


Testing in Higher Education: Decisions on students' performance

Iris E. Yocarini



Testing in Higher Education:
Decisions on students' performance

Iris E. Yocarini

Colophon

Copyright original content © 2019 Iris E. Yocarini

All rights reserved. Neither this book nor any part may be reproduced or transmitted in any form of by any means, electronic or mechanical, including photocopying, micro-filming, and recording, or by any information storage and retrieval system, without prior written permission form the author.

Cover design: Iris E. Yocarini, background *Building Forest* 2017 by Minjung Kim (Gwangju, 1962)

Layout: Iris E. Yocarini

Printed by: Ridderprint BV, the Netherlands

ISBN: 978-94-6375-446-0

Testing in Higher Education: Decisions on students' performance

**Toetsen in het hoger onderwijs:
Beslissingen over de prestatie van studenten**

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
vrijdag 27 september 2019 om 13:30

door

Iris Eleni Yocarini

Geboren te Zeist

Promotiecommissie

Promotoren: Prof. dr. L.R. Arends

Prof. dr. G. Smeets

Copromotor: Dr. S. Bouwmeester

Overige leden: Prof. dr. C.J. Albers

Dr. I. Visser

Dr. P.P.J.L. Verkoeijen

Contents

Chapter 1	General Introduction	1
Chapter 2	Systematic Comparison of Decision Accuracy of Complex Compensatory Decision Rules Combining Multiple Tests in a Higher Education Context	11
Chapter 3	Allowing Course Compensation in Higher Education: A latent class regression to evaluate performance on a sequel course	47
Chapter 4	Correcting for Guessing in Estimating True Scores in Higher Education Tests	75
Chapter 5	Comparing the Validity of Different Cut-Score Methods for for Dutch Higher Education	107
Chapter 6	General Discussion	143
Summary		151
References		155
Samenvatting	Summary in Dutch	165
Curriculum Vitae		171
Dankwoord	Acknowledgements	175

1

General Introduction

Sara is a first-year Psychology student who just finished the exam of her final course. Mark is an associate professor of educational psychology and just finished teaching the final course in the first-year of the bachelor. He now has to decide which of his students, like Sara, passed the test and also his course. Carol is the head of the educational program and implemented an academic dismissal policy at the end of the first year of the bachelor. In this policy, she decides which first-year students, such as Sara, are allowed to continue their bachelor studies and which students are not. With this policy, Carol wishes to motivate students in the first year while at the same time she tries to ensure that students who do not meet the requirements and are not likely to obtain their diploma in the future, are dismissed.

In higher education curricula, tests are administered so that decisions about students' performance, such as those described in the example, can be made. As portrayed, different stakeholders make different decisions based on students' performance (e.g., decisions to pass or fail students and decisions to allow students to continue their studies). Although each stakeholder makes their decisions to the best of their ability, they have different objectives and available resources that may be in conflict with each other. For example, Carol needs to make a decision based on multiple tests to select students who are motivated and who have the right capacities. She only wants to allow students who truly meet all the requirements to continue their studies such that the educational quality of the study program is guaranteed. However, she understands that tests are not perfectly reliable and valid and that wrong decisions are inevitable. Whether her decisions are valid, such that students who are allowed to continue their bachelor study meet all the necessary study program requirements, depends on many aspects of which the quality of the individual tests is an important one. Although teachers like Mark aim to construct high quality tests, such that the test score estimates a student's underlying ability level well, he is constrained in his time and budget to design the test, which may limit the tests' quality.

To preserve the educational quality of the diploma of a study program, the decisions made about students' performance should be valid, such that students who receive the diploma meet the requirements to obtain the diploma. This is important for decisions made at each level. Valid decisions are made when the decision is accurate. In

psychometrics, students are assumed to have a certain underlying (that is, unobserved, latent) ability level, also referred to as a student's *true score*. By administering a test, a test administrator wishes to estimate this latent ability. This true ability level, or true score, is the test score we would obtain when the test would measure the true ability level perfectly. Notably, this true score of a specific person applies to a specific test at a specific moment in time, and would be stable across different administrations of the test under similar circumstances (i.e., if one assumes the student would start each repeated test administration with a clean slate, that is, *tabula rasa*). Unfortunately, the test score may not perfectly reflect a student's true ability level because random luck is of influence and may result in a test score that is higher (due to luck) or lower (due to bad luck) than the true score. The larger the degree of luck that is reflected in the test score, the larger the discrepancy between the latent true score and the observed test score, and the less *reliable* the test score is. When this is true, the decision based on the test is more likely to be inaccurate. Furthermore, for a test to result in valid decisions on students' performance it should measure what it intended to measure (i.e., the test itself should be *valid*).

Having higher education tests that do not measure a student's true ability level perfectly, in terms of both reliability and validity, two types of inaccurate decisions can be made. On the one hand, a decision based on the unobserved true score should be positive while the decision based on the observed test score shows to be negative. This is referred to as a false negative and would mean that we dismiss or fail a student based on his or her test score(s) while his or her underlying ability is actually sufficient. On the other hand, an inaccurate decision may occur when the decision based on the unobserved true score should be negative while the decision based on the observed test score is positive. This is referred to as a false positive, students who are not dismissed or pass a test while they are, based on their underlying ability, not truly sufficiently skilled yet. In this dissertation, the accuracy and consequences of decisions on students' performance in higher education are evaluated, both for decisions based on multiple tests (such as those made by Carol) and for decisions based on individual tests (such as those made by Mark).

Academic Dismissal Policy

In the Netherlands, among many other countries (e.g., in the USA, Germany, Finland, Australia, Ireland, Scotland, and Denmark as well; de Boer et al., 2015), higher education institutions obtain performance-based funds from the government.

Different types of performance-based funds exist, where funds may vary with an institution's past performance or are based on expected performance (through so-called performance agreements). In these performance agreements specific goals are agreed upon for a given time period which, if not met, may result in less funding for institutions. Important indicators in these performance goals are students' dropout rates after the first year and completion rates for bachelor students. As a consequence of these agreements, among other objectives, improving student success (that is, reducing dropout and increasing completion rates) has become a core focus in higher education institutions.

One way to boost student success is through the design of the testing system that is employed. Herein an academic dismissal (AD) policy may be implemented to dismiss students who do not meet certain criteria. Studies have shown that, although AD policies seem to particularly benefit teachers and institutions by retaining talented and motivated students who are likely to succeed, AD policies are beneficial to students as well. Students are more likely to succeed when an AD policy is in place through increasing their efforts when a dismissal is in sight or by switching to another (more suitable) study program in time (Cornelisz, Levels, van der Velden, de Wolf, & van Klaveren, 2018; De Koning et al., 2014). In the Dutch higher education, the AD policy that is in place is called the binding study advice (BSA), in which students who do not meet the required number of course credits obtained in their first year of the bachelor are dismissed. For its BSA requirements, the Erasmus University Rotterdam (EUR) decided to increase the number of course credits required to the maximum of 60 ECTS¹ in 2011 for the Psychology bachelor and later expanded this requirement to other study programs (Vermeulen et al., 2012).

¹ ECTS is a standardized grading system common in Europe and stands for the European Credit Transfer and Accumulation System

Increasing the BSA requirements to the maximum study credits sparked the media's attention, which was sparked again after the Dutch Minister of Education acclaimed her plans to lower the maximally allowed study credits required within the BSA to a maximum of 40 out of 60 ECTS (Rijksoverheid [Dutch government], 2018). In discussions on the BSA requirements the EUR is often mentioned as an example as it has the highest BSA requirements for most of its study programs. This discussion on the BSA requirements should however not solely focus on the 60 ECTS requirement, as there are other measures that were simultaneously implemented with the increased BSA requirements (for a detailed description thereof see Arnold & van den Brink, 2009; Vermeulen et al., 2012). Part of these additional measures, for example, were a cap on the number of tests students were allowed to retake and the use of a compensatory decision rule to calculate the number of course credits a student obtained in the first year. Together, these measures were an attempt to decrease student procrastination behavior and to increase student success through the adjustment of the institutional academic environment. In this dissertation, focus lies on the latter measure, allowing compensation between courses.

Traditionally, course credits are assigned to individual courses and students receive these credits when they obtain a passing course grade, which is when the student's test score is above the pass-fail test score (referred to as the cut-score). Assigning course credits in this way means that a so-called conjunctive decision rule is in place. Alternatively, in a compensatory decision rule, course credits are assigned based on a student's average grade (that is, the grade point average [GPA]). In this way, students are allowed to compensate a low score on one course with a high score on another, as long as their average grade meets the requirements. Noticeably, compensation in a higher education context, in which a certain minimum level of performance is expected from students, is usually allowed within certain boundaries. This is often referred to as a complex compensatory decision rule, where 'complex' refers to additional conjunctive requirements such as requiring each individual test score to be above a certain criterion in addition to the requirements for the average grade.

Compensatory Decision Rule

Whether compensation should be allowed or not depends on the context of the decision. In higher education, one could argue that compensation should be allowed only for those courses which are believed to be to a large extent interchangeable in the sense that students still meet the overall end qualification requirements of the study program. For example, in a Psychology bachelor program, first year courses might all be considered introductory courses covering the broad fundamentals of psychology and compensating one of these courses might not be considered problematic for later performance as a psychologist. However, in a Psychology master program where courses are highly specialized, focusing on a small area in psychology, compensation between courses would not be recommended as each course covers a fundamental aspect and students would need this knowledge for becoming a successful expert in this specialized field. Similarly, this logic applies to the formation of cluster of courses within which compensation is allowed. These clusters could be formed based on the courses' content or difficulty, resulting in courses that are believed to be interchangeable.

Overall, the discussion and decision to allow compensation is mostly a consideration of students scoring close to the cut-score instead of high performing students as they will likely pass regardless of the decision rule (Van Rijn, Béguin, & Verstralen, 2009). In this discussion, each stakeholder has their own view and opinion on allowing compensation between courses. Taking the view of Carol the policy maker, compensation may be favored as it may decrease the number of retakes, trying to encourage students to speed up their study progress and in this way increase students' study success. From the perspective of both Carol and Mark the teacher, compensation may be favored as it may encourage students to increase their effort on individual courses as it pays off to get a grade that is higher than the cut-score. Alternatively, Mark may be reluctant towards compensation as he believes students should not be able to pass his course with a low grade, viewing it as a devaluation of his course (Rekveld & Starren, 1994). Similarly, students such as Sara may be happy because she can compensate a low grade with a higher grade but may at the same time worry that

compensation may decrease the quality of an educational program and result in a devaluation of her diploma as well (Bakker, 2012; Cohen-Schotanus, 1995).

Regardless of the perspective one takes, what should be central in the discussion of whether to allow compensation is the *accuracy* of the decision that is made. One argument related to the accuracy that is often put forth by proponents of compensation is that the average grade is more reliable than individual course grades (Vermeulen et al., 2012). Whereas several studies evaluated the consequences of allowing compensation within a higher education curriculum (see e.g., Arnold & van den Brink, 2009; Cohen-Schotanus, 1995), most studies did not evaluate the accuracy of this decision rule. Where a few studies exist that evaluated the decision accuracy of different decision rules (e.g., Douglas & Mislevy, 2010; Hambleton & Slater, 1997; Van Rijn, Béguin, & Verstralen, 2009, 2012), none of these studies were placed in the context of higher education curricula. Studying the decision accuracy is difficult because an assessment of whether the decision based on the observed test scores is accurate requires the students' true ability level to be known. As mentioned, students' true ability is the test score we would obtain when the test would measure the true ability level perfectly. As tests and its administrations are not free of error, true scores remain unknown.

In **Chapter 2** the accuracy of a decision based on multiple tests (such as the BSA decision) in higher education is evaluated by performing a systematic comparison of the decision accuracy of different complex compensatory decision rules. In order to obtain students' true ability levels and to mimic different realistic higher education contexts, real-data-guided simulations are performed. By comparing different compensatory and conjunctive decision rules, one of the arguments for allowing compensation, that the average grade is more reliable than individual course grades, is evaluated as well. This is done within different realistic settings by varying the requirements in the complex compensatory decision rules as well as the characteristics of the testing system such as the correlation among tests, average test reliability, the number of tests, and the number of retakes allowed.

One of the criticisms of allowing compensation in a first-year higher education curriculum is that compensation might result in second-year students who have knowledge gaps for courses they were allowed to compensate in the first year. Specifically, this concern holds when knowledge is accumulated across courses, such that a sequel course builds on material from previous, so-called precursor, courses. By studying these course combinations, the consequences of allowing compensations with respect to hiatuses in knowledge can be evaluated. In **Chapter 3** an extension on previous studies in which the performance on sequel courses was evaluated, is made by evaluating the performance on sequel courses for different groups of students based on their unobserved (i.e., latent) study processes. A latent class regression analysis is applied to student data from a Psychology bachelor and a Law curriculum to identify students who show low performance on sequel courses, in which students' first-year average, variability in first-year grades, number of compensated courses, and number of retaken courses are used to form these latent classes.

Testing in Higher Education

Regardless of the specific testing system that is implemented (i.e., the decision rule for the combination of tests), the proportion of inaccurate decisions will be high if the quality of the individual tests, on which the decision is based, is not sufficient. However, ensuring the quality of individual tests in higher education is challenging due to the limited time and budget that is available to course instructors (such as Mark). Several studies have shown that the quality in instructor-constructed multiple choice tests in higher education indeed may be low (e.g., Brown & Abdulnabi, 2017; DiBattista & Kurzawa, 2011). In **Chapter 4** and **Chapter 5** therefore, different methods are evaluated to assess how true score estimation in individual higher education tests, and given their quality, could be improved.

In higher education, tests are administered to assess students' knowledge or skills on a specific topic. Although testing is known to support learning (e.g., Roediger & Karpicke, 2006) and might be directed towards learning, most tests offered in higher education are end-of-course tests in which the goal is to measure students' true ability on the course. This type of testing is commonly referred to as summative tests (Black & Wiliam, 2003). Although true score estimation on educational tests has been

studied extensively in the educational measurement literature, most tests studied in the literature are different from the type of tests found in higher education curricula in several ways. This makes it difficult to generalize results found in the literature to the tests in (Dutch) higher education which are studied in this dissertation.

Whereas the literature mostly focuses on large-scaled standardized tests, such as the Dutch end-of-primary school tests (e.g., CITO) and the college-entry Scholastic Aptitude Test (SAT) commonly used in the US, tests in Dutch higher education are not standardized. Consequently, most tests are designed in-house by individual course instructors. Different from standardized tests, these instructors are limited in their time and budget and therefore cannot pre-test their test items. This constrained time and budget also limits the use of panels to determine the cut-score in higher education, which is the most common method described in the literature, leaving this task to the instructor. Moreover, course instructors have not received formal training in designing and analyzing test items (Draaijer, 2016), making it difficult for them to safeguard the quality of the test. Still, even when trained psychometricians are available to analyze the test items, tests in higher education are often too small to obtain stable item and person parameters using item response theory (IRT) models which limits true score estimation in higher education tests. Given all these differences between the tests studied in the literature and those found in higher education, different challenges exist in students' true score estimation in higher education tests, making it a relevant subject of study.

Whereas many aspects determine whether a true score is estimated correctly, **Chapter 4** focusses on the accuracy of different methods to correct for guessing in higher education multiple choice (MC) tests. Specifically, MC tests in which students are not directly penalized for wrong answers (that is, a wrong answer does not result in deducted points) and consequently students' optimal strategy is to guess instead of omit answers, are investigated. Psychometrically, guessing is problematic for the estimation of a student's true score as we cannot be sure whether a correct answer is due to knowledge or a lucky guess (Bar-Hillel, Budescu, & Attali, 2005; Budescu & Bar-Hillel, 1993). Although there has been a recent shift towards not correcting for guessing in large-scaled tests such as the SAT (Guo, 2017), MC tests in higher

education are often corrected for guessing. Here, the total number of correct items is adjusted by subtracting a proportion of items assuming that test-takers would have randomly guessed among the given response options. Problematically, partial knowledge is not considered in this correction, possibly resulting in an overestimation of students' underlying true score. Other methods to correct for guessing exist, such as the extended classical correction, (extended) beta binomial correction methods, and models from IRT, that take sample information into account. The aim of the study in Chapter 4 is to evaluate these different methods that correct for guessing to see if students' true score estimation might be improved. Hereby, the accuracy of each method is compared for which a simulation study is performed. By varying several aspects of the higher education test context, performance within different realistic test settings is evaluated.

