others' research as well as your own selection procedure, assessments, and curriculum. A practical implication was the affirmation of the procedure for medicine that was developed at Maastricht University, which was helpful in showing the university-stakeholders that the selection procedure was worth the hassle in terms of prediction as well as cost-effectiveness. Furthermore, the application of CDM was very helpful in practice, as it helped the selection committee understand what they were actually asking of their applicants. This method may have additional future applications as well, such as exporting an overview showing on which competencies each applicant is above or below average to help students use the selection procedure as a learning moment. Last, the application of modern validity frameworks was useful in practice; it stimulated more critical reasoning about selection and provided a 'roadmap' on how to create procedures for which validity evidence will actually be supportive.

In order to refrain from repetition, and in order to pursue more impact on the Dutch 'selection versus lottery' discussion specifically, the remainder of this valorization is written in Dutch, as a plea to the different levels of stakeholders (personal, institutional, and governmental).

In dit proefschrift is de validiteit van een nieuwe aanpak van selectie onderzocht: het expliciet toepassen van '*constructive alignment*', beginnend bij de selectie, door het curriculum en de toetsing heen, tot in de eindtermen van de opleiding. Hierbij is gebruik gemaakt van een zogenaamde 'job analysis', een overzicht van alle

competenties die artsen nodig hebben om hun beroep goed te kunnen uitoefenen. Voor dit doel zijn de nationale, Nederlandse, eindtermen gebruikt: het Raamplan Artsopleiding 2009. Door de competenties en eindtermen beschreven in dit raamplan goed door te nemen en te vertalen naar het niveau van een geneeskunde selectie-kandidaat van 18-19 jaar (zonder ervaringen in de gezondheidszorg of met Probleem Gestuurd Onderwijs [PGO]), kon een op eindtermen gebaseerde selectieprocedure ontwikkeld worden. De keuze van de toetsen/middelen om de competenties te meten stond in dienst van die competenties in plaats van andersom, zoals tot nu toe vaak gebruikelijk was.

Om deze nieuwe aanpak uitgebreid te kunnen evalueren zijn in dit proefschrift niet alleen de 'standaard' evaluaties van selectie uitgevoerd, gericht op voorspellende en incrementele waarde, maar was er sprake van een veel bredere focus op validiteit. Dit betekent niet dat de voorspellende waarde minder belangrijk is, dat is immers het hoofddoel van selectie. Het laat enkel zien dat er meer belangrijk is dan alleen voorspelling van studiesucces. Een voorbeeld: als het strikken van veters voorspellend zou zijn voor prestaties in de geneeskundeopleiding, ook al weten we niet hoe of waarom, zou het dan ethisch zijn om daarop te selecteren? Tot op heden wordt er gebruik gemaakt van toetsen waarvan niet echt onderzocht is of zij ook daadwekelijk de competenties meten die van belang zijn. In andere woorden: een toets die samenwerking zou moeten meten, maar waarvan niet gecontroleerd wordt of dit daadwerkelijk het geval is, zou -bij wijze van spreken- ook kunnen meten of iemand graag sport. Daarom is die bredere visie op validiteit binnen selectie belangrijk: om een eerlijke, betrouwbare en valide selectieprocedure op te stellen moet niet alleen rekening worden gehouden met de voorspellende waarde, maar ook met de inhoud, reactieprocessen van de kandidaten, interne structuren en de consequenties van de procedures.

Het onderzoek dat in het kader van dit proefschrift is uitgevoerd heeft tot een aantal belangrijke conclusies geleid. Als eerste: '*constructive alignment'* van selectie, curriculum, toetsing en eindtermen leidt tot een selectieprocedure die voorspellend is voor de behaalde resultaten in zowel de pre-klinische bachelorfase als de klinische masterfase van de opleiding geneeskunde. Verder is het expliciet en grondig vaststellen van de competenties die gemeten dienen te worden een onmisbare stap in de creatie van de selectieprocedure. Dit leidt tot interne structuren van selectietoetsen die goed passen bij deze competenties. In andere woorden: door een goede blauwdruk van de inhoud van de selectieprocedure te maken, kunnen items ontwikkeld worden die daadwerkelijk meten wat ze zouden moeten meten. Ook rondom de consequenties van de selectieprocedure werden belangrijke resultaten gevonden: een 'dure' selectieprocedure kan kosteneffectief zijn ten opzichte van een 'goedkope' lotingsprocedure, waardoor er onder de streep meer geld over blijft om te investeren in kwalitatief goed onderwijs.