Often, after correcting for guessing on MC items, grades are assigned to test scores in higher education as an indication of students' underlying ability level. The process of transforming test scores into grades using certain rules is referred to as setting standards (Reckase, 2006). In higher education, this process is often simplified compared to panel methods as one instructor is responsible for setting the standard and consensus is easily reached in this way. Although simplified, the cut-score in Dutch higher education is often set at a prefixed percentage of items to answer correct without much consideration of the underlying ability level required for a passing grade. In **Chapter 5** the accuracy of different standard setting methods that are feasible in small-scaled non-standardized higher education tests is evaluated. In addition to the pre-fixed percentage method, which is an absolute method, two compromise methods were included which take students' performance into account as well (i.e., it has a relative component): the Cohen and Hofstee method. Again, simulations are performed to obtain students' true scores and assess the accuracy of estimated true scores across different methods. Also, through the use of simulations different type of tests and samples are evaluated.

Finally, **Chapter 6** provides an overall summary of the findings of chapter two to five as well as a discussion and conclusion on the implications for educational measurement research and higher education policy making.

2

Systematic Comparison of Decision Accuracy of Complex Compensatory Decision Rules Combining Multiple Tests in a Higher Education Context

This chapter has been published as:

Yocarini, I. E., Bouwmeester, S., Smeets, G., & Arends, L. R. (2018). Systematic Comparison of Decision Accuracy of Complex Compensatory Decision Rules Combining Multiple Tests in a Higher Education Context. *Educational Measurement: Issues and Practice*, 37, 24-39.

doi: 10.1111/emip.12186

Abstract

This real-data-guided simulation study systematically evaluated the decision accuracy of complex decision rules combining multiple tests within different realistic curricula. Specifically, complex decision rules combining conjunctive aspects and compensatory aspects were evaluated. A conjunctive aspect requires a minimum level of performance whereas a compensatory aspect requires an average level of performance. Simulations were performed to obtain students' true and observed score distributions and to manipulate several factors relevant to a higher education curriculum in practice. The results showed that the decision accuracy depends on the conjunctive (required minimum grade) and compensatory (required GPA) aspects and their combination. Overall, within a complex compensatory decision rule the false negative rate is lower and the false positive rate higher compared to a conjunctive decision rule. For a conjunctive decision rule the reverse is true. Which rule is more accurate also depends on the average test reliability, average test correlation, and the number of reexaminations. This comparison highlights the importance of evaluating decision accuracy in high-stake decisions, considering both the specific rule as well as the selected measures.

Keywords: high-stake decision, multiple measures, conjunctive decision rule, compensatory decision rule, decision accuracy.

Introduction

In the academic year of 2011-2012 a new compensatory testing system was introduced in the first year of the Psychology bachelor at the Erasmus University Rotterdam (EUR) in the Netherlands. In this compensatory testing system students are allowed to compensate, within certain boundaries, a low test score on one course with a high test score on another course. Contrary, students in a conjunctive testing system are required to pass each individual course (Chester, 2003). Given that a conjunctive testing system is commonly applied in higher education programs in the Netherlands, the introduction of this new compensatory testing system has been ground for some debate. Critics argue that allowing compensation creates hiatuses in knowledge and consequently leads to a devaluation of the diploma (Arnold, 2011). Within this context, an academic dismissal policy exists in Dutch higher education in which a decision, called the binding study advice (BSA), is made at the end of the first year of the bachelor. In this decision it is determined whether students meet the required number of study credits to be allowed to continue their bachelor studies. When allowing compensation between courses, this BSA decision is based on the average grade over courses instead of individual course grades. In other words, the average grade serves as a decision-making tool in a situation in which the stakes are high. Consequently, the accuracy of this decision is of great importance. The aim of this study is to compare the accuracy of different compensatory, conjunctive, and complex decision rules within different realistic higher education contexts.

Comparing the decision accuracy of these rules implies comparing the degree of erroneous decisions made, based on the decision rule applied (Douglas & Mislevy, 2010). One such erroneous decision is a false positive. In this case a student is allowed to continue to their second bachelor year while he or she is not sufficiently skilled. The other incorrect decision is a false negative. Here, a student is not allowed to progress to the second year while he or she is actually competent. As shown in Table 1, evaluating the type of incorrect classifications implies comparing the decision based on a student's latent true score to the decision based on a student's observed test score. Since a student's true score cannot be observed directly, this study includes simulations to obtain students' latent true scores using the classical test theory (CTT)

framework. A clear disadvantage of this simulation method is that many assumptions have to be made about both test and student characteristics. To ensure these assumptions are as accurate as possible we explicitly evaluated their tenability by using empirical information. Still, a difficult problem remains as students' behavior is dynamic and responsive (see e.g., Budescu & Bo's [2015] study on test-taking behavior within a test). Unfortunately, students' strategic behaviors in response to different decision rules is not modeled in the simulations. Instead, this behavior is assumed to be constant across decision rules. Despite this required assumption, the simulations are valuable because they allow us to evaluate the decision accuracy in a broad range of educational contexts. Here, aspects of the curriculum are varied (such as the correlation between tests, the number of tests, the average reliability of tests at an average true score, and the number of reexaminations² allowed).

Table 1: Classification Decisions

Decisions Based on <i>Observed Score</i>	Decision Based on <i>True Score</i>	
	Fail	Pass
Fail	Correct classification	Misclassification False negative
Pass	Misclassification False positive	Correct classification

Furthermore, decision rules applied in a higher education curriculum are rarely completely compensatory but rather a combination of conjunctive and compensatory aspects (a so-called complex decision rule; Douglas & Mislevy, 2010). To ensure the studied decision rules are realistic, we used the complex compensatory-conjunctive decision rule applied in the first year of the Psychology bachelor at the EUR³ and the traditional conjunctive decision rule applied in most Dutch universities as reference points. In additional complex decision rules, we varied the specific components around these reference rules.

² In this study the number of reexaminations refers to the number of tests a student is allowed to retake within a curriculum, assuming each test in the curriculum is allowed to be retaken only once. Note, that this differs from the situation in which students are allowed to retake a test multiple times within a curriculum.

³ See the Method section for an overview of the specific requirements in this decision rule.

Psychometric Motivation for Implementing Compensation

The implementation of a (complex) compensatory decision rule in a higher education study program may be partly motivated by psychometric arguments. As Lord (1962) showed, a conjunctive decision rule is suboptimal for observed scores that include measurement error, even if a conjunctive decision rule is assumed for the true scores. To illustrate, Lord derived the optimal decision rule for observed scores when combining two tests.⁴ Additionally, the psychometric argument for choosing a compensatory decision rule notes that decisions based on average scores are more reliable than those based on single scores (Vermeulen et al., 2012). This argument follows from CTT (see Appendix A for a detailed elaboration of this argument). This line of reasoning heavily relies upon CTT's assumptions of equal error variance across tests and true scores and CTT's assumption of the number of tests approaching infinity (Lord & Novick, 1968). Also, the argument implies test scores to be highly correlated (Haladyna & Hess, 1999). These assumptions can be problematic in practice.

First, tests of different courses are likely to vary with respect to the variance of the measurement error. Second, it is unlikely that the variance of the measurement error is equal for different values of the true scores. For example, in many first year Psychology curricula multiple choice (MC) tests are administered. In taking these MC tests, students with low true scores are expected to guess more often than students with high true scores. Therefore, random measurement error will have more influence in the observed scores of students with low true scores. Third, CTT assumes measurement error over different tests for one individual to cancel out over a large number of tests. However, in practice the number of tests included in a first year curriculum might not be large enough for the measurement error to cancel out and become zero for the average test score. Fourth, tests of different courses aim to measure different kinds of knowledge so the test scores might not be highly correlated. This makes it less likely that the reliability of a total score is high (Haladyna & Hess, 1999) as the confidence interval around the average grade increases as inter-correlations decrease, resulting in a less accurate average grade.

⁴ Special thanks to our anonymous reviewer who pointed us to this interpretation.

Given these likely violations of the assumptions in practice, it remains questionable whether the psychometric argument for allowing compensation between tests is generally tenable and the average grade is more reliable in practice. Consequently, the compensatory decision rule was included in our comparison of the decision accuracy of different (complex) decision rules.

Reliability and Decision Accuracy

The psychometric argument concerning the reliability described in the previous section is important as it relates to the decision accuracy. As mentioned before, evaluating the decision accuracy involves the comparison of the decision based on the latent true score and the decision based on the observed test score. Here, the true score corresponds to the average test score a student would obtain when he or she would take a parallel test infinity times. For a dichotomous decision this results in four quadrants of decision accuracy, as displayed in Table 1. A correct classification (i.e., an accurate decision) is made when both decisions align. If a selection instrument is more reliable, less measurement error is included in the observed test score. This means that the true score and observed score are more similar, which results in fewer false positives and false negatives.

Given our aim to evaluate the decision accuracy of different decision rules in realistic higher education settings, several variables are varied to mimic realistic settings. These variables were selected for their relevant influence on the decision accuracy either directly or indirectly through test reliability. Variables influencing the reliability of the selection instrument are the correlations between tests, the individual test reliability, and the number of tests, as described before. Practically relevant factors that influence the decision accuracy directly are the number of reexaminations and the required average and minimum grade. Assuming that only students who failed the test on the first attempt retake a test, reexaminations decrease the number of false negatives and increase the number of false positives. This is because students who partake in the reexamination were classified as either false negatives or true negatives on the first attempt. At the reexamination students who were classified as false negatives may become true positives and students who were classified as true negatives may become false positives. Secondly, the specific requirements in the

decision rule are relevant as misclassifications are especially present for true scores close to the cut-off score (Van Rijn, Béguin, & Verstralen, 2009). When a student's true score is further removed from the cut-off score, measurement error in the observed test score is less likely to cause a misclassification as decisions based on the true score and observed score are still likely to align.

Previous Studies

Previous studies examined the decision accuracy of different combinations of multiple tests as well as the influence of different factors on the decision accuracy of these combinations. Overall, these studies indicate that using a conjunctive, compensatory, or a complex decision rule results in different levels of decision accuracy. From his simulations, Lord (1962) concluded that, in the face of fallible measures, one better opts for some sort of compensation rather than using multiple cutting scores (i.e., a conjunctive decision rule). Hambleton and Slater (1997) conducted a simulation study to assess the accuracy of combining exercises within a test and found that with a compensatory and a complex compensatory-conjunctive rule false positives were more likely than false negatives. More recently, Douglas and Mislevy (2010) showed that using a complex decision rule, results in fewer decisional errors compared to a conjunctive rule, in terms of both false negatives and false positives. Furthermore, Van Rijn, Béguin, and Verstralen (2012) found that including conjunctive aspects in a complex decision rule in a secondary education context resulted in a higher percentage of misclassification compared to adding a condition that combined individual cut-off scores in the decision rule.

In addition, the influence of several factors on the decisional accuracy has been studied. For example, McBee et al. (2014) studied the decision accuracy in the context of identifying gifted students and evaluated the consequences of test reliability and correlations between tests. Their study shows that given their decision rule (which combines several scores by means of a conjunctive and a complementary rule, i.e., 'or' rule) lower test correlations and test reliability are associated with a higher proportion of decisional errors. Here, relatively more false negative classifications existed than false positives. In addition, Douglas and Mislevy (2010) showed that the number of false negatives and false positives was higher for a conjunctive decision rule compared

to a compensatory rule and that this effect was exaggerated when more tests were used. Also, their study showed that increasing the number of opportunities to pass increased the false positive rates. Notably, with three reexaminations, no false negatives were present in case of a compensatory decision rule. Hambleton and Slater (1997) also found that higher correlations between exercises and more items included in a test resulted in higher decision accuracy of a (complex) compensatory decision rule.

Research on the decision accuracy of different decision rules is still sparse yet informative (Haladyna & Hess, 1999). Several studies included a complex compensatory-conjunctive decision rule, however, none of the studies evaluated the influence of varying the specific conjunctive and compensatory requirements within a complex rule. Although part of the results might be intuitively theorized, the size of the difference in the accuracy of different complex decision rules may not. Also, none of the previous studies were placed in the context of higher education curricula. Practitioners might need to specify the requirements in a complex decision rule in a higher education curriculum and previous results might not provide easy guidance for this purpose. To enable evidence-based curriculum implementations, this study evaluates the proportions of false negatives and false positives across different complex decision rules within realistic higher education curricula.

Hypotheses

In light of the aim of this simulation study to compare the accuracy of different compensatory, conjunctive, and complex decision rules within realistic higher education settings, several variables were varied. We included specifically these variables for three reasons. First, we wish to replicate previous findings by evaluating the influence of correlation between tests and the number of tests. Importantly, we extend these findings by adding higher levels of correlations between tests. This is interesting as it informs practitioners how to form clusters of courses in which compensation is allowed. Second, we evaluate the test reliability and the number of reexaminations to see if these factors influence the decision accuracy as expected. Although McBee et al (2014) also evaluated the test reliability, they did not evaluate how and whether test reliability differently influenced a conjunctive, compensatory,

and complex rule. This is interesting as measurement error may cause the conjunctive decision rule to be more inaccurate (i.e., produce relatively more false negatives) than a compensatory decision rule. Third, by including all these variables this study provides a comprehensive overview of the different influences on decision accuracy for practitioners.

Specifically, the number of tests, the number of reexaminations, the test reliability, and the correlations between tests were varied in our simulations. Moreover, the studied decision rules differed in their compensatory (i.e., the average grade) and conjunctive (i.e., the minimum grade) requirement. Overall, in line with previous studies, it was predicted that more decision errors are made using a conjunctive decision rule compared to a compensatory decision rule. Specifically, in line with our reasoning above, it was hypothesized that more misclassifications occur when the cut-off score approaches the average (true) score.

Furthermore, measurement error (which is related to the test reliability) was expected to have a stronger influence on the decision accuracy of conjunctive decision rules than on the accuracy of compensatory rules. For conjunctive rules an unreliable test may easily result in a classification error. In a compensatory rule the result of an unreliable test may be compensated by the other tests in the curriculum, making it less likely to result in a classification error compared to a conjunctive rule. Given that the average grade becomes less accurate with low inter-correlations we also expected the differences between the conjunctive and compensatory rules to be more explicit for low correlations between tests. In line with CTT and previous studies (Douglas & Mislevy, 2010; Hambleton & Slater, 1997), it was hypothesized that increasing the number of tests increases the accuracy of compensatory decision rules, as measurement error is more likely to cancel out and result in a more reliable average grade. Alternatively, with more tests it becomes more likely that measurement error on a single test administration causes an individual test score to be lower or higher than the true score. Consequently, we expected the false negative and false positive rate to increase for conjunctive rules. Finally, following Douglas and Mislevy (2010) and our previous discussion, increasing the number of reexaminations was expected to decrease the false negative rate and increase the false positive rate. In the

(complex) compensatory rule fewer reexaminations are required as compensation is allowed, so here it was expected that reexaminations had a smaller influence on the decision accuracy compared to the conjunctive decision rule.

Method

Simulation Model

The procedure for performing our simulation study was in line with the simulation method developed by Douglas (2007) as applied in Douglas and Mislevy (2010). Broadly, the simulations were structured through the following steps: (1) simulate a true score distribution for each test, (2) simulate observed scores for each student by simulating error around the true scores, (3) simulate replicate scores for the reexaminations, and (4) evaluate the decision accuracy by computing the appropriate indices.