Alle bovengenoemde informatie, zowel in deze valorisatie paragraaf als in de rest van het proefschrift, brengt ons tot de volgende aanbevelingen:

1) Selectie voor de opleiding geneeskunde heeft heel veel potentie, en verdient de kans doorontwikkeld te worden. Teruggaan naar een lotingprocedure, terwijl er nog relatief weinig onderzoek is naar wat wel werkt in selectie en wat niet, en vooral waarom dit al dan niet werkt, zou betekenen dat we het metaforische kind met het badwater weggooien. Een selectieprocedure kan zowel voorspellend als kosteneffectief zijn, en daarnaast ook meten wat er beoogd wordt te meten.

2) Om belanghebbenden te ondersteunen in het maken van weloverwogen beslissingen is er meer nodig dan alleen informatie over de voorspellende waarde van verschillende, individuele selectieprocedures. Denk hierbij bijvoorbeeld aan de kosten en baten van de procedures, de eerlijkheid, betrouwbaarheid en validiteit.

3) Het toepassen van moderne validiteitstheorieën in het onderzoek naar selectie en de dagelijkse praktijk van selectie kan positieve effecten hebben op de verdedigbaarheid, eerlijkheid, betrouwbaarheid en validiteit van de selectieprocedures. De moderne validiteitstheorieën stippelen als het ware een route uit die gevolgd kan worden om tot een selectieprocedure te komen. Door de volgorde te volgen van (1) het grondig uitzoeken van de inhoud, (2) het onderzoeken van de reactieprocessen waarvan de deelnemers gebruik maken, (3) het analyseren van de interne structuren en bekijken of deze daadwerkelijk passen bij de beoogde inhoud, (4) het vaststellen van de relaties met andere variabelen en (5) het onderzoeken van de consequenties die samengaan met het toepassen van de selectieprocedure, zal de constructvaliditeit van selectieprocedures verbeteren.

4) Innovatieve methoden van onderzoek zijn nuttig in de context van selectie; zij leveren een duidelijke meerwaarde. Selectie is complex, en complexe vraagstukken worden niet opgelost met simpele antwoorden, en dus ook niet met simpele test theorieën. Het vaststellen van de interne structuur van een selectietoets, bijvoorbeeld, kan niet met een simpele (of complexe) factoranalyse uitgevoerd worden, en zelfs niet met '*Item Response Theory*'. Omdat de verschillende items verschillende competenties meten, maar ook binnen een enkel item meerdere competenties worden gemeten, zijn deze theorieën niet toereikend. Het kunstmatig uit elkaar trekken van competenties die in sommige situaties erg aan elkaar gerelateerd zijn komt de validiteit niet ten goede, en dus moet de testtheorie om kunnen gaan met multidimensionaliteit binnen en tussen items. Ook het berekenen van de kosten en baten was nieuw binnen het veld van selectie, maar levert resultaten op van groot belang. Deze creativiteit en innovativiteit zijn nodig om met het complexe onderwerp van selectie om te gaan.

In conclusie: selectie heeft veel potentie, en hoewel het complexe vraagstukken met zich meebrengt, is er vooral ook nog heel veel te winnen in dit veld. Door selectieprocedures te baseren op de eindtermen van de opleiding, de principes van '*constructive alignment*' toe te passen, het onderzoek te verbreden en innovatieve theorieën en methoden toe te passen, kan de aanpak van selectie uiteindelijk resulteren in een voorspellende, betrouwbare, valide, eerlijke, acceptabele en kosteneffectieve selectie. Maar daarvoor moeten wij, als selectiecommissies, selectieonderzoekers, programma- en onderwijsdirecteuren, overheid en samenleving, wel de kans grijpen selectieprocedures verder te ontwikkelen, van fouten te leren en nieuwe inzichten toe te passen.