First, T true score distributions were simulated for each test. The mean of T was assumed to vary for each test. Data from three cohorts of first year Psychology students at the EUR were used to obtain a realistic simulated mean true score. Specifically, data were obtained from eight tests of 246 students in cohort 2011, 245 students in cohort 2012, and 330 students in cohort 2013. In total eight tests were used which each had 40 multiple choice items with four answer categories. These samples included students who had obtained at least one test score throughout the year. For the total sample, mean observed test score were calculated for each test, see Table 2 for descriptive statistics of the empirical data. The standard deviation and mean of these mean observed test scores were estimated to define the distribution from which mean true scores were sampled for each simulated test⁵. The true score variance was assumed to be equal across tests, which means that the true scores within each course were assumed to vary by the same amount across different courses. A realistic value for the true score variance was estimated by calculating the variance in the observed test scores for each test and taking the mean of these variances. Importantly, the true scores were truncated between 1.0 and 10.0, to

⁵ Note that true scores were not varied systematically across simulated datasets, meaning that we did not evaluate decision accuracy for different student ability levels.

mimic the Dutch higher education grading system. Consequently, the T distributions were simulated from a multivariate truncated normal distribution to simulate different levels of correlations between the tests. See Appendix B for a detailed outline on the simulation procedure, the specific assumptions, and an example of code to perform the simulations in R (R Core Team, 2015).

Table 2: Descriptive Statistics Empirical Data

Descriptive Statistic	Test							
	1	2	3	4	5	6	7	8
N	817	797	758	727	719	706	687	678
Min	1.9	1.0	1.0	2.0	2.3	2.9	1.8	3.1
Max	9.3	10.0	10.0	9.7	9.7	10.0	9.8	9.5
Mean	5.89	6.70	6.11	6.85	6.71	6.64	6.77	6.43
SD	1.16	1.34	1.70	1.26	1.20	1.11	1.15	1.04

Correlation between tests. The correlations between tests were manipulated to evaluate the optimal degree of cohesion between courses that results in the most accurate decision. The latter helps to construct guidelines on forming clusters of courses wherein students are allowed to compensate. Varying these correlations ensured that the true scores on different tests were more or less alike. Taking the first year Psychology at EUR and the correlations used by Hambleton and Slater (1997) as an example, a realistic average correlation between courses was .3. As other study programs might have more or less cohesion between courses, the correlation was manipulated to be .1, .3 .5, or .7.

Average true score test reliability. Secondly, error was simulated around the true scores to produce the simulated test scores. This error variance was estimated using the test reliability. Following our discussion of assuming equal measurement error variances in CTT in the Introduction, we assumed the test reliability to vary as a function of the true score; the higher the true score, the lower the measurement error variance, the higher the test reliability. In defining the test reliability at a specific true score, the following functions were used: $b_1 = \left(\frac{R_{xx} - .99}{\bar{T} - 10}\right)$, $b_0 = R_{xx} - (b_1 * \bar{T})$, and consequently $R_{xx \text{ at } T} = b_0 + b_1 * T$. Here, R_{xx} refers to the test reliability at an average true score, \bar{T} , which was manipulated to be 0.4, 0.6, and 0.8. Since R_{xx} has a

maximum of $R_{xx} = 1$, which indicates no measurement error, the maximum reliability at a maximum true score of $T = 10$ was set at 0.99⁶. Consequently, the error variance at T was defined as: $s_E^2 = \left(\frac{s_T^2}{R_{xx \text{ at } T}} \right) - s_T^2$. By this definition, there is more error variance at lower true scores and less error at higher true scores.

Number of tests and reexaminations. Finally, to study the influence of the number of reexaminations, replicate observed scores were drawn as well. As noted, students were assumed to retake a test only once in a first-year curriculum. For these replicate observed scores, it was assumed that someone's true score had increased between the first test administration and the reexamination as students gained knowledge within this time interval. An estimate of the increment in true score (set at 0.5) was obtained from available data of reexaminations taken by first year Psychology students at the EUR. To analyze the influence of the number of reexaminations, several conditions were simulated; no reexaminations, 1, 2, 3, 4, or all tests in the curriculum. In addition to varying the number of reexaminations, the number of tests was also varied to be 8 or 12. Both situations are realistic in a first-year curriculum.

Measure of Decision Accuracy

The decision accuracy of using different decision rules was evaluated by looking at four measures of classification accuracy. First, we evaluated the total proportion of misclassification. This is the proportion of misclassified students relative to the overall group of students, N : $P(\text{misclassification}) = \frac{P(X < c | T > c) + P(X > c | T < c)}{N}$. Here, c indicates

the cut-off score. Secondly, we evaluated the false negative rate which is the conditional probability that someone with a qualifying true score is misclassified:

$P(X < c | T > c) = \frac{P(X < c \ \& \ T > c)}{P(T > c)}$. The sensitivity rate can be easily obtained using the

false negative rate: sensitivity rate = 1 – false negative rate. Thirdly, we evaluated the false positive rate. This is the conditional probability that a student with a

disqualifying true score is misclassified: $P(X > c | T < c) = \frac{P(X > c \ \& \ T < c)}{P(T < c)}$. The specificity

⁶ A sensitivity analysis in which we also evaluated the results where the maximum reliability at a maximum true score was set at 0.90 as well as a classical test theory interpretation of reliability (not varying across true scores) showed the results were robust under these alternative error variance methods of simulation. See <https://osf.io/8pgyt/> for the results of the sensitivity analysis.

rate can be easily obtained using the false positive rate: specificity rate = 1 – false positive rate. Finally, we evaluated the positive predictive value. This is the conditional probability that someone with a qualifying true score is identified correctly $P(T > c | X > c) = \frac{P(X > c \& T > c)}{P(X > c)}$. In accordance with Van Rijn et al. (2012) the negative predictive value was not considered.

Decision Rules

In this study, different realistic decision rules were evaluated and compared; see Table 3 for an overview. For the complex compensatory-conjunctive decision rules we used the rule applied in the Psychology bachelor at the EUR as a reference point. For the conjunctive decision rules, the rule used among most Dutch universities was used as a reference point. In additional complex decision rules, we varied the specific conjunctive and compensatory components around these reference rules. As the test scores were allowed to range between 1.0 and 10.0, a rule that requires a minimum grade of 1.0 is similar to using a compensatory rule because only the required GPA is relevant in this case. Furthermore, the curriculum aspects were evaluated in a fully crossed design. In total 144 conditions existed. For each of these conditions 500 datasets of 2000 students were simulated to obtain stable results. Finally, the decision accuracy measures were computed for each decision rule and dataset.

Table 3: Decision Rules

Decision Rule	Score Requirements	
	GPA	Minimum grade
1. Compensatory rule	5.5	1.0
2. Complex compensatory rule	5.5	3.0
3. Complex compensatory rule	5.5	4.0
4. Complex compensatory rule	5.5	5.0
5. Conjunctive rule	5.5	5.5
6. Compensatory rule	6.0	1.0
7. Complex compensatory rule	6.0	3.0
8. Complex compensatory rule ¹	6.0	4.0
9. Complex compensatory rule	6.0	5.0
10. Conjunctive rule	6.0	6.0
11. Compensatory rule	6.5	1.0
12. Complex compensatory rule	6.5	3.0
13. Complex compensatory rule	6.5	4.0
14. Complex compensatory rule	6.5	5.0
15. Conjunctive rule	6.5	6.5

¹Decision rule as applied in the first year Psychology at the EUR.

By studying these specific decision rules, using data as a basis for the simulations, several assumptions were made with respect to the setting and structure of the educational program. The students included in the observed data had eight knowledge tests in a year, programmed in a sequential format. Also, the observed test scores in the data all originate from MC tests. In the complex compensatory decision rule at the EUR students were only allowed two reexaminations when their GPA was below a 6.0 or when an individual test score was below a 4.0 and these reexaminations took place at the end of the academic year.

Results

In discussing the results of our simulation study, we focus on comparing the decision accuracy of the different decision rules, averaged over all manipulated conditions. These mean values are displayed in Table 4. In addition, the representativeness of these mean values for the simulated conditions is described. An elaborate description of the results per manipulated factor is provided in Appendix C with an overview of

the results per simulated condition in Table C1 to C4. For results on specific conditions, researchers can evaluate these themselves using data of our simulations that is freely available from the Open Science Framework (OSF) directory at <https://osf.io/zmvbh/>.

In the next paragraphs the influence of the required GPA and minimum grade on the decision accuracy of a complex compensatory decision rule is evaluated first. Second, the accuracy of the compensatory rules is compared to that of the conjunctive decision rules. Finally, the mean values observed in Table 4 are compared to the results for each separate condition in Table C1 to C4, which illustrate the most important deviations from the patterns observed in Table 4.

Table 4: Mean Values for Each Outcome Measure per Decision Rule

Decision Rule	GPA	Minimum	Mean Proportion Misclassifications	Mean False Negative Rate	Mean False Positive Rate	Mean Positive Predictive Value
1	5.5	1	.06(.04)	.02(.02)	.62(.24) ²	.95(.03)
2	5.5	3	.10(.08)	.07(.09)	.49(.23)	.96(.03)
3	5.5	4	.17(.11)	.14(.14)	.41(.20)	.94(.02)
4	5.5	5	.26(.08)	.26(.16)	.29(.14)	.81(.10)
5	5.5	5.5	.24(.06)	.31(.17)	.21(.11)	.68(.16)
6	6	1	.14(.06)	.03(.03)	.55(.25)	.87(.06)
7	6	3	.15(.06)	.06(.08)	.48(.22)	.88(.05)
8	6	4	.18(.08)	.12(.12)	.41(.20)	.89(.04)
9	6	5	.25(.08)	.25(.16)	.29(.14)	.80(.10)
10	6	6	.17(.06)	.37(.17)	.14(.08)	.55(.20)
11	6.5	1	.23(.10)	.05(.05)	.44(.25)	.73(.11)
12	6.5	3	.22(.10)	.06(.07)	.42(.23)	.74(.10)
13	6.5	4	.22(.09)	.10(.10)	.38(.20)	.75(.10)
14	6.5	5	.23(.08)	.20(.15)	.28(.15)	.74(.11)
15	6.5	6.5	.10(.05)	.42(.18) ¹	.07(.05)	.44(.22) ³

Note: *SD* over simulations given in brackets. Darker shades of grey implicate increased accuracy (i.e., lower proportion of error, false negative rate and false positive rate, and higher positive predictive value). When the required GPA equals the required minimum, the decision rule is conjunctive. When the required minimum equals 1, it is a compensatory decision rule. The remaining rules are complex compensatory- conjunctive decision rules. ¹ $N=71954$ instead of $N=72000$, ² $N=71997$, ³ $N=71952$.

Proportion of Misclassifications

As shown in the mean proportion error column in Table 4, the proportion of misclassifications depended on both the specific required GPA and required minimum grade. As expected, increasing the required minimum grade increased the mean proportion of misclassifications in the (complex) compensatory decision rules when the GPA was not too strict. This means that the compensatory rule resulted in the most accurate decision. At a strict GPA, the required minimum grade did not influence the decision accuracy of compensatory rules. Overall, increasing the GPA resulted in a large to moderate increase of the proportion of misclassifications (except when the minimum grade was high and increasing the GPA had a small negative influence).

Comparing the decision accuracy of the compensatory and conjunctive decision rules with a similar required GPA shows that the (complex) compensatory rules were generally more accurate when the required GPA was low. When the required minimum grade in the complex compensatory rules was high, the conjunctive rule was more accurate. Furthermore, when the GPA was closest to the average population true score (i.e., high), the conjunctive decision rule resulted in fewer total misclassifications.

Table C1 in Appendix C shows the results for each factor separately which show that for most conditions the results are consistent the pattern observed in the mean proportion of error in Table 4. Some exceptions exist. The differences in accuracy for the different decision rules were smaller when the test correlation or test reliability was high. Also, the accuracy was higher when the test reliability was high. Finally, when no reexaminations were allowed or when the average test reliability was low, the minimum grade had a more pronounced influence on the decision accuracy than seen in the average pattern. In light of our hypotheses, the results in Table C1 show that the average test reliability mostly had a larger influence on the proportion of misclassifications for (complex) compensatory decision rules than for conjunctive rules given a specific GPA. As expected, higher test correlations resulted in a smaller proportion of misclassification than lower test correlations in complex compensatory decision rules. Also, the differences in proportion of misclassifications for the different decision rules were larger at lower test correlations.

The False Negative Rate

The false negative rate of the different decision rules shown in Table 4 illustrate a clear pattern: the higher the required minimum grade, the higher the false negative rate. So, the compensatory decision rules were the most accurate. The required GPA had a small positive influence if a compensatory decision rule was used, and a small negative influence when a complex compensatory decision rule was applied. Overall, the pattern is consistent, such that the conjunctive decision rules had higher false negative rates than the (complex) compensatory rules requiring the same GPA.

Comparing the pattern in the mean values of the false negative rate in Table 4 to the patterns observed over the different conditions in Table C2 in Appendix C shows that the mean values were very representative. The only differences were observed when the test reliability was low, no reexaminations were allowed, or when the correlation between the tests was low. In these conditions, the influence of the minimum grade was slightly more pronounced, such that there were larger differences in the false negative rates across different decision rules. Regarding our hypotheses for the false negative rate, the results in Table C2 show that the false negative rate increased for conjunctive rules when more tests were included. In addition, increasing the number of reexaminations decreased the false negative rate. The influence of the number of reexaminations was larger for conjunctive decision rules compared to (complex) compensatory rules.

The False Positive Rate

Similarly, the false positive rates in Table 4 show a consistent pattern: the higher the minimum grade, the lower the false positive rate. Consequently, the compensatory decision rules were the least accurate. Furthermore, increasing the GPA resulted in a decrease in the false positive rate. Hereby, the negative influence of the GPA was large for compensatory decision rules and became small as the required minimum grade increased. Overall, the conjunctive decision rules were the most accurate.

In addition, the pattern observed in the mean values of the false positive rate in Table 4 is comparable to the patterns observed in Table C3 in Appendix C. The only differences are observed for the condition in which no or one reexamination is

allowed. Here, the overall false negative rate was lower than observed in the mean values and the differences in the false positive rates across rules was smaller. In line with our hypothesis, increasing the number of reexaminations increased the false positive rate. Contrary to expectations, the number of reexaminations had a larger influence on the false positive rate of (complex) compensatory decision rules than conjunctive rules.

Positive Predictive Value

The mean positive predictive values provided in Table 4 show that the positive predictive values of the different decision rules mostly depended on the required GPA. The higher the GPA, the lower the mean positive predictive value. This influence became smaller as the minimum grade increased. Overall, the minimum grade had a small negative influence. When the required GPA was strict, the influence of the minimum grade on the positive predictive value of the complex compensatory rules disappeared. Overall, the positive predictive value of a complex compensatory decision rule was higher than that of a conjunctive decision rule with a similar required GPA.

Table C4 in Appendix C shows the positive predictive value results for each manipulated factor. The pattern illustrated resembles the pattern observed in Table 4. Differences are mainly observed when the test correlation or test reliability was high, or when reexaminations were not allowed. In these conditions, the differences in the positive predictive value of the different decision rules were less pronounced than the differences observed in Table 4.

Discussion

The purpose of this study was to compare the accuracy of different compensatory, conjunctive, and complex decision rules within different realistic higher education contexts. Overall, the results indicate that the accuracy of the decision rules depends on the degree of compensation allowed. For the total proportion of misclassifications, the results show that the required minimum grade and GPA interplay. Specifically, at a low GPA the compensatory decision rule was the most accurate, while at a high GPA the conjunctive decision rule was the most accurate. This result can be explained by

the proportion of false positives which dramatically decreased when the requirements within the conjunctive rule were closer to the average true score. For the remaining outcome measures, the results were more consistent. Overall, conjunctive decision rules had a higher false negative rate and a lower false positive rate compared to compensatory decision rules requiring a similar GPA. In addition, the compensatory decision rules had a higher positive predictive value than conjunctive decision rules requiring a similar GPA.

The patterns in the overall results displayed in Table 3 were representative of the patterns observed in the separate settings. Deviations from the overall pattern were mainly observed when the test reliability was high or low, the test correlation was high or low, or whether none or many reexaminations were allowed. As hypothesized, the differences between the decision rules became more explicit when correlations were low. Contrary to expectations the average test reliability had a larger influence on the proportion of misclassifications for (complex) compensatory decision rules than for conjunctive rules. This finding shows that test reliability has an important influence on the decision accuracy and is as important for compensatory as for conjunctive decision rules. Adding tests to the curriculum increased the false negative rate for conjunctive rules as hypothesized. Also, the number of reexaminations decreased the number of false negatives and increased the number of false positives. As expected, the influence of the reexaminations was larger for conjunctive rules than for (complex) compensatory decision rules. On the contrary, the reexaminations had a larger influence on the false positive rate of (complex) compensatory rules than conjunctive rules. This is because false positives are in general more likely in compensatory decision rules than conjunctive rules.

Overall, the results from this study are in line with previous findings. As Douglas and Mislavy (2010) found, a combination of a conjunctive and compensatory decision rule results in less decision errors. Our results show that this depends on the specific requirements in the decision rule; the complex rule was more accurate than the conjunctive decision rule when the required GPA and minimum grade were not too strict. Furthermore, the results from our study are similar to McBee et al. (2014) their finding that with lower test correlations and lower test reliability a higher proportion

of false negatives and false positives is present. Hereby, the influence of test reliability on the false positive rate was somewhat stronger than the influence of the correlation between the tests. Furthermore, Douglas and Mislevy (2010) found that increasing the number of tests exaggerated the difference in the number of false positives and false negatives of the conjunctive and compensatory decision rules. The current results did not show such a clear pattern for increasing the number of tests. A possible explanation for this difference originates in the different factors that were included in this study. As additional factors were manipulated, the influence of the number of tests might not be a main effect but instead be moderated by other factors.

As a whole, the findings from this study indicate that it is not only the manner in which the multiple measures are combined that is important for the accuracy of a decision, the measures selected are just as important. These findings support of Chester's (2003) conclusion. Mostly, a selection of measures in terms of average reliability and correlation between the tests seems important.

Recommendations

Although the results suggest decision accuracy to be context dependent, some recommendations for implementing a (complex) compensatory decision rule might be possible based on these results. Mostly, decision makers have to determine the specific trade-off between false positives and false negatives. Consequently, in practice, choosing the appropriate decision rule implies a discussion of the relative emphasis put on preventing false positives or false negatives. This is highly dependent on the context in which the decision is placed (i.e., the stakes involved) as well as the perspective one takes (see e.g., Mehrens, 1990, for an overview of when (not) to use composite scores in decision making). For example, as courses become more advanced and specialized it is recommended to allow for less compensation as the prevention of false positives would become increasingly important.

Furthermore, the results show that one should allow compensation within a cluster of courses that are correlated. In highly correlated clusters the differences in accuracy between different decision rules becomes smaller and the overall accuracy is higher. Selecting courses to obtain a highly correlated cluster can be done based on, for

instance, content or difficulty level. Overall, with low correlation between tests, allowing compensation between the tests should be carefully considered as it becomes questionable whether these tests could compensate one another content-wise.

Considerations

Several assumptions were made in this simulation study, see Appendix A for a detailed outline thereof. For example, it was assumed that all students employed a similar strategy and choose to retake the course on which their observed score was lowest. In real life situations different groups of students might employ different strategies. One might for instance argue that students opt a more optimal retake strategy and choose those tests where the discrepancy between their observed and true score is highest. Because students might not be good in defining their true score accurately and consequently the discrepancy between their observed and true score in general, it was chosen to simulate a strategy in which students retook the test that had the lowest observed score.

Furthermore, an empirical approach was taken in this study by using empirical data as the basis for the simulations. This data only includes Dutch first year Psychology students at the EUR. Consequently, the specific accuracy levels might differ for other programs or similar bachelor programs in different cities or countries and therefore one should not focus on these specific values. Alternatively, this study aims at analyzing overall effects of having a higher or lower minimum required grade, not the specific value ascribed to it as this might vary in different testing systems.

Accordingly, interpreting the results as such, the results are more easily generalized to other testing systems as well as other decision-making situations.

As mentioned in the Introduction, it was assumed that students behave similarly under each of the decision rules by means of similar true and observed score distributions. Hereby, specific learning strategies that students possibly apply were ignored. As argued by Van Rijn et al. (2012), this is not to say that in practice these exact accuracy levels will automatically occur once a specific decision rule is applied. Students are able to react to different testing systems by, for instance, allocating their study time accordingly. In this context it remains questionable whether students are

capable of directing their study activities in such a way that they would exert a strong influence on the specific true score they achieve. Further studies should explore the possibility of incorporating alternative study strategies for different decision rules.

Future Directions

Although there is a vast amount of literature on the decision accuracy of single assessments (Cronbach, 1951; Lee, 2010; Lee, Hanson, & Brennan, 2002; Rudner, 2005; Sijtsma, 2009) this research is not easily generalized to situations in which multiple assessments are combined (Douglas & Mislavy, 2010; Van Rijn et al., 2012). Likewise, studies into the measurement precision of composite scores (e.g., He, 2009; Wheadon & Stockford, 2013) do not easily apply to situations in which composite scores are not easily computed or useful. Consequently, future studies should examine the decision accuracy of using multiple measures and in particular focus on the plausibility of the assumptions that were made in the current study. The assumptions regarding the reexaminations should be tested in further studies to see if the results would change considerably when other reexamination strategies are assumed.

Overall, the results suggest that the accuracy of complex decision rules depends on the specific requirements set within a complex decision rule. Consequently, it seems that one should carefully consider the strictness of the GPA and minimum grade required in a complex decision rule. This implies that the educational decision maker should make a trade-off between the emphases put on preventing false negatives versus false positives. Hereby, the specific context of the decision is important as well as the tests that are used to make the decision. In making these trade-offs, this study might aid as a guideline.

Appendix A. Psychometric Argument: Average grades more reliable

Within the CTT framework it is assumed that each individual observed test score, X , is a summation of someone's true score, T , and random measurement error, E (Lord & Novick, 1968);

$$X = T + E.$$

Here, both T and E are unobserved theoretical variables. Moreover, because CTT assumes the correlation between T and E to be zero,

$$r_{T,E} = 0,$$

the variance of X is defined as

$$s_X^2 = s_T^2 + s_E^2.$$

Consequently, the theoretical definition of reliability equals

$$r_{xx} = \frac{s_T^2}{s_X^2} = \frac{s_T^2}{s_T^2 + s_E^2}.$$

From this definition it follows that a test is more reliable when the variance of measurement error is low. Now let us assume that X is a composite score defined as the average test score of a series of courses. Then T is the average true score and E is the average of the individual measurement errors. For the sake of simplicity, let's assume an equal measurement error variance for each course. Because E_i is a random variable with a population mean of zero, the average E_i of an individual student will approach zero when the number of tests that are combined in the composite becomes large. As this is true for all individuals it follows that the σ_E^2 of a composite score is smaller than the σ_E^2 of a single test score. Therefore, the reliability of a composite score is more reliable than that of a single test.

Appendix B. Detailed Outline Simulation Procedure

In this appendix, the simulation procedure including the implied assumptions is discussed in detail. Here, the exam scores of 821 first year students on eight courses were used to obtain several estimates.

First, a covariance matrix was computed that included the variance and covariance of each of the tests included in the decision. Consequently, an R function `sig` was written that enables the manipulation of the cohesion (`cor.mean`) and the number of tests (`n`):

```
> sig <- function(cor.mean, n, s2T){
+   c <- c(rep(cor.mean, n*n) #creating correlation vector with
+     similar correlations
+   sigma <- matrix(c,n,n)
+   diag(sigma) <- 1 #correlation matrix
+   sigma <- sigma*(sqrt(s2T)*sqrt(s2T)) #correlation to covariance
+     matrix
+   return(as.matrix(sigma))}
```

The argument `s2T` indicates the variance in true scores on a test and was estimated from the available data as the average variance in test scores over all courses. This variance in true score was assumed to be similar for each course. In addition, the correlation between each test in the curriculum was assumed to be similar.

Furthermore, the function `sig` returned a symmetric covariance matrix.

Secondly, the simulated covariance matrix was used as input for the sampling of a true score distribution using the function `true.score`. These true scores varied between 1 and 10 and were therefore simulated from a truncated multivariate normal distribution, meaning that the underlying true score distribution was assumed to be normally distributed. The R function `rtmvnorm` from package `tmvtnorm` (Wilhelm & Manjunath, 2014) was used for this purpose:

```
> true.score <- function(N, n, m, s, sigma, a, b){
+   require("tmvtnorm")
+   a = c(rep(a, n)) #lowerbound
+   b = c(rep(b, n)) #upperbound
+   mean <- rnorm(n=n, mean=m, sd=s) #random mean true score for
+     each test
+   true.score <- rtmvnorm(n=N, mean=mean, sigma=sigma, lower=a,
+     upper=b, algorithm="rejection") #simulate true
+     score distribution for each test
+   corcheck <- c(mean(cor(true.score)) #check correlation in output
```

```
+ return(list(true.score=true.score, cor=corcheckt))}
```

Again, the arguments included in the function allowed for manipulation of several parameters; the sample size (N), the number of tests (n), the mean true score value (m), the variability in mean true score values (s), covariance matrix (σ), and the lower (a) and upper (b) bound. Both the mean true score and the standard deviation of these means were estimated from the available test scores. Subsequently, these values were used to randomly sample mean true scores for each test. Which were consequently used to sample the true scores from a truncated multivariate normal distribution. In addition to a student his or her true score for each test, the function included a check for the strength of the correlations of the final true scores between tests to see if the manipulation of the correlations between tests was successful (`corcheckt`). Comparing the output to the input showed that the simulations with a correlation $r = .1$ resulted in an average correlation of $r = .19$, for the $r = .3$ simulations it was $r = .36$, for the $r = .5$ simulations it was $r = .54$, and finally for the $r = .7$ simulations the simulated datasets had an average correlation of $r = .72$. These differences were caused by the truncation of the true score distribution using a rejection algorithm. Because of the truncation some sampled distributions were rejected as they did not fit in the specified lower and upper bounds and this caused a different correlation in the remaining samples compared to the input.

Thirdly, the observed test scores were simulated. To do so, the true scores were used as the mean, and the measurement error functioned as the standard deviation to randomly define the observed scores, using the function `obsscore`. This function included the parameter test reliability (R) that could be manipulated. Notably, this average true score test reliability referred to the test reliability at an average true score. For other true scores however, the reliability varied as it depends on the specific true score. Consequently, given the reliability at a specific true score, the error variance was calculated and used as an estimate of the measurement error. Again, scores were bounded to fall between 1.0 and 10.0.

```
> obsscore <- function(R, m, s2T, true.score){
+   t = as.vector(true.score)
+   n = length(t)
+   R10 = 0.99 #reliability at true score 10
+   Rmu = R #reliability at mean true score
```

```

+ b1 = ((Rmu-.99)/(mu-10))
+ b0 = Rmu - (b1*mu)
+ Rt = b0 + b1*t #linear equation to define reliability at each
                true score
+ Rt <- replace(Rt, Rt <=0, 0.01) #replace reliability of 0 or < 0
+ errorT = (s2T/Rt)-s2T #error variance at t
+ obs.score <- rnorm(n, mean = t, sd = sqrt(errorT))
+ obs.score <- replace(obs.score, obs.score > 10.0, 10.0)
+ obs.score <- replace(obs.score, obs.score < 1.0, 1.0)
+ obs.score <- matrix(obs.score, nrow(true.score),
                    ncol(true.score))
+ return(obs.score = obs.score)}

```

Subsequently, the procedure of taking observed test score was duplicated to obtain a replicate observed score in case a student chooses to retake the test, using the `replicatescore` function. Here, one difference compared to the `obsscore` function existed. Given that the test was taken again at the end of the academic year, it was assumed that a student his or her true score increased as students were assumed to have obtained more test taking skills and relevant knowledge in the interval between the first attempt and the reexamination. An estimate of the increase was obtained from available data on reexaminations by first year Psychology students (approximately 0.5) and was set equal for all students. In simulating the change in the true score at the reexamination, all true scores increased by the same amount that was estimated from the data (approximately 0.5).

Fourthly, the `retakes` function was used. This function determined whether a student passed or failed his or her first year. Hereby, `x` is the input score, which refers to the observed score, `z` refers to the number of reexaminations allowed, `ret` refers to observed score on the reexamination, produced by the `replicatescore` function. Finally, `min` and `GPA` refer to the required minimum grade and GPA in the decision rule that is applied. Importantly, a test could only be retaken once and the retake was restricted to a test that had not been retaken before, the highest grade (of first attempt or reexamination) was used, and students were only allowed to retake a test if it was below the minimum grade or their GPA was below the required GPA. In the latter case, the course with the lowest observed score was retaken.

```

> retakes <- function(x,z,ret,min, GPA){
+   count <- matrix(0, nrow(x), ncol(x)) #matrix to identify which
+                                       test has been retaken

+   result <- c()
+   max <- z
+   r <- c(rep(0, nrow(x))) #number of tests retaken
+   for (i in 1:nrow(x)){
+     r[i] = 0
+     for (j in 1:ncol(x)){ #checking minimum grade
+       if (x[i,j] < min & count[i,j]==0 & r[i] < z){
+         if (x[i,j] < ret[i,j]){
+           x[i,j] <- ret[i,j]}
+         else {
+           x[i,j] <- x[i,j]}
+         count[i,j] = 1
+         r[i] = r[i] +1}}
+     for (n in r[i]:max){
+       if(mean(x[i,]) < GPA & r[i]<z){ #checking GPA
+         j <- which(x[i,]==min(x[i,][count[i,]==0]))[1]
+         if (x[i,j] < ret[i,j]){
+           x[i,j] <- ret[i,j]}
+         else {
+           x[i,j] <- x[i,j]}
+         count[i,j] = 1
+         r[i] = r[i] + 1}}}}
+   for (i in 1:nrow(x)){
+     if(min(x[i,])< min){ #if a score is below required minimum
+                           student fails
+
+       result[i] = 0}
+     else if(mean(x[i,])< GPA){ #if GPA is below required GPA
+                               student fails
+
+       result[i] = 0}
+     else {
+       result[i] = 1}}
+   return(list(r, count = count, result = result))}

```

Consequently, the matrix that defined which test was retaken was used to define whether a student passed or failed based on his or her true score. For this the function `trueretakes` was used. Here, the inputs were the true scores (`truescore`), the increased true score at the reexamination (`trueretake`), the matrix that defines which tests are retaken (`count`), the required minimum grade (`min`), and the required GPA (`GPA`).

```

> trueretakes <- function(truescore,trueretake,count, min,GPA){
+   result <- c()
+   for (i in 1:nrow(truescore)){
+     for (j in 1:ncol(truescore)){
+       if (count[i,j] == 1){
+         truescore[i,j] <- trueretake[i,j]}}
+   for (i in 1:nrow(truescore)){
+     if(min(truescore[i,])< min){
+       result[i] = 0}
+     else if(mean(truescore[i,])< GPA){
+       result[i] = 0}
+     else {
+       result[i] = 1}}
+   return(list(result = result))}