Toolbox with exemplary manners in which to support the sources of evidence for validity

Evidence for	*What?*	*How?*
Purpose	The reason why the assessment needs to be conducted in the first place	Explain why the assessment is necessary (e.g. to guarantee learning in students, to prevent graduating incompetent doctors, etc.)
	The knowledge, skills and attitudes to be assessed	Conduct a job analysis (see chapter one), gather support from the literature and use theories on the construct to be measured
	The interpretations of and decisions based on the scores	Explain how results from the measurement are to be interpreted and how high-stakes the decision is that is to be based on the measurement
Content	A blueprint establishing the link between the test domain and real performance	Explain the logical, theoretical and/or empirical relationships between the items, (sub)scales and the domain to be measured
	Representativeness	Determine the relevance and authenticity of all items for the domain and target population (whether stakeholders feel like the measurement is indeed a measurement of the domain to be measures)
	Coverage of domain	Are all relevant parts of the domain (as defined in the purpose) represented in a large enough sample of items
	Quality of items & writers	Who the writers of the items were and how and why they were chosen and trained. General item demands to keep in mind: comprehensibility, relevance, acceptability, feasibility and completeness
Response process	Scoring models	Explain the theory behind the construct-guided construction of scoring criteria and rubrics
	Summarizing and weighting of scores	Rationale behind and procedure for combining related elements and separating unrelated items as well as for the weighting of different items
	Correspondence between the item and the response options	The observation format or type of measurement should logically align with how the construct can be expected to be measured most validly, based on the definition of the construct (e.g. written versus computer) Employ process measures (e.g. eye-tracking) that support the relation between items and responses as was expected based on the definition of the construct
	Quality control	Manner in which (electronic) scanning and/or scoring takes place, how scores are calculated and reported Ensuring the accuracy of the scoring keys and getting rid of poorly performing items by assessing preliminary scores
	Data security	Documentation of quality-control procedures

Internal structure	Reliability	Test reliability (for the test as a whole, and/or established subscales): * Internal consistency; e.g. Cronbach's alfa with a unidimensional construct, and Omega in case of multidimensionality * Test-retest, split-half, and parallel forms reliability * Rater reliability: Inter-rater and intra-rater reliability
	Item functioning	Explore items with special attention item distributions, variation, difficulty and discrimination, and floor and ceiling effects; explore the item importance by looking at factor loadings and including expert opinions and the theoretical value
	Fairness	Review the items for cultural sensitivity, later check for Differential Item Functioning (DIF), conduct subgroup analyses to check for inexplicable differences between groups
	Generalizability	Generalizability and/or decision study
	Psychometric model	Determine the dimensionality, check scale correlations and homogeneity, then compare this model to the expected model (the blueprint that was created beforehand). Examples of possible analyses are: * Factor Analysis, either exploratory (if there is no theory on the construct) or confirmatory (to test your blueprint); your items must have continuous scores, be unidimensional and normally distributed. * Item Response Theory, related to factor analysis, but the items can be nominal or even ordinal (latent variable remains continuous); gives information on the difficulty of each item, and depending on which model is used, also on the item's discrimination * Cognitive Diagnostic Modeling, a confirmatory model in which multidimensionality within items can be modeled, and a predetermined blueprint can be compared to the data
Relation to other variables	Concurrent and/or predictive and convergent and/or divergent measures, correlation patterns with these measures	Check whether measures that are supposed to measure the same construct indeed correlate positively with the measurement and whether measures that are supposed to measure a different construct indeed do not correlate with the measurement, either at the same time or at any time in the future

Consequences	Intended and unintended, positive and negative, short-term and long-term consequences	An account of the impact the test scores/results and the decisions based on them have on each of the stakeholders involved in the process (e.g. medical students, the university); In how far the positive consequences outweigh the negative consequences; An important long-term consideration is the costs, cost-benefit analyses and cost-effectiveness analyses comparing multiple alternatives
	Pass/Fail standard	Reasonableness of the method that was used to establish the cut score; Investigate the agreement of experts with the final decisions about applicants/students; Accuracy with which the pass-fail decision is applied
	Implications of failing the measurement	What happens when an applicant/student fails the measurement, and is this a reasonable result? The more implications a measurement has, the more support should be gathered to support the construct validity
	False positives / negatives	Frequency of occurrence

References to Appendix A

1. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, United States of America: American Educational Research Association; 2014.
2. De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide: Cambridge University Press; 2011.
3. Royal KD. Four tenets of modern validity theory for medical education assessment and evaluation. Adv Med Educ Pract. 2017;8:567-70.
4. Messick S. Validity of Psychological-Assessment - Validation of Inferences from Persons Responses and Performances as Scientific Inquiry into Score Meaning. American Psychologist. 1995;50(9):741-9.
5. Sweet RM, Hananel D, Lawrenz F. A unified approach to validation, reliability, and education study design for surgical technical skills training. Arch Surg. 2010;145(2):197-201.
6. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 2015;49(6):560-75.
7. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-7.
8. Tavakol M, Dennick R. The foundations of measurement and assessment in medical education. Med Teach. 2017;39(10):1010-5.
9. Kane MT. An Argument-Based Approach to Validity. Psychological Bulletin. 1992;112(3):527-35.
10. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19(4):539-49.