```

Finally, the decision vector was converted into a classification table from which the appropriate measures were calculated using the function `classtable`. This function required the results from the decision rule (whether students passed or failed) based on the true and observed score respectively.

```

> classtable <- function(Tr, X){ #for each decision rule you get
                                classification table
+   v = 0
+   w = 0
+   x = 0
+   y = 0
+   for (i in 1:length(Tr)){
+     if (Tr[i]== 0 & X[i]== 0){
+       v = v+1}
+     if (Tr[i] == 1 & X[i]==0){
+       w = w+1}
+     if (Tr[i] == 0 & X[i] == 1){
+       x = x+1}
+     if (Tr[i] == 1 & X[i] == 1){
+       y = y+1}}
+   class <- matrix(c(v,x,w,y),2,2)
+   sensitivity <- (class[2,2]/(class[1,2]+class[2,2]))
+   specificity <- (class[1,1]/(class[1,1]+class[2,1]))
+   totalmiss <- (class[1,2]+class[2,1])/(sum(class))
+   pospred <- (class[2,2]/(class[2,1]+class[2,2]))
+   return(matrix(c("v" = v, "w" = w, "x"=x,"y"=y,
+                   "sens"=sensitivity,
+                   "spec"=specificity,
+                   "total"=totalmiss,"pos"=pospred),1,8))}

```

Appendix C. Detailed Elaboration of the Results per Factor

Table C1: Mean Proportion of Misclassifications for Decision Rules and Factors Test Correlation, Test Reliability, Number of Tests, and Number of Reexaminations

Decision Rule	GPA	Minimum	Mean	Proportion Error	Average		Number					Number								
					Test Correlation	Test Reliability	of Tests					of Reexaminations								
					0.1	0.3	0.5	0.7	0.4	0.6	0.8	8	12	0	1	2	3	4	max	
1	5.5	1	.06(.04)	.03(.02)	.06(.02)	.08(.03)	.09(.03)	.08(.04)	.06(.03)	.04(.02)	.07(.04)	.07(.04)	.06(.03)	.07(.04)	.06(.03)	.06(.03)	.06(.03)	.06(.03)	.06(.04)	.07(.04)
2	5.5	3	.10(.08)	.10(.11)	.10(.08)	.10(.06)	.09(.04)	.17(.09)	.08(.04)	.05(.02)	.09(.04)	.09(.04)	.10(.09)	.18(.13)	.09(.06)	.08(.04)	.08(.04)	.08(.04)	.08(.04)	.09(.04)
3	5.5	4	.17(.11)	.21(.12)	.18(.11)	.15(.09)	.13(.08)	.26(.11)	.15(.07)	.09(.04)	.15(.09)	.15(.09)	.18(.12)	.29(.15)	.17(.10)	.14(.07)	.13(.06)	.13(.06)	.14(.06)	.14(.06)
4	5.5	5	.26(.08)	.32(.06)	.27(.06)	.24(.06)	.20(.07)	.32(.06)	.26(.06)	.19(.05)	.25(.08)	.25(.08)	.26(.07)	.29(.08)	.24(.07)	.24(.07)	.25(.08)	.25(.08)	.26(.08)	.26(.08)
5	5.5	5.5	.24(.06)	.27(.08)	.25(.06)	.23(.04)	.20(.05)	.28(.06)	.24(.05)	.19(.04)	.25(.07)	.25(.07)	.22(.06)	.20(.06)	.20(.05)	.23(.05)	.25(.06)	.25(.06)	.26(.06)	.27(.06)
6	6	1	.14(.06)	.11(.05)	.14(.05)	.15(.06)	.15(.06)	.19(.05)	.13(.04)	.09(.03)	.14(.06)	.14(.06)	.13(.05)	.12(.05)	.12(.05)	.13(.05)	.14(.05)	.14(.05)	.15(.05)	.17(.05)
7	6	3	.15(.06)	.15(.08)	.15(.06)	.15(.05)	.14(.05)	.20(.06)	.14(.04)	.09(.03)	.15(.06)	.15(.06)	.14(.07)	.16(.10)	.12(.05)	.13(.05)	.14(.05)	.14(.05)	.15(.05)	.17(.04)
8	6	4	.18(.08)	.22(.10)	.19(.08)	.16(.06)	.14(.05)	.25(.08)	.17(.05)	.11(.04)	.17(.07)	.17(.07)	.18(.09)	.24(.13)	.16(.08)	.15(.06)	.16(.06)	.16(.06)	.17(.05)	.19(.05)
9	6	5	.25(.08)	.32(.06)	.27(.06)	.23(.06)	.19(.06)	.31(.06)	.25(.06)	.19(.05)	.25(.08)	.25(.08)	.25(.08)	.28(.08)	.23(.07)	.23(.08)	.24(.08)	.24(.08)	.25(.08)	.26(.07)
10	6	6	.17(.06)	.16(.09)	.17(.07)	.18(.05)	.17(.04)	.20(.07)	.17(.06)	.14(.05)	.20(.06)	.20(.06)	.14(.05)	.11(.05)	.13(.04)	.16(.05)	.19(.05)	.19(.05)	.21(.05)	.22(.05)
11	6.5	1	.23(.10)	.27(.10)	.23(.10)	.21(.10)	.19(.10)	.32(.08)	.22(.08)	.15(.07)	.24(.10)	.24(.10)	.22(.11)	.14(.06)	.17(.08)	.21(.09)	.24(.09)	.24(.09)	.26(.09)	.34(.08)
12	6.5	3	.22(.10)	.27(.09)	.23(.09)	.20(.09)	.18(.09)	.30(.07)	.22(.08)	.15(.07)	.23(.09)	.23(.09)	.21(.10)	.15(.07)	.17(.07)	.20(.08)	.23(.08)	.23(.08)	.26(.08)	.32(.07)
13	6.5	4	.22(.09)	.28(.08)	.23(.08)	.20(.08)	.18(.08)	.29(.06)	.22(.07)	.15(.07)	.23(.09)	.23(.09)	.21(.09)	.17(.08)	.17(.07)	.20(.08)	.23(.08)	.23(.08)	.25(.08)	.31(.06)
14	6.5	5	.23(.08)	.30(.07)	.25(.06)	.21(.06)	.17(.06)	.29(.06)	.24(.07)	.17(.06)	.24(.08)	.24(.08)	.23(.08)	.21(.07)	.19(.07)	.21(.07)	.24(.08)	.24(.08)	.25(.08)	.29(.06)
15	6.5	6.5	.10(.05)	.06(.05)	.09(.05)	.11(.05)	.12(.04)	.12(.06)	.10(.05)	.08(.04)	.12(.05)	.12(.05)	.07(.04)	.05(.03)	.06(.03)	.09(.04)	.11(.05)	.11(.05)	.13(.05)	.14(.05)

Note: *SD* over simulations given in brackets. The darker the shade of grey, the lower the proportion of misclassifications, and the higher the decision accuracy.

Table C4: Mean Positive Predictive Value for Decision Rules and Factors Test Correlation, Test Reliability, Number of Tests, and Number of Reexaminations

Decision Rule	GPA	Minimum	Mean Positive Predictive Value	Average Test Correlation										Average Test Reliability										Number of Tests					Number of Reexaminations				
				0.1	0.3	0.5	0.7	0.4	0.6	0.8	0.8	8	12	0	1	2	3	4	max	0	1	2	3	4	max								
1	5.5	1	.95(.03)	.99(.01)	.96(.02)	.94(.03)	.92(.03)	.94(.04)	.95(.03)	.96(.02)	.95(.03)	.96(.02)	.95(.03)	.95(.03)	.96(.03)	.97(.02)	.97(.02)	.97(.02)	.96(.03)	.95(.03)	.95(.03)	.96(.03)	.95(.03)	.94(.04)	.93(.04)								
2	5.5	3	.96(.03)	.98(.01)	.97(.02)	.95(.02)	.94(.03)	.97(.02)	.95(.03)	.96(.02)	.96(.02)	.96(.02)	.96(.03)	.96(.03)	.97(.02)	.98(.01)	.98(.01)	.97(.02)	.96(.02)	.96(.02)	.96(.02)	.96(.02)	.95(.03)	.95(.03)	.95(.03)								
3	5.5	4	.94(.02)	.93(.02)	.94(.02)	.95(.02)	.95(.02)	.95(.03)	.94(.02)	.95(.02)	.95(.02)	.95(.02)	.95(.02)	.95(.02)	.94(.03)	.97(.02)	.97(.02)	.95(.02)	.94(.02)	.94(.02)	.94(.02)	.94(.02)	.93(.02)	.93(.02)	.93(.02)								
4	5.5	5	.81(.10)	.68(.08)	.78(.06)	.85(.04)	.91(.04)	.79(.12)	.80(.10)	.83(.08)	.83(.08)	.82(.09)	.82(.09)	.79(.12)	.79(.12)	.86(.10)	.86(.10)	.83(.10)	.80(.10)	.80(.10)	.80(.10)	.79(.10)	.78(.10)	.78(.10)	.78(.10)								
5	5.5	5.5	.68(.16)	.49(.11)	.64(.08)	.76(.07)	.85(.06)	.64(.18)	.68(.15)	.73(.12)	.73(.12)	.70(.14)	.70(.14)	.67(.17)	.67(.17)	.77(.16)	.77(.16)	.72(.15)	.68(.15)	.68(.15)	.66(.15)	.64(.15)	.64(.15)	.64(.15)	.64(.15)								
6	6	1	.87(.06)	.91(.04)	.87(.05)	.86(.06)	.85(.07)	.83(.06)	.87(.05)	.91(.04)	.91(.04)	.87(.06)	.87(.06)	.88(.06)	.88(.06)	.93(.04)	.93(.04)	.90(.05)	.87(.05)	.86(.05)	.86(.05)	.85(.06)	.85(.06)	.82(.06)	.82(.06)								
7	6	3	.88(.05)	.92(.04)	.89(.05)	.87(.05)	.87(.06)	.87(.05)	.88(.05)	.91(.04)	.91(.04)	.88(.05)	.88(.05)	.89(.05)	.89(.05)	.94(.02)	.94(.02)	.91(.03)	.89(.04)	.87(.04)	.87(.04)	.86(.05)	.86(.05)	.84(.05)	.84(.05)								
8	6	4	.89(.04)	.89(.04)	.89(.04)	.89(.05)	.89(.05)	.88(.04)	.88(.05)	.90(.04)	.90(.04)	.88(.05)	.88(.05)	.89(.04)	.89(.04)	.94(.02)	.94(.02)	.91(.03)	.89(.03)	.87(.03)	.87(.03)	.86(.03)	.85(.04)	.85(.04)	.85(.04)								
9	6	5	.80(.10)	.68(.08)	.78(.06)	.85(.05)	.90(.04)	.78(.12)	.79(.10)	.82(.08)	.82(.08)	.81(.09)	.81(.09)	.79(.11)	.79(.11)	.86(.10)	.86(.10)	.82(.10)	.80(.10)	.80(.10)	.78(.09)	.77(.09)	.76(.09)	.76(.09)	.76(.09)								
10	6	6	.55(.20)	.31(.12)	.50(.11)	.65(.09)	.77(.08)	.50(.22)	.55(.19)	.62(.16)	.62(.16)	.57(.18)	.57(.18)	.54(.22)	.54(.22)	.67(.21)	.67(.21)	.60(.20)	.55(.19)	.55(.19)	.52(.18)	.50(.18)	.49(.18)	.49(.18)	.49(.18)								
11	6.5	1	.73(.11)	.70(.11)	.72(.11)	.75(.11)	.76(.11)	.64(.09)	.74(.09)	.81(.08)	.81(.08)	.72(.11)	.72(.11)	.74(.11)	.74(.11)	.85(.08)	.85(.08)	.78(.09)	.74(.10)	.71(.09)	.71(.09)	.68(.09)	.63(.08)	.63(.08)	.63(.08)								
12	6.5	3	.74(.10)	.71(.10)	.73(.10)	.75(.10)	.77(.10)	.67(.09)	.74(.09)	.81(.08)	.81(.08)	.73(.10)	.73(.10)	.75(.11)	.75(.11)	.86(.07)	.86(.07)	.79(.08)	.75(.09)	.72(.08)	.72(.08)	.69(.08)	.64(.07)	.64(.07)	.64(.07)								
13	6.5	4	.75(.10)	.71(.10)	.75(.09)	.77(.09)	.78(.09)	.70(.09)	.74(.09)	.81(.08)	.81(.08)	.74(.10)	.74(.10)	.76(.10)	.76(.10)	.87(.06)	.87(.06)	.80(.07)	.75(.08)	.72(.08)	.72(.08)	.70(.07)	.66(.06)	.66(.06)	.66(.06)								
14	6.5	5	.74(.11)	.62(.09)	.72(.08)	.79(.08)	.82(.08)	.70(.12)	.73(.11)	.78(.09)	.78(.09)	.73(.10)	.73(.10)	.74(.12)	.74(.12)	.84(.10)	.84(.10)	.78(.10)	.74(.10)	.71(.10)	.71(.10)	.69(.09)	.66(.08)	.66(.08)	.66(.08)								
15	6.5	6.5	.44(.22) ¹	.18(.13) ²	.37(.12)	.53(.11)	.68(.10)	.37(.23) ³	.43(.22) ⁴	.52(.19) ⁵	.52(.19) ⁵	.45(.21) ⁶	.45(.21) ⁶	.43(.24) ⁷	.43(.24) ⁷	.56(.25) ⁸	.56(.25) ⁸	.49(.22)	.44(.21)	.40(.20)	.38(.19)	.37(.19)	.37(.19)	.37(.19)									

Note: SD over simulations given in brackets. The darker the shade of grey, the lower the proportion of misclassifications, and the higher the decision accuracy.

¹N=71952 instead of N=72000. ²N=17952 instead of N=18000. ³N=23994 instead of N=24000. ⁴N=23985. ⁵N=23973. ⁶N=35999 instead of N=36000.

⁷N=35953. ⁸N=11952 instead of N=12000.

In this appendix, the results of the simulation study are displayed per factor in Table C1 to C4. In addition, the next paragraphs discuss these findings in detail by focusing on the observed effects for each factor separately. First, the direction of the influence of each factor on the decision accuracy is discussed, after which the strength of the influence is described.

Influence of Average Test Correlation

The average test correlation columns in Table C1 to C4 show the mean values for the four levels of test correlations that were simulated. Overall, the direction of the influence of the test correlation on the proportion of misclassifications and positive predictive value depended on the required GPA and minimum grade. Although mostly a negative influence of the test correlation on the proportion of misclassifications was observed, it was positive for compensatory decision rules in which the GPA was low. Furthermore, increasing the test correlation increased the positive predictive value if the minimum grade was high and decreased the positive predictive value if the minimum grade was low. The test correlation had a negative influence on the false negative and false positive rate.

Similarly, the size of the observed effect of the test correlation depended on the specific decision rule applied. The test correlation strongly influenced the false positive rate. Here, increasing the minimum grade or GPA, resulted in a decrease in the influence of the test correlation. Secondly, the test correlation had a large influence on the false negative rate and the proportion of misclassifications. For the false negative rate, increasing the minimum grade strongly increased the negative influence of the test correlation. Increasing the GPA slightly increased the influence of the test correlation as well. Similarly, for the proportion of misclassifications, the negative influence of the test correlation was largest as the minimum grade increased. Interestingly, the influence of the test correlation on the positive predictive value was small except when the minimum grade was high. In this case, the influence of the test correlation was very large.

Influence of Average True Score Test Reliability

Evaluating the direction of the influence of the test reliability at an average true score on the decision accuracy shows that the direction only depended on the specific decision rule for the false positive rate. Here, mostly a negative influence was observed, except when the minimum grade was low or the GPA was high. For these decision rules, increasing the test reliability at an average true score resulted in an increase in the false negative rate. Furthermore, the test reliability had a negative influence on the proportion of misclassifications and the false negative rate; increasing the test reliability resulted in fewer classification errors. Contrary, test reliability had a positive influence on the positive predictive value.