Dankwoord / Acknowledgements

First and foremost, I would like to thank my team. Mirjam, Kitty and Jennifer, you have been the best team I could have imagined, and I could not have done this without you. Mirjam, bedankt dat je altijd tijd voor me maakte, ook al had je die eigenlijk niet. Bedankt voor je -uitermate nodige- pragmatische blik als ik weer eens verdwaalde in alle mogelijkheden of verzeilde in veel te lange uitleg. Bedankt voor het feit dat je me een betere wetenschapper en schrijver hebt gemaakt. Kitty, bedankt voor het feit dat ik altijd bij je kon aankloppen en voor al je steun, op zowel professioneel als persoonlijk vlak. Bedankt voor je voorbeeld in wat doorzetten is, en het vertrouwen dat je me gaf in het onderzoek en in mezelf. Bedankt voor alle discussies waardoor ik zowel leerde om voor mijn standpunt op te komen als om mijn punten beter te maken, en waarbij we het uiteindelijk stiekem toch altijd wel eens waren. Jennifer, thank you for your great ideas for our research, some of these studies would not have existed without you. Thank you for your swift responses with a critical eye for the methods, but also for language. Without you, this thesis would likely have been even longer. Thank you for your enthusiasm, your openness and your examples.
Thank you all for having been part of my team, and sticking with me through these four years (and then some).

DECSEL Geneeskunde (voor de mensen die deze afkorting niet kennen: de selectiecommissie), bedankt voor 4 jaar aan fantastische vergaderingen, leerzame momenten, en heerlijke etentjes. De open sfeer in deze groep is een inspiratie. Dymphie, zonder jou had ik er niet gezeten. Heel erg bedankt voor je begeleiding in mijn masterstage, voor je aanbeveling bij Kitty, en voor je niet aflatende positiviteit! Je bent een voorbeeld voor me geweest sinds ik je ontmoette op de geheugenpoli. Kitty, binnen de context van DECSEL wil ik je ook nog een keer bedanken, voor de manier waarop je deze club geleid hebt. Zonder jou waren de vergaderingen minder gezellig geweest, en zeker ook minder productief. Ben, jouw interesse in mijn onderzoek en mij als persoon was ontzettend motiverend. Je keek altijd mee voor artikelen die interessant waren en mogelijkheden om mijn onderzoek te laten zien, en dat waardeer ik ontzettend. Frans, hoe moet ik jou omschrijven? Je bent zonder twijfel de grappigste hoogleraar die ik ooit heb ontmoet, en ik verwacht dat dat niet meer gaat veranderen. Je weet altijd precies wat je moet zeggen, en voelt feilloos aan hoe ver je kunt gaan. Daarnaast ben je ook nog eens ontzettend scherp en kritisch, en weet je altijd de vinger op de zere plek te leggen, maar dan op de een of andere manier zonder dat dat daadwerkelijk pijn doet. Ik kijk uit naar je vraag! Ik ben al veel te veel aan het schrijven, maar de rest van de commissie, Roy, Marian, Martijn, Laura, Ester, Nikki, Neslihan, Ilse, Fabienne, Juanita en Noud, heel erg bedankt voor alles.

Dan al mijn collega-PhD's, collega's en oud-collega's bij SHE. Ik kan niet goed uitdrukken waar ik jullie allemaal dankbaar voor ben, dus bedankt voor alles. Al mijn oud kamergenoten, Ellen, Jorrick, Lorette, Koos, Andrea, Carolin, Sanne, Guusje, Jolien, Derk, Georg en Fatemeh, en alle andere PhD's, ontzettend bedankt voor alle

gezelligheid, steun, hulp, feedback, en heel veel afleiding! Ik zou voor jullie allemaal een heel stuk kunnen schrijven, want zonder jullie had ik het misschien niet eens volgehouden, maar ik hou het bij een ontzettend grote 'Thank you'! Jullie zijn de beste collega's die ik me had kunnen wensen.