Evaluating the size of the influence of the test reliability at an average true score, shows that the reliability had a medium to large influence on the false positive rate. Here, the influence was strongest if a compensatory decision rule was applied. Furthermore, the influence of the test reliability decreased at a higher minimum grade. Here, the influence of the minimum grade on the influence of the test reliability was smaller if the required GPA was low. The test reliability also had a large influence on the false negative rate. Here, the influence of the test reliability increased as the minimum grade increases as well, especially at a low required GPA. Furthermore, a large influence of the test reliability on the proportion of the misclassification was observed. This influence increased as the minimum grade increased for rules in which the required GPA was low. Contrary, the influence of the test reliability on the proportion of misclassifications decreased as the minimum grade increased when the required GPA was high. Also, the positive influence of the GPA on the influence of the test reliability was strongest for a compensatory rule. Finally, the test reliability had a small to medium influence on the positive predictive value. Increasing the required GPA or increased the influence of the test reliability, while increasing the minimum grade slightly decreased the influence of the test reliability on the positive predictive value.

Influence of Number of Tests

The number of tests columns in Table C1 to C4 show the mean values for the simulations of curricula with 8 or 12 tests. As can be seen, the direction of the influence of the number of tests on the proportion of misclassifications and the positive predictive value depended on the required minimum grade and GPA. A negative influence of the number of tests on the proportion of misclassification was observed when the minimum grade was low or when the required GPA was high. Additionally, mostly a positive influence of the number of tests on the positive predictive value was observed. Only if the minimum grade was high and the GPA low, a negative influence was observed. Overall, increasing the number of tests increased the false negative rate and decreased the false positive rate.

The results in Table C1 to C4 show that the size of the influence of the number of tests only had a small influence on the decision accuracy relative to the other measures. Mostly, it influenced the false positive rate. Additionally, increasing the GPA slightly increased the moderate influence of the number of tests. Secondly, there was a small influence of the number of tests on the false negative rate if the minimum grade was high. If the minimum grade was low, there was no effect of the number of tests. Finally, the influence of the number of tests on the proportion of misclassifications and positive predictive value was very small and not consistently influenced by the specific decision rule applied.

Influence of Number of Reexaminations

Finally, the last columns in Table C1 to C4 display the decision accuracy measures for the simulations with the different number of reexaminations. Looking at the direction of the influence of the number of reexaminations on the proportion of misclassifications shows that the influence depended on the specific decision rule. Specifically, if the minimum grade was high and the GPA low, a negative influence of the number of reexaminations on the proportion of misclassifications existed. For the remaining decision rules, increasing the number of reexaminations, resulted in an increase in the proportion of decisional errors. Furthermore, the number of

reexaminations had a negative influence on the false negative rate and the positive predictive value and a positive influence on the false positive rate.

Focusing on the size of the influence of the number of tests shows that the number of reexaminations mostly influenced the false positive rate. Here, the influence decreased as the minimum grade increased. Also, the number of reexaminations had a very large influence on the false negative rate, especially for the rules in which a high minimum grade was required. For both decision accuracy measures, the minimum grade was more important in determining the size of the influence of the number of reexaminations than the GPA. Furthermore, the number of reexaminations had a medium to large influence on the positive predictive value. Here the influence was largest as the decision rule required a high GPA. The number of reexaminations also had a small to large influence on the proportion of misclassifications. Here, the influence was highest if the required GPA was high and the minimum grade was low. Notably, the influence of the GPA on the size of the influence of the number of tests on the proportion of misclassifications and the positive predictive value was larger than the influence of the minimum grade.

3

Allowing Course Compensation in Higher Education: A latent class regression to evaluate performance on a sequel course

This chapter is submitted as:

Yocarini, I. E., Bouwmeester, S., Smeets, G., & Arends, L. R. (submitted). Allowing Course Compensation in Higher Education: A latent class regression to evaluate performance on a sequel course.

Abstract

In Dutch higher education, an academic dismissal policy is in place in which a compensatory decision rule might be used to assign study credits. In this study, the consequences of allowing compensation are evaluated by examining performance on a second-year sequel course that builds on material from a first-year precursor course. Up to now, differences in the consequences of compensation on student performance across latent groups of students were not considered. This study uses a latent class regression model to distinguish between students who portray different unobserved study processes. Data from a Psychology and a Law undergraduate curriculum were used and latent classes were formed based on similar patterns of first-year averages, variability in first-year grades, the number of compensated first-year courses, and the number of retakes in the first year. Results show that students can be distinguished by three latent classes. Although the first-year precursor course is compensated in each of these latent classes, low performance on the precursor course results in low performance on the second-year sequel course for Psychology students who belong to a class in which the average across first-year courses is low and the average number of compensated courses and retakes are high. For these students, compensation on a precursor course seems more likely to relate to insufficient performance on a sequel course.

Keywords: compensation, higher education, latent class regression, academic performance.

Introduction

Student success in higher education is an important issue. This is underlined by the goal set in the Europe 2020 strategy to have at least 40% of 30-34-year-olds complete higher education. Reducing student dropout and increasing study completion rates is hereby one of the main strategies to improve student success (Vossensteyn et al., 2015). Similarly, in the US an increased focus on student attainment exists as US colleges and universities try to prevent high dropout rates (Barefoot, 2004).

Improving study success in higher education has been approached in many ways, using different interventions (Sneyers & De Witte, 2018). One successful intervention, which we focus on in this study, is the use of an academic dismissal (AD) policy, that is, a performance-based selection mechanism through which students may be dismissed from an academic program (in The Netherlands referred to as the binding study advice [BSA]; Sneyers & De Witte, 2018). In an AD policy, students' progress is evaluated, for example, after the first year of the bachelor to assess whether the requirements to continue their studies are met. In making this decision, different decision rules may be applied by the higher education institutions. These rules may vary with respect to the number of study credits required. Traditionally, a conjunctive decision rule is applied, in which students either pass or fail an individual course and study credits are assigned to individual grades. Alternatively, a compensatory decision rule may be applied in which students are allowed to compensate, within boundaries, a low score on one course with a high score on another course. In this situation, students receive study credits based on their average score. In this study, the latter approach is examined as we aim to evaluate performance of students who are allowed to compensate courses in the academic dismissal policy.

Allowing Course Compensation

Different reasons may motivate the implementation of a compensatory decision rule instead of a conjunctive decision rule. First, compensation might be allowed to improve students' grade goals and motivation to perform well on tests as it pays off to get a grade that is as high as possible when the average grade serves as the selection instrument, instead of just passing a test. Indeed, Kickert, Stegers-Jager, Meeuwisse, Prinzie, and Arends (2017) showed that this might increase students' study efforts.

Furthermore, compensation might be a consequence of the intention to decrease students' procrastination by limiting the number of retakes allowed. When allowing compensation, a failing grade on a course does not necessarily need to be retaken and subsequent study delay might be decreased. By decreasing study delay, study success in terms of time-to-degree increases (Vermeulen et al., 2012). Third, students are trained for a profession that is compensatory by nature. In a job, an employee may compensate his or her lacking in one area by outshining others in another area of expertise, assuming that a minimum (Rekveld & Starren, 1994). Notably, lacking here is relative as an employee holding a diploma should have a minimum degree of competency in each area as stated in the end qualification requirements for the diploma. Fourth, the argument that the average grade is more reliable than individual test scores and consequently guards against making incorrect decisions in the AD policy, may motivate the implementation of a compensatory rule (De Gruijter, 2008; Vermeulen et al., 2012).

Whereas there may be several reasons to implement compensation, there is also critique on allowing course compensation in higher education. Opponents argue that compensation might result in unfavorable study behavior and processes that might result in lower academic performance. By using a decision rule in which compensation is allowed, strategic study behavior might be elicited in which students may make certain strategic study choices in terms of their resource allocation (such as in time and effort; see e.g., Van Naerssen, 1970) that might hamper their academic performance. Certain resource allocation strategies, such as for example, focusing more on easier courses and less on difficult ones, within a compensatory decision rule might possibly create problematic hiatuses (i.e., gaps) in students' knowledge, thereby decreasing educational quality (Arnold, 2011). Specifically, this concern applies to the situation in which courses accumulate on knowledge obtained in previous courses (so-called sequel courses). By allowing compensation, students might not obtain sufficient knowledge to perform well on a sequel course and possibly graduate with hiatuses. Furthermore, opponents argue that students who are allowed to compensate courses might strategically compensate more complicated courses with easier ones (Rekveld & Starren, 1994).

To prevent such undesired situations, an educational program should be designed accordingly and, for example, form clusters in which compensation is allowed (Rekveld & Starren, 1994). By forming clusters based on, such as for example, the course content or the course difficulty level or by giving difficult courses more weight, one could avoid having graduates with hiatuses in their knowledge. Consequently, decision rules in an educational context are rarely fully compensatory (Douglas & Mislevy, 2010). Rather, there are some conjunctive aspects included in which a minimum level of performance is required. Furthermore, we evaluated in an earlier study the fourth argument (that the average grade is more reliable) motivating the implementation of compensation in a simulation study in which relevant higher education contexts were mimicked (Yocarini, Bouwmeester, Smeets, & Arends, 2018). By varying the conjunctive and compensatory aspects in a decision rule, the accuracy of each decision was evaluated. From these results it can be concluded that the compensatory decision rule in a higher education context is not always more reliable than a conjunctive rule. Instead, its relative accuracy depends on the context (i.e., the test reliabilities, correlation between tests, and the number of resists allowed). Consequently, choosing a specific decision rule should involve the evaluation of the decision accuracy as well as the characteristics of the tests that are combined. For conjunctive decision rules, false negatives, those students who failed but are truly competent enough to pass, are more prominent than false positives, students who passed but are not sufficiently skilled yet. For compensatory decision rules the reverse is true and more false positives occur compared to false negatives. The discussion of whether to allow course compensation is therefore also a discussion of preferring false negatives over false positives or vice versa.

With the increased focus on student success in higher education, the debate on allowing compensation in an AD policy has gotten more attention as well (see e.g., Smits, Kelderman, & Hoeksma, 2015). Several studies have focused on the evaluation of compensatory decision rules. For example, Arnold (2011) evaluated the consequences in an Economics bachelor program and showed that whether the compensated course grade was obtained after one or multiple tries (i.e., was retaken) was important for later performance; that is, the number of retakes were negatively

related to performance on the sequel course. Whereas the pros and cons of compensation have been discussed from both a scientific and policy perspective, up to now the debate has not yet touched upon the question whether the discussion of the consequences of compensation (e.g., possible hiatuses in knowledge) applies to specific groups of students. Although this point was raised by Smits et al. (2015), no study has evaluated this. The higher education student population is diverse, containing students with varying levels of cognitive abilities who may portray different study strategies. Students' grades and choices to compensate or retake courses in a curriculum may vary as a consequence.

The Present Study

The purpose of this study is to take into account these unobserved study processes for students who are allowed to compensate courses. In this way, our study extends previous studies that have studied the relation between performance on a first-year precursor and second-year sequel course (e.g., Arnold, 2011). To our knowledge, there is no published study that evaluated differences between groups of students that show similar study processes. As these groups have yet to be detected, a latent class model could be applied to explore whether groups of students exist who are characterized by similar unobserved choices in study resource allocation.

Variables that were used to distinguish these latent classes are the first-year average grade, the variation in first-year grades, the number of courses that were compensated, and the number of courses that were retaken. Courses are qualified as compensated when the course grade (e.g., 5.0, on a 1-10-point scale, as is common in Dutch higher education) is below the required average grade set in the compensatory decision rule (e.g., 6.0). These variables were selected as these are expected to be able to make a distinction between groups of students who make different (unobserved) choices with regards to their study resource allocation in the first year of the bachelor, and which groups may display different relations between first-year precursor and second-year sequel course performance. Although students with very high and very low cognitive abilities will probably perform similarly under different decision rules (i.e., pass or fail in either situation), students with average abilities just above the cut-score are of most interest in the compensation discussion. Here, two students (let's

call them Ann and Peter) with similar average grades might be very different in their underlying latent true scores per course. Where Ann might have quite similar true scores across courses and obtained her average grade by consistently scoring around this average, Peter might have high variation in his true scores across courses, resulting in high variability in his grades. Study choices as to which course to compensate or retake and to alter study resource allocation accordingly, might be more useful to Peter compared to Ann. Consequently, performance on a second-year sequel course is more likely to be low for Peter compared to Ann, as he is more likely to have had a low true score, and possibly compensated the first-year precursor course. By compensating this course, it might be that Peter lacks knowledge of the first-year material required in the second-year sequel course, resulting in lower academic performance in the second-year compared to Ann. Up to now, the consequences of compensation have been studied as being similar for Ann and Peter by not making a distinction in latent student groups. However, what applies to the aggregate does not necessarily mean the finding holds in general and applies to each and every student (Hamaker, 2012).

The overall aim of this study is to evaluate the consequences of course compensation by evaluating performance on a second-year sequel course across different latent groups. Specifically, the relation between the first-year precursor course and second-year sequel course performance is allowed to vary across latent student groups who are characterized by a similar pattern in first-year grades, variability in their first-year grades, the number of compensated first-year courses, and the number of retakes in the first year. For this purpose, a latent class regression model (Vermunt & Magidson, 2002; Wedel & DeSarbo, 1994) is applied on data from a Psychology university bachelor in the Netherlands in which compensation is allowed within boundaries in the first year of the bachelor. As a second purpose, the generalizability of the results is assessed by replicating the analyses on data from a Dutch Law university bachelor program in which course compensation is also allowed. Here, a latent class model has the advantage that groups of students can be formed which show similar grade characteristics and study choices without a priori assumptions about the specific formation of these groups (e.g., in terms of the number of classes or class sizes;

Vermunt & Magidson, 2002). In this way, the regression coefficients are allowed to vary across heterogeneous groups (Wedel & DeSarbo, 1994).

Method

Sample

Test scores from students' first- and second-year courses in a Psychology bachelor program at a Dutch university were used. This data was obtained from the Erasmus Educational Research (EER) database. Specifically, cohorts were included in which compensation was allowed and in which a sequel second-year course was present, meaning that the material in the second year built upon the material in a first-year course. For this selection, the content of the courses was considered by consulting the exam regulations and course descriptions, as well as the course coordinators or program executives. Overall, only students who passed the AD policy requirements to continue to the second year of the bachelor were included. This implies that each included student obtained a grade on each first-year course (i.e., no missing values were allowed). Dutch students are graded on a scale of 1 to 10, with 5.5 serving as the cut-score for a passing grade. Compared to American grading scales, a grade of 8 or higher corresponds to an 'A', a grade of 7 to a 'B+', and a grade of 6.5 to a 'B' (Nuffic, 2009). Following these selection criteria, the cohorts 2011 to 2015 were selected, including 1077 Psychology students in the study. Psychology students were required to score 6.0 on average (rounded from 5.95 on a 1-10-point scale) over eight courses with a minimum required grade of 4.0 on each individual course. Of these eight tests only two were allowed to be retaken once. Overall, one course combination existed in which the second-year course very clearly and explicitly built on first-year material for these Psychology cohorts, namely that of Statistics I in the first year and Statistics II in the second year.

To assess the generalizability of our findings in the Psychology data, students' grades from the Law bachelor were selected to replicate the analyses. If one would expect similarities in latent classes across study programs, these similarities are expected to be most pronounced in a study program that is most similar to the Psychology bachelor program. For this reason, law students were selected as the Law curriculum

has a similar organisation (i.e., eight consecutive courses that each have a similar number of course credits) and didactic approach (i.e., problem-based learning), is about similar in size, and employs an AD policy decision rule that is most similar to that of the Psychology bachelor. Following a similar exclusion procedure, the cohorts 2012 to 2015 were selected, including 1120 Law students. Law students were required to score 6.0 on average (unrounded) over eight courses with a minimum required grade of 4.5 on each individual course. Similarly, two out of these eight tests were allowed to be retaken once. Within the Law curriculum several course combinations existed in which the second-year sequel course built on the first-year precursor course. The combination in which the first-year course was compensated most often was selected: Introduction to constitutional and administrative law in the first year and Constitutional law in the second year. See Table 1 and Table 2 for descriptive information of the sample per study program.