Arno en Carlos, jullie gigantische expertise wordt ontzettend gewaardeerd. Arno, jouw statistische hulp zal ik nooit vergeten, je wist het altijd duidelijk te maken en werd nooit moe van al mijn vragen. Heel erg bedankt voor alles. Carlos, zonder jou was deze thesis methodologisch een stuk minder spannend geweest! Bedankt dat je me het vertrouwen gaf om met R te gaan werken, en bedankt dat je allerlei nieuwe dingen met me wilde uitzoeken en het CDM-avontuur met me aan wilde gaan.

Het liefste secretariaat ooit: zonder jullie was mijn PhD-leven een stuk moeilijker geweest! Lilian, onze mama van de afdeling, bedankt dat ik altijd bij je terecht kon, voor vragen (ook domme vragen), voor hulp met alles dat los en vast zit, en voor knuffels! Nicky, Ryan, Audrey en Hennie, bedankt voor al jullie hulp en ideeën, en voor alle gezelligheid! En onze IT-mannen, Maurice en Ruud, bedankt voor alle ondersteuning, die ik vrij vaak nodig had.

Iedereen bij DocProf, heel erg bedankt voor alle kansen die jullie me gegeven hebben! Herma, mee mogen draaien met de tutortrainingen gaf me wat broodnodige afwisseling van het schrijven, bedankt voor je vertrouwen. Maarten en Lianne, jullie zijn fantastische mensen om mee samen te werken, het is altijd leuk met jullie, en ik kan voor alles bij jullie terecht. Wilma, bedankt dat je me hebt willen adopteren op je kamer! Pascal, bedankt voor je inspiratie voor mijn discussie, zonder jou had ik misschien nu nog vastgezeten. Ruth, Sophie, Cintha en Anne, ontzettend bedankt voor jullie ondersteuning bij alle logistieke uitdagingen, en voor alle gezelligheid! Heel fijn dat jullie deur altijd openstaat!

En het Study Smart team, bedankt voor de nieuwe uitdaging na mijn PhD! Anique, ik ben je ontzettend dankbaar dat je aan me dacht en zo veel moeite hebt gedaan om me toe te voegen aan het Study Smart team! En ontzettend bedankt dat je al die jaren voor ons als PhD's opkwam als PhD-coördinator. Je bent een fantastisch voorbeeld. Felicitas, het is heerlijk om met jou samen te werken. Je enthousiasme voor je PhD en het Study Smart project zijn aanstekelijk. Ik ben heel blij dat jullie me deze kans gegeven hebben, het is een geweldig project met een ontzettend leuk team, en ik kijk ernaar uit om te zien wat we hiermee allemaal kunnen bereiken!

Mijn paranimfen, Margo en Max. Hoe anders zijn jullie, en wat hou ik van jullie allebei! Margo, met jou kan ik alles delen, of ik nou wil klagen over nutteloze dingen, ergens mee zit, gewoon heel vrolijk ben, of kleren aan het shoppen ben, jij weet altijd goed te reageren. De afgelopen vier jaar hebben we heel wat meegemaakt, en mijn leven zou niet hetzelfde (en lang niet zo leuk) zijn geweest zonder jou! Max, ook bij jou kan ik altijd alles kwijt, maar vooral als ik even een monoloog moet houden over vervelende dingen/gebeurtenissen, weet ik dat ik bij jou moet zijn! Je bent een inspiratie als

vriend, hoe je altijd voor iedereen klaar staat en hoe ontzettend loyaal je bent. Gelukkig mag ik je er dan af en toe aan herinneren dat je soms ook aan jezelf moet denken (hypocrisie is geen drogreden).