Table 1: Descriptive Statistics Continuous Variables

Study program	Course	Year	Variable	Mean	Median	<i>SD</i>	Min	Max
Psy		1	Yearly average ¹	6.77	6.60	0.65	5.95	9.25
		1	Yearly <i>SD</i>	0.89	0.88	0.24	0.28	1.67
	Statistics I	1	Course grade	6.40	6.40	1.28	4	10
	Statistics II	2	Course grade ²	6.66	6.80	1.49	1	10
Law		1	Yearly average ¹	6.82	6.75	0.61	6	9
		1	Yearly <i>SD</i>	0.80	0.79	0.22	0	1.51
	Intro cons and admin. law	1	Course grade	6.54	7	1.03	5	10
	Con. law	2	Course grade ³	6.26	6	1.08	3	10

¹The average includes students who received a grade on all first-year courses. ²Note that second-year courses do not have the requirements in the AD policy as in the first year, so grades run from 1 to 10, *NA* = 77. ³*NA* = 192.

Table 2: Descriptive Statistics Categorical Variables

Study Program	Course	Year	Variable	0	1	2	3	4	5
Psy		1	Number of compensated courses ¹	25.5% (275)	22.9% (247)	20.3% (219)	20.3% (219)	8.6% (93)	2.2% (24)
		1	Number of resits	71.7% (772)	11% (119)	17.3% (186)			
	Stat I	1	Course compensated	62.1% (669)	37.9% (408)				
		1	Course retaken	90.3% (972)	9.7% (105)				
Law		1	Number of compensated courses ¹	52.3% (586)	27.1% (304)	16.8% (188)	3.4% (38)	0.4% (4)	
		1	Number of resits	50.7% (568)	34.6% (387)	14.7% (165)			
	Intro con. and admin. law	1	Course compensated	83.2% (932)	16.8% (188)				
		1	Course retaken	8.9% (995)	11.2% (125)				

¹Compensated courses are courses on which the grade was below the required average of 6.0. This thus differs from the number of insufficient grades (below the Dutch cut-off of 5.5).

Statistical Analyses

To assess the relation between the first-year precursor course and second-year sequel course grade across different groups of students a latent class (LC) regression model is performed. In a LC regression model, a single dependent variable exists that is assumed to be class dependent, which is the grade on a second-year sequel course. In this way, unobserved heterogeneity with respect to the distribution of second-year grades can be described by the latent classes. Different variables are included in the latent class regression analysis, namely predictors that influence the dependent variable and variables that influence the latent variable (referred to as covariates; Vermunt & Magidson, 2013a). This results in the LC regression model:

$$(1) f(y|z_1^{cov} z_2^{cov} z_3^{cov} z_4^{cov}, z_1^{pred}) = \sum_{x=1}^K P(x|z_1^{cov}, z_2^{cov}, z_3^{cov}, z_4^{cov}) f(y|x, z_1^{pred}).$$

In this model, $f(y|x, z_1^{pred})$ denotes the distribution of the grades on the second-year course, given a student his or her class membership x and predictor value z_1^{pred} , which is the grade on the first-year precursor course. These second-year grades are allowed to be class dependent, where it is assumed that the probability of belonging to latent class x , of all K classes, depends on the values of the four covariates z_1^{cov} , z_2^{cov} , z_3^{cov} , and z_4^{cov} . In this study, the covariates include: the yearly average, variation in first-year grades, the number of compensated courses in the first year (i.e., the number of course grades below the required minimum grade), and the number of resits in the first year. In this way, latent classes are formed that depend on the first-year grade patterns as well as the grade on a second-year sequel course.

The LC regression was performed using Latent GOLD 5.0 (LG; Vermunt & Magidson, 2013a, see Appendix A for the syntax used). The number of classes were determined using various information criteria (IC): the Akaike information criterion (AIC), the Akaike information criterion 3 (AIC3), and the Bayesian information criterion (BIC). Generally, models with more parameters provide a better absolute fit (McCutcheon, 2002). However, each of these IC apply a different penalty on the log-likelihood statistic for the number of model parameters, sample size, or both (Nylund, Asparouhov, & Muthén, 2007). As a consequence, each IC might point towards different models as the relatively best fitting model (that for which the IC is lowest)

and thus the ICs might not be in agreement. For this reason, choosing the number of classes is one of the most difficult aspects of LC analyses and consequently a main research topic in LC analyses (Vermunt & Magidson, 2002). Therefore, in addition to the ICs, the classification error was reported, a proportion that indicates how well the latent classes are separated, that is, their distinctiveness (Vermunt & Magidson, 2013b). Furthermore, in addition to these fit indices, the parsimoniousness of the model, class size, and class interpretability were also taken into account in determining the number of classes. The first two are important to ensure the selected model is generalizable to other students as well. The latter, the substantive consideration of the model, is very important as the latent classes selected should substantively add value to our evaluation of the relation between first- and second-year grades.

Note that the *validation* ICs and classification errors were used to prevent overfitting. Using the sample ICs to validate the model might lead to overoptimistic assessments as the same data is used for validation as well as model building (Skrondal & Rabe-Hesketh, 2004). To remedy this overfitting, cross-validation was used in which a 10-fold validation procedure was applied. In this procedure, students were randomly assigned to one of the ten validation subsamples. The model of interest was then estimated ten times for which each time one of the ten subsamples was excluded in the estimation (Vermunt & Magidson, 2015). In each of the 10 folds, the estimated model parameters were used to obtain the log-likelihood and prediction statistics for the subsample left out of the estimation procedure. Consequently, validation ICs and the validation classification error were obtained by summing the statistics over the ten subsamples. This validation procedure was run in LG simultaneously with estimating our models. Additionally, the interpretability of the classes and class sizes were also taken into account in determining the number of latent classes.

Finally, after selecting the best fitting model, class assignments were used to describe the latent classes. Students were assigned to classes using the posterior probability to be in a latent class x given their response pattern y (i.e., their values on each of the four covariate variables). These posterior probabilities are obtained using the

estimated parameters from the latent classes in the LC regression analysis (Vermunt & Magidson, 2002):

$$(4)P(X = x|Y = y) = \frac{P(X=x)P(Y=y|X=x)}{P(Y=y)}$$

Consequently, a modal classification method was used in which students were assigned to the class with the largest posterior membership probability (Vermunt & Magidson, 2013b).

After this evaluation, the LC model was further validated by performing the same analysis on the Law data. In this way, the generalizability of our findings across study programs was assessed.

Results

LC Regression Analysis

Several LC regression models were fitted to the Psychology data, with an increasing number of classes. The validation fit statistics and proportion of classification errors for the LC models are displayed in Table 3.

Table 3: Validation Information Criteria and Classification Errors for Different LC Models for Psychology

Model	LL^1	BIC	AIC	AIC(3)	Number of Parameters	Proportion of Classification Errors
1- class	-1732.44	3485.61	3470.87	3473.87	3	0
2- class	-1632.06	3340.16	3286.12	3297.12	11	0.12
3- class	-1605.28	3341.91	3248.57	3267.57	19	0.2
4- class	-1616.03	3418.71	3286.07	3313.07	27	0.25
5- class	-1611.22	3464.39	3292.45	3327.45	35	0.22
6- class	-1615.40	3528.05	3316.80	3359.80	43	0.22

¹ LL = Log-Likelihood. Note that the Log-Likelihood slightly increases at the 4- and 6-class model, this is possible because of the holdout validation procedure used to estimate these values.

First, the BIC values in Table 3 are lowest for the two-class model (and only slightly smaller than the BIC of the three-class model). The AIC and AIC(3) values are lowest for the three-class model. Consequently, the two- and three-class models will be

considered. As can be seen by the proportion of classification errors of the different models, which indicate how distinct the latent classes are, these are higher for the three-class model and models with more classes. These classifications are high because some students are not easily classified in one of the three classes within the LC model, which seems mostly true for students classified in class one and two, as the mean classification error was slightly smaller in the third class. Investigating the class sizes shows that in both models there is one relatively larger class and one or two smaller classes. Where there is one class that has a high first-year average and one that has a low first-year average in the two-class model, the three-class model makes an additional distinction resulting in two classes with low and average first-year averages, each having different average number of compensated and retaken tests. Because of this additional distinction, which adds valuable information for our LC regression analysis, the three-class model was selected.

Table 4 shows the descriptive statistics for the covariate variables, the predictor variable, and the dependent variable for each class, as well as the class sizes. The parameter estimates of the covariates in the LC regression can be tested to see whether the influence of the covariate has a significant influence on the classes. The results for these tests showed that the yearly average, Wald statistic (2) = 29.16, $p < .001$, the yearly number of compensated courses, Wald statistic (2) = 8.30, $p = .016$, and the yearly number of retaken tests, Wald statistic (2) = 13.69, $p < .001$, were significantly different across the latent classes. As shown in Table 4, the first class can be interpreted as the students with an average low performance, a high average number of compensated courses and a high average number of retaken tests. About a quarter of the students are classified in class one. The second class, which is about half of the sample, are students with average performance levels, a moderate average number of compensated courses and a low number of retaken tests on average. Finally, the third class, about a fifth of the sample, consists of the high performing students, who have a low number of compensated courses on average and no retakes on average.

Table 4: Descriptive Statistics for the Three-Class Model for Psychology

Variable		Class		
		1	2	3
First-year average ¹	Average	6.20	6.69	7.61
	Standard deviation	0.20	0.43	0.62
	Max	7	8	9
First-year standard deviation ²	Average	0.88	0.91	0.86
	Standard deviation	0.22	0.23	0.25
	Number of compensations ³	Average	2.86	1.61
Number of retaken tests ⁴	Standard deviation	1.07	1.22	1.09
	Average	1.35	0.21	0.09
First-year precursor course grade ⁵	Standard deviation	0.87	0.51	0.33
	Average	5.64	6.35	7.32
	Standard deviation	1.03	1.14	1.25
Second-year sequel course grade	Max	9	9	10
	Average	4.98	6.68	8.47
	Standard deviation	1.26	0.90	0.61
Class size	Min	1	4	7
	Max	9	9	10
	N^6	0.23	0.56	0.21
		234	563	208

¹The minimum first-year average is 6 for all classes. ²The range of the standard deviation in first-year grades is similar across classes: from 0 to 2. ³The number of compensated courses ranges between 0 and 5 for each class. ⁴The number of retaken tests ranges between 0 and 2 for each class. ⁵The minimum grade is 4 across all classes. ⁶Sample size here is smaller as reported in the Results section as there were 72 missing values on the second-year grade.

Subsequently, the class-dependent relations between the first-year precursor grades on the second-year sequel course grades were evaluated. Here, a Wald test indicated the relation of the predictor with the second-year sequel course grade to be significant: Wald statistic (3) = 135.09, $p < .001$. When the precursor course grade was high, the grade on the second-year sequel course was high as well. Yet, a Wald test comparing the parameters across classes was not significant, Wald statistic (2) = 0.20, $p = .90$, indicating that the parameters did not significantly differ across classes. This implies that the positive relation that is found between the first-year precursor and second-year sequel grade did not vary statistically significant across different latent classes. Furthermore, the variances of the dependent variable, the second-year

sequel course grades, were significantly different across the three classes: Wald statistic (2) = 42.05, $p < .001$. This implies that the variability in second-year sequel course grades differs across the different latent classes, as shown by the standard deviation of the second-year precursor grade per class in the bottom half of Table 4. Because second-year grades may be more similar within a specific class, the parameters of the predictors might not differ across classes. Therefore, it is important to evaluate the average performance on the first-year precursor and second-year sequel course across classes.

The bottom half of Table 4 shows these estimates and as shown by the variability of the second-year grades, variability is high in the first class, but smaller in the second and third class. Importantly, in the first class, the average grade on the first-year precursor grade was just sufficient at the Dutch cut-off score of 5.5 (5.6), yet below the required average grade of 6.0 in the first-year compensatory decision rule. For these students in the first class, the average grade on the second-year sequel course was even lower and on average insufficient (5.0). For the second class, the average grade on the first-year course was sufficient and above the required average of 6.0 (6.4), while on average performance on the second-year sequel course was good (6.7). Finally, for the third class, which contains the high performing students, the average grade on the first-year course was high (7.3) and the average grade on the second-year course was very high (8.5). The ranges of the first-year and second-year grades show that while in every class (some) students compensated the first-year precursor course grade, the second-year sequel course was only compensated by (some) students from the first and second class.

LC Regression Analysis Law Curriculum

To assess whether the results generalize to other study programs, data from a Law bachelor program were analysed. A few differences exist in the analyses as the dependent variable, grades on a sequel second-year course, is treated as ordinal here as rounded grades are used in the Law program. The dependent variable therefore consists of only eight levels (grades ranging from 3 to 10). If the dependent variable would be considered continuous in this case, the resulting classes would be focused too much on these eight levels and not result in relevant and insightful latent classes.

As a consequence, second-year grades are treated as ordinal and the class dependent variances are not included in the model. Table 5 shows the validation information criteria and classification errors for the different LC models.

Table 5: Validation Information Criteria and Classification Errors for Different LC Models for Law

Model	LL^1	BIC	AIC	AIC(3)	Number of Parameters	Proportion of Classification Errors
1- class	-1321.90	2698.47	2659.81	2667.81	8	0
2- class	-1268.43	2680.35	2578.86	2599.86	21	0.11
3- class	-1269.22	2770.77	2606.44	2640.44	34	0.23
4- class	-1270.50	2862.15	2635.00	2682.00	47	0.17
5- class	-1268.05	2946.08	2656.10	2716.10	60	0.23
6- class	-1272.71	3044.23	2691.42	2764.42	73	0.19

¹ LL = Log-Likelihood. Note that the Log-Likelihood slightly increases at the 4- and 6-class model, this is possible because of the holdout validation procedure used to estimate these values.

The BIC, AIC, and AIC(3) values in Table 5 are all lowest for the two-class model indicating that this model fits the data best. Next, descriptive statistics for the covariate variables, the predictor variable, and the dependent variable for both classes, as well as the class sizes, are shown in Table 6. Tests to assess the influence of the covariates on the latent classes showed that the yearly average and the variation in first-year grades had a significant influence on the latent classes: Wald statistic (1) = 14.56, $p < .001$ and Wald statistic (1) = 8.76, $p = .003$, respectively. Students belonging to the first class had on average lower average grades than students in the first class. Also, variation in first-year grades was on average higher in the second class than in the first class. Furthermore, the grade on a first-year precursor course was a statistically significant predictor of grades on a second-year sequel course: Wald statistic (2) = 20.14, $p < .001$. Differences in the parameters across the two latent classes, however, were not statistically significant, Wald statistic (1) = 1.62, $p = .200$.

As the three-class model fitted best in the Psychology data, it is interesting to evaluate the three-class model, which had the second-best fit, in the Law data as well. In this three-class model, only the first-year average was significantly different across the latent classes: Wald statistic (2) = 20.56, $p < .001$. Furthermore, the relation

between the first-year precursor course and second-year sequel course grades was significant: Wald statistic (3) = 17.29, $p < .001$. The differences in the positive relation across the three classes were not statistically significant: Wald statistic (2) = 3.82, $p = .150$. These results are similar to those found in the three-class model of the Psychology data.