Daphne, bescheiden als je bent vind je helemaal niet dat je thuishoort in een dankwoord van een boekje waar je 'toch niks aan gedaan hebt'. Ik denk daar anders over. Mijn beste vriendin die het al het langst met me uithoudt, wat ben ik blij dat al mijn hele leven blind op jou mag en kan rekenen. Melanie, the best German friend anyone could ever have. I am so proud at you, and so happy to call you my friend. You don't know half how special you are, and how much you deserve in life. But I am confident that you will, and I love being there to see it: *I'm gonna stand by you*. Yvonne, de allerliefste blije doos! Onze gedeelde passie voor wijn, stickers en dansen heeft ons heel wat memorabele avonden gebracht, en ik ben blij die met jou gedeeld te hebben. Op naar nog heel veel meer wijn, dansjes en weekendjes! Dunja, lekker ding van me, dankjewel dat je me voorgegaan bent naar het 'Noorden', en me hebt laten zien dat het best wel mogelijk is om dan lekker Limburgs te blijven en je vrienden te blijven zien. Maar goed dat Gelderland ook sushi heeft! Mark, jij houdt het stiekem ook al een hele tijd met me uit! Bedankt voor alle duizenden SMSjes (en later ook appjes), voor de goede gesprekken, etentjes en vooral voor de cocktails in Barcelona. Sanne, in die vier jaar PhD ben je een ontzettend groot deel van mijn leven geweest, soms bijna 24 uur per dag mijn steun en toeverlaat. Jij hebt me laten inzien dat reizen buiten Europa eigenlijk ook best wel leuk is! Ik ben heel blij dat we zo veel gedeeld hebben, en dat jij je plekje gevonden hebt. Ellen, jij bent zo veel meer dan een oud-collega. Mijn wetenschappelijke mama, degene die me de kans gaf om meer te kunnen dan ik zelf dacht, degene die me door moeilijke tijden heen gesleept heeft, en me altijd weer wist te motiveren. Je bent met vlag en wimpel geslaagd voor je cursus emoties! En ook Kevin en Finn, bedankt dat jullie me praktisch opnemen in jullie gezin!

Mijn familie, die veel te groot is om iedereen los te kunnen bedanken. Ik hou van jullie allemaal, en ik ben met jullie allemaal heel erg blij! Lieke, mijn grote zus, mijn grote voorbeeld. Dankjewel voor je 'voeten op de grond'-instelling, waar ik nog steeds van kan leren. Je hebt me altijd geleerd dat je niet moet zeuren, maar gewoon door moet gaan, en dat heeft me ontzettend geholpen tijdens mijn PhD. En dankjewel voor dit boekje, ook niet onbelangrijk natuurlijk. Daarnaast moet ik Vitor en jou natuurlijk ook bedanken voor Tigo en Mirte! Wat een heerlijke (eigenwijze, energieke) kinderen hebben jullie, en wat word ik vrolijk van ze. Ook al heeft vooral Mirte het zo nu en dan wel nog wat spannend gemaakt, ik zou het niet anders willen. Jullie geven me het vertrouwen dat kinderen toch wel leuk zijn. Yvo, bruder, jij maakt het me niet altijd makkelijk, maar ik hou van je met heel mijn hart. Je bent mijn broer, en ik zal altijd voor je klaarstaan en van je houden. Anke, big sis, soms zijn wij veel te veel hetzelfde, en dat is niet altijd goed voor de omgeving. Wat kan ik met je lachen, maar ik kan ook altijd hele serieuze dingen bij je kwijt! Michiel, zonder jou was ik tijdens mijn PhD nooit bereikbaar geweest. Dankjewel voor alle elektronische ondersteuning, en voor de etentjes! Lieve Dily, ik ben braaf gebleven, en ik heb zelfs netjes een promotie afgerond. En Happa, je hebt alleen het begin van mijn PhD mogen meemaken, maar

wat was je trots. Je hebt me geïnspireerd dit project te beginnen en met de gedachte aan jou is het nu dan afgelopen: deze is voor jou!

Jorian, je bent pas ongeveer halverwege mijn PhD ingevallen, maar gelukkig heb je volledig kunnen genieten van de ergste stress en crises. Je weet mij altijd weer rustig te krijgen en hebt me altijd alle mogelijkheden gegeven om in alle rust aan mijn PhD te werken, vooral als dat totaal niet uitkwam (bijvoorbeeld als we moesten verhuizen, schoonmaken, koken, of een tuin planten). Je geeft me meer zelfvertrouwen dan ik ooit gehad heb. 'Just seeing you makes me happy'. Ik hou van je! En ook jouw familie, die mij als 'rare Limbo' toch maar (grotendeels) geaccepteerd heeft: heel erg bedankt voor alle avondjes, etentjes en spelletjes!

En als laatste, mijn eeuwige steun en toeverlaat, pap en mam! Pap, jij hebt me geleerd hoe je omgaat met moeilijke situaties. Als het echt lastig wordt, word jij de rust zelve en heb je al je prioriteiten volledig op orde: familie voor alles. Vrij snel daarna wel werk, natuurlijk. Je bent er altijd voor ons, en we weten allemaal hoe veel je eigenlijk van ons houdt. Zoals de gevleugelde uitspraak van Happa: 'was sich liebt, das neckt sich', en pap, ik hou van je! En mama, ik kan de woorden niet eens vinden om jou te bedanken. Jij hebt me gemaakt wie ik ben, laat me zien wie ik wil zijn en moedigt me vooral aan om altijd bij mezelf te blijven. Je hebt me altijd gesteund in alles wat ik gedaan heb en wilde doen, zelfs als jij het zelf moeilijk had, en ik kan je op geen enkele manier genoeg bedanken. Ik ben er trots op om jou mijn moeder te kunnen noemen, en ik blijf bij de overtuigingen die ik van kinds af aan al had: ik heb de allerliefste mama van de hele wereld. Ik hou ontzettend veel van je, en ik ben blij dit met jou te mogen delen.

About the author

Sanne Schreurs was born in Roermond on July 8, 1992. She received her Bachelor's degree in Psychology at Maastricht University in 2013. She received her Master's degree (cum laude) in Neuropsychology from Maastricht University in 2015. She worked as a research assistant at Maastricht University Medical Center in the memory clinic, before starting her PhD project in July 2015. This PhD project was conducted at the School of Health Professions Education at Maastricht University. Currently, she is a teacher with the department of Faculty Development and project manager for the Study Smart project.

Academic work

Articles

Schreurs, S., Cleutjens, K. B., Muijtjens, A. M., Cleland, J., & Oude Egbrink, M. G. (2018). Selection into medicine: the predictive validity of an outcome-based procedure. *BMC Medical Education*, *18*(1), 214.

Schreurs, S., Cleland, J., Muijtjens, A. M., Oude Egbrink, M. G., & Cleutjens, K. (2018). Does selection pay off? A cost–benefit comparison of medical school selection and lottery systems. *Medical Education*, *52*(12), 1240-1248.

Schreurs, S., Cleutjens, K., Collares, C. F., Cleland, J., & Oude Egbrink, M. G. (2019). Opening the black box of selection. *Advances in Health Sciences Education*, 1-20.

Schreurs S, Cleutjens K, Cleland J, oude Egbrink MGA. Outcome-based selection can predict performance in the clinical years of medical school: The proof is in the pudding. Academic Medicine. 2019; accepted for publication.

Schreurs S, Cleutjens K, Oude Egbrink MGA. (2019). Increasing value in research: Cost evaluations in health professions education. *Medical Education*; *53*(12):1171-1173.

Oral presentations

What	*Where*	*When*	*Subject*
SHE presents	School of Health Professions Education, Maastricht	March 1st 2016	Complete project
Conference	AMEE, Helsinki	August 29th 2017	Chapter 2
Broodje verstand	HAG, Maastricht	November 2nd 2017	Chapter 2 + future plans
Conference	NVMO, Egmond aan zee	November 17th 2017	Chapter 2
Conference	AMEE, Basel	August 27th 2018	Chapter 5
Conference	NVMO, Egmond aan zee	November 15th 2018	Chapter 5
Conference	Scientific Research Community, Leuven	October 18th 2018	Chapter 2
Lunch lecture	School of Health Professions Education, Maastricht	February 12th 2019	Complete project
National Spring School	ICO, Amstelveen	March 14th 2019	Chapter 4
Conference	AMEE, Vienna	August 27th 2019	Chapter 4
Conference	NVMO, Rotterdam	November 22nd 2019	Chapter 3
Invited lecture	UCAT Consortium; London	December 6th, 2019	Overall results of the project

Other presentations

What	*Where*	*When*	*Subject*
Round table	ICO international fall school, Bad Schussenried	November 1st 2016	Weighting article
Poster presentation	SHE academy, Maastricht	March 27th 2017	Chapter 2
Round table	NVMO, Egmond aan zee	November 16th 2018	'Zelfselectie of selectie'
Podcast	Medical education	2018	Chapter 5

Completed courses

Course	*ECTS*	*Date*	*Place*
SHE writing course	12	October 2015 & March 2016	Maastricht
ICO introductory course	5	Spring 2016	Utrecht
Statistics 2	3	Summer 2016	Maastricht
ICO international fall school	3	Winter 2016	Bad Schussenried
Masterclass writing – Lorelei Lingard	3	Spring 2017	Maastricht
Multilevel Analysis of Longitudinal Data	3	Summer 2017	Maastricht
ICO Learning and instruction; basics and beyond	3	Summer 2018	Online
EBMA Psychometrics	3	Summer 2018	Maastricht
University Teaching Qualification	6	May – October 2018	Maastricht
ICO National Spring School	1	Spring 2019	Amstelveen
Total	42		

Further activities

Activity	*Context*	*Date*	*Subject*
Peer review	AHSE	2016	Statistics
		2019	Writing crafts
	Med Ed	2017	Widening access
		2018	Selection for medical school
Education	Health sciences	2017-18	Introductie Wetenschappelijk onderzoeksmethoden
		2019	Mentor
	Medicine	2015-19	Member of the selection committee
	Faculty development	2017-19	Tutor training / SGTT PBL principles training
Organization	SHE	2017	HAM dagen
		2017	SHE presents
		2019	SHE academy
Representative	EBMA conference	2017	ASME

SHE dissertation series

The SHE Dissertation Series publishes dissertations of PhD candidates from the School of Health Professions Education (SHE) who defended their PhD theses at Maastricht University. The most recent ones are listed below. For more information go to: https://she.mumc.maastrichtuniversity.nl

- ✓ Kikukawa, M. (17-12-2019) The situated nature of validity: Exploring the cultural dependency of evaluating clinical teachers in Japan
- ✓ Kelly, M. (10-12-2019) Body of knowledge. An interpretive inquiry into touch in medical education
- ✓ Klein, D. (06.11.2019) The performance of medical record review as an instrument for measuring and improving patient safety
- ✓ Bollen, J. (01.11.2019) Organ donation after euthanasia: medical, legal and ethical considerations
- ✓ Wagner-Menghin, M. (25-09-2019) Self-regulated learning of history-taking: looking for predictive cues
- ✓ Wilby, K. (02-07-2019) When numbers become words: Assessors' processing of performance data within OSCEs
- ✓ Szulewski, A. (20-06-2019) Through the eyes of the physician: Expertise development in resuscitation medicine
- ✓ McGill, D. (29-05-2019) Supervisor competence as an assessor of medical trainees
- ✓ Evaluating the validity and quality of supervisor assessments
- ✓ Van Rossum, T. (28-02-2019) Walking the tightrope of training and clinical service; The implementation of time variable medical training
- ✓ Amalba, A. (20-12-2018) Influences of problem-based learning combined with community-based education and service as an integral part of the undergraduate curriculum on specialty and rural workplace choices
- ✓ Melo, B. (12-12-2018) Simulation Design Matters; Improving Obstetrics Training Outcomes
- ✓ Olmos-Vega, F. (07-12-2018) Workplace Learning through Interaction: using socio-cultural theory to study residency training
- ✓ Chew, K. (06-12-2018) Evaluation of a metacognitive mnemonic to mitigate cognitive errors
- ✓ Sukhera, J. (29-11-2018) Bias in the Mirror. Exploring Implicit Bias in Health Professions Education
- ✓ Mogre, V. (07-11-2018) Nutrition care and its education: medical students' and doctors' perspectives
- ✓ Ramani, S. (31-10-2018) Swinging the pendulum from recipes to relationships: enhancing impact of feedback through transformation of institutional culture
- ✓ Winslade N. (23-10-2018) Community Pharmacists' quality-of-care metrics. A prescription for improvement

- ✓ Eppich, W. (10-10-2018) Learning through Talk: The Role of Discourse in Medical Education
- ✓ Wenrich, M. (12-09-2018) Guided Bedside Teaching for Early Learners: Benefits and Impact for Students and Clinical Teachers
- ✓ Marei, H. (07-09-2018) Application of Virtual Patients in Undergraduate Dental Education
- ✓ Waterval, D. (26-04-2018) Copy but not paste, an exploration of crossborder medical curriculum partnerships
- ✓ Smirnova, A. (04-04-2018) Unpacking quality in residency training and health care delivery