Table 6: Descriptive Statistics for the Two-Class Model for Law

Variable		Class	
		1	2
First-year average ¹	Average	6.57	7.5
	Standard deviation	0.41	0.52
	Max	8	9
First-year standard deviation ²	Average	0.78	0.87
	Standard deviation	0.20	0.22
	Max	1	2
Number of compensations ³	Average	0.91	0.23
	Standard deviation	0.93	0.49
	Max	4	2
Number of retaken tests ⁴	Average	0.79	0.25
	Standard deviation	0.74	0.50
	Max		
First-year precursor course grade ⁵	Average	6.30	7.22
	Standard deviation	0.90	1.07
	Max	9	10
Second-year sequel course grade	Average	5.78	7.32
	Standard deviation	0.84	0.76
	Min	3	6
	Max	8	10
Class size		0.69	0.31
N^6		286	642

¹The minimum first-year average is 6 for both classes. ²The minimum of the first-year standard deviation is 0 in both classes. ³The minimum number of compensations is 0 for both classes. ⁴The range of the number of retaken tests ranges between 0 and 2 for both classes. ⁵The minimum of the first-year precursor course grade is 5 in both classes. ⁶Sample size here is smaller as there were 192 students for who second year grade was missing.

Table 7: Descriptive Statistics for the Three-Class Model for Law

Variable		Class		
		1	2	3
First-year average ¹	Average	6.42	6.71	7.25
	Standard deviation	0.37	0.46	0.63
	Max	8	8	9
First-year standard deviation ²	Average	0.88	0.70	0.87
	Standard deviation	0.18	0.17	0.21
	Max	1	1	2
Number of compensations ³	Average	1.5	0.43	0.45
	Standard deviation	0.93	0.63	0.72
	Max	4	2	3
Number of retaken tests ⁴	Average	0.68	0.86	0.39
	Standard deviation	0.70	0.78	0.60
First-year precursor course grade ⁵	Average	6.04	6.54	6.96
	Standard deviation	0.95	0.84	1.12
	Max	9	9	10
Second-year sequel course grade	Average	5.2	5.86	7.26
	Standard deviation	0.61	0.61	0.70
	Min	4	3	6
	Max	6	7	10
Class size		0.25	0.35	0.40
N ⁶		230	327	371

¹The minimum first-year average was 6 for all classes. ²The minimum first-year standard deviation was 0 for all classes. ³The minimum number of compensations was 0 for all classes. ⁴The number of retaken tests ranges between 0 and 2 for each class. ⁵The minimum first-year precursor grade was 5 for all classes. ⁶Sample size here is smaller as there were 192 students for who second year grade was missing.

Table 7 shows the descriptive statistics for the three-class model for Law. In comparison to the two-class model in which the low performing class had a first-year average of 6.4, the three-class model has two classes that have a first-year average around this value. One class has a lower average first-year average of 6.2, while the average first-year average in the second class is slightly higher at 6.7. These two classes show differences in their average first-year precursor course grade as the average is just above the required average of 6.0 in the first class and around 6.5 on average in the second-class. Interestingly, for the lowest performing class, the average grade on the second-year sequel course is below the Dutch pass-fail cut-score of 5.5 (5.2) on average, while that of the second class is just below the required average

grade of 6.0 (5.9) on average. These results show that with three classes, a similar pattern is observed in the Law data as was in the Psychology data, where low performance on the first-year precursor course relates to an even lower performance on the sequel course on average for students whose performance in the first-year was low.

Discussion

In an academic dismissal (AD) policy, such as the binding study advice (BSA) in Dutch higher education, decisions are made about students' performance through the combination of multiple tests. In making these decisions, higher institutions may choose to allow course compensation when combining courses. The aim of this study was to evaluate performance on a second-year sequel course that builds on material from a first-year precursor course when students were allowed to compensate courses in the first-year of their undergraduate curriculum. This study extends on prior research by allowing the relation between a first-year precursor course and a second-year sequel course to be different for different latent groups of students. Data of an undergraduate Psychology program and a Law program were analysed using a latent class (LC) regression approach. These latent classes were expected to portray different unobserved study processes and choices with regard to students' study resource allocation within a complex compensatory testing system. Specifically, the latent classes were formed based on similar patterns in first-year averages, variability in first-year grades, the number of compensated first-year courses, and the number of retakes in the first year.

The best fitting latent class model for the Psychology data was a three-class model in which three groups of students could be distinguished in terms of their patterns in first-year averages, number of compensated courses, and number of retakes in the first year. Most students, a little over a half of the sample, belonged to the second class. Students in the second class performed well on average (6.7 on a 1- to 10-point scale), compensated about 1.6 courses (out of eight) on average (i.e., the number of grades below the required average grade of 6.0 in the complex compensatory decision rule), and on average did not need to retake many courses (about 0.2). Then, about a quarter of the sample belonged to the first class, which were the students with low

performance. These students on average had a first-year average just above the required average of 6.0 (6.2), had the highest number of compensated courses (2.9), and the highest number of retakes (1.4) on average. Finally, about a fifth of the sample belonged to the third class which were the best performing students. These students had a very high first-year average (7.6), a low number of compensated courses (0.6), and almost no retakes on average (0.1). For each of the classes, the relation between the average first-year grade, and the number of compensations or retakes is as expected, the higher the first-year average, the lower the number of compensated courses or retakes. Overall, these three latent classes distinguish students with low, moderate, and high performance.

Subsequently, testing the relation between the first-year precursor course grades and the second-year sequel course grades showed that differences in this relation were not statistically significant across classes. However, as this result is possibly due to the smaller variation in the dependent variable for the second and third class, evaluating the average grades on the first-year precursor course and the second-year sequel course across the different classes provided more insight. The average grade on the precursor course for the low performing students in the first class was below the required 6.0 average and just sufficient (5.6; the cut-score in Dutch education is at 5.5), for students from the second class the average grade was sufficient (6.4), and for the high performing students from the third class the average grade was high (7.3). The ranges of the precursor course across classes showed that in each class some students compensated the precursor course (i.e., scored below 6.0). The average grade on the second-year sequel course was insufficient (5.0) for the lowest performing students from the first class, good (6.7) for students from the second class, and very high (8.5) for the high performing students from class three. Here, the range of the sequel course grade showed that some students compensated the second-year sequel course in the first and second class, while no student needed to do so in the third class (i.e., the high performing students).

Overall, the positive relation that was found between the grades on the first-year precursor course and second-year sequel course is similar to findings from previous studies (e.g., Arnold, 2011). To assess the generalizability of the results, the model

was validated by replicating the analyses using data from a Law program. Here, a two-class model fitted best. In this model, the largest class (about two thirds) contained students with a moderate first-year average and the smaller class included students with a high first-year average and slightly more variation in their first-year grades. Similar to Psychology, differences in the relation between the precursor course grade and sequel course grade were not statistically significant, yet the average grade on the precursor course in the first class was moderate while performance on the sequel course on average was below the required average grade of 6.0 (5.8). For comparison purposes, the three-class model, the second-best fitting model in the Law data, was also evaluated. Here, a distinction was found similar to the Psychology data where the first class had a moderate first-year average, yet an average grade on the first-year precursor course that was just at the required average grade of 6.0. For this group, average performance on the second-year sequel course (5.2) was not only below the required average grade of 6.0 (as found in the two-class model results) but also below the Dutch pass-fail cut-score of 5.5. These results show that although the results in the Law data are not as extreme as in the Psychology data, that is, the average performance on the precursor and sequel course are slightly higher in the Law data, similar patterns are found in both datasets. The differences across the study programs might be due to differences in the course combinations, such that the second-year sequel course in the Psychology curriculum more extensively builds on the precursor course than in the studied course combination in the Law program. Also, grades on the Psychology courses might be lower because the courses might be considered more difficult within the curriculum than the courses evaluated in the Law program.

In this study, performance on a second-year sequel course was of interest because it explicitly builds on materials from a precursor course. Performance on these courses gives an indication of the consequences of allowing compensation in a first-year curriculum in terms of knowledge accumulation. The results from our latent class regression show that for the lowest performing class, which across study programs was about a quarter of the sample, performance on the second-year sequel course is on average a failing grade. These results suggest that allowing course compensation

might result in low performance on a sequel course when students' performance in the first-year is low as well (for Psychology students characterized by a low first-year average and a high number of compensated first-year courses and retakes). This seems to suggest that allowing compensation might on average have negative consequences for the students in the class with overall low first-year performance, such that performance on later courses is not sufficient. However, the results also show that the precursor course is compensated by students in each of the three classes, yet performance on the sequel course on average is higher (i.e., not insufficient) in the second and third class. This seems to suggest that these students (who compensated the precursor course yet had higher first-year performance than students in class one) are able to accumulate knowledge and skills on other courses that may transfer to the sequel course, resulting in sufficient performance.

Overall, the LC regression showed the data to be summarized best by three distinct latent classes. Whereas a positive relation between the precursor and sequel course was observed for all classes, this is the best fitting relation on average and does not imply this relation occurs for all students in the class. As policy makers might want to obtain guidelines for allowing compensation or not, students' performance in the low performing class in the Psychology dataset was further explored to see if patterns could be identified for which performance on the second-year course is more likely to be low. The exploration shows that the second-year sequel course grade in the first class is on average (i.e., the lowest performing students) lower when students in this class did not retake a test (26.5% of class 1, they had an average of 4.45 with *SD* of 0.88) or just had one test retaken (12% of class 1, they had an average of 4.73 with *SD* of 0.99) compared to those who had two retaken tests (61.5% of class 1, they had an average of 5.26 with *SD* of 1.37). Taken together, these patterns suggest that students who accumulated knowledge by retaking first-year tests or by having moderate to high performance on other first-year courses, might have gained additional knowledge that helped them perform well on the sequel course (even if performance on the first-year precursor course was low). Note that this elaboration is only meant to explore trends, in the LC regression only one group of low performing

students was identified and further trends should therefore be interpreted with caution, as these might, for example, not be large enough to be relevant.

Furthermore, these results suggest that first-year course compensation does not necessarily result in lower performance on a second-year sequel course. However, for the low performing group of students, compensation showed to have negative influences such that performance on a second-year sequel course was more likely to be insufficient. In observing this, it is important to keep in mind that this study only evaluated performance of students who were allowed to compensate courses in their first-year curriculum. Especially, given the debate on whether to allow compensation or to use a traditional conjunctive decision rule in the first-year of the bachelor program, the discussion of our results is consequently limited to students who were allowed to compensate and retake a maximum of two tests. That compensation might have negative influences for a group of students does not imply that students in a conjunctive system might not have low performance on a sequel course when their overall performance is low and around the cut-score as well. However, even if these relations between first-year precursor course grades and second-year sequel course grades are similar across testing systems, the group for which performance might be low on sequel courses would be larger when the minimum required grade on a course is lower (as is often the case in compensatory systems). Furthermore, the results from our study are limited as empirical data is used, including higher education tests. Although these tests meet a specific quality level, tests in higher education are not always able to discriminate highly among students (see e.g., Brown & Abdulnabi, 2017; DiBattista & Kurzawa, 2011) and consequently our results are limited by the quality of the tests that were used.

In this study, a first exploration of differences in patterns in students' study results was done. These patterns might be indicative of differences in study processes. As proponents of course compensation in an AD policy believe compensation to positively influence students' study processes, it would be interesting for future studies to focus explicitly on these study processes. A good starting point for studying these processes could be an evaluation of how students allocate their study time. Here, experienced based sampling methods (also known as ecological momentary assessment) might

provide a convenient method for measuring study time allocation. Ideally, these choices in study time allocation would be evaluated across different testing programs (i.e., compensatory or conjunctive) and in response to different curriculum aspects such as assessments and retakes. Also, it would be interesting to expand the model fitted in this study to different study programs to assess its generalizability, as well as to future students to assess its predictive ability. This latter aspect would be interesting to policy makers as one would like to identify students who will likely have a low performance on later courses in time. Based on our findings, which are limited only to our compensatory decision rule, it seems that students whose first-year performance is low and consequently compensate and/or fail the precursor course in the first-year have a higher likelihood to have low performance on the sequel course as well. Consequently, these students might require additional attention within a compensatory decision rule in the first-year of a curriculum to prevent gaps in knowledge, late drop-out, or increased time-to-degree.

Appendix A. Latent GOLD syntax

```
options
  maxthreads=all;
  //this option ensures that all computer cores are
  employed in the model computations.
  algorithm
    tolerance=1e-008 emtolerance=0.01 emiterations=500
nriterations=100;
  //latentGOLD uses both the expectation maximization (EM)
  and the Newton-Raphson (NR) algorithm, alternating the
  two to obtain the optimal model parameter estimates.
  First the EM is used up till the maximum number of EM
  iterations (emiterations) or the EM convergence
  criterion (emtolerance) is reached. Then, NR iterations
  are used till the maximum number of NR iterations
  (nriterations) or convergence tolerance) is reached. Here,
  we doubled the default values to ensure convergence to a
  global optimum.
startvalues
  seed=0 sets=16 tolerance=1e-005 iterations=100;
  //latentGOLD generates random start values automatically,
  thereby generating multiple sets to avoid convergence to local
  minima. Iterations here were doubled compared to the default of
  50.
bayes
  categorical=1 variances=1 latent=1 poisson=1;
montecarlo
  seed=0 sets=0 replicates=500 tolerance=1e-008;
quadrature nodes=10;
missing excludeall;
output
  parameters=effect betaopts=wl standarderrors profile
  probmeans=posterior
  bivariateresiduals estimatedvalues=regression
  validationLL;
  //validationLL ensures that a validation procedure is performed
  in which the sample is randomly split into ten folds. Through
  each of the ten runs, one of the ten subsamples is the holdout
  set and therefore not included in the model estimation part.
  The holdout set is consequently used to obtain the output
  statistics using the parameter estimates from the model
  estimation on the other 9 subsamples. Finally, the output
  includes the validation statistics which is the sum for these
  statistics over all ten folds.
outfile "class3psy.sav" classification
  keep STUDENT_VOLGNR;
  //this option specifies a file as output that has all posterior
  prediction probabilities for each cluster per row. In this way
  we obtain the classification of students to the cluster for
  which this posterior predictive probability is highest.
variables
  dependent JR2_GELDEND_RESULTAAT continuous;
  independent yearly_average numeric, yearly_stddev
  numeric, yearly_compensated numeric, yearly_herkansing numeric,
  JR1_GELDEND_RESULTAAT numeric;
  latent
```

```
Class nominal 3;
equations
Class <- 1 + yearly_average + yearly_stddev + yearly_compensated
+
    yearly_herkansing;
//This equation specifies the variables used to define the
latent classes.
JR2_GELDEND_RESULTAAT <- 1|Class + JR1_GELDEND_RESULTAAT|Class;
//This equation specifies the regression analysis, where the
influence of first-year grades on second-year grades is allowed
to vary across classes.
JR2_GELDEND_RESULTAAT | Class;
//This equation allows the error variances for the dependent
variable to vary across classes.
```


4

Correcting for Guessing in Estimating True Scores in Higher Education Tests

This chapter is under revision as:

Yocarini, I. E., Bouwmeester, S., & Jongerling, J. (submitted). Correcting for guessing in estimating true scores in higher education tests.

Abstract

In small-scale multiple choice (MC) tests, as used in higher education, a correction for guessing is often applied when calculating test scores. A classical method adjusts the number of correct items by subtracting a proportion of items examinees answered correctly assuming they would have purely guessed (i.e., formula scoring).

Problematically, the guessing probability may not be accurate as students may have partial knowledge. In this simulation study the performance of the classical and alternative correction methods were evaluated. Results from two studies showed that the estimation of true scores might be improved by using the extended classical correction method proposed by Calandra (1941) and Hamilton (1950) or by using a method, such as our proposed weighted item difficulty correction, that incorporates item characteristics in the true score estimation.

Keywords: Guessing correction, formula scoring, higher education, MC items, true score estimation.

Introduction

In higher education classrooms, multiple choice (MC) tests are often small-scaled, designed in-house for each course. The goal of these tests is to estimate students' true scores as accurately as possible. Whether true scores are estimated correctly depends on various factors; for example, it depends on the number of items in the test, the quality of the items, the number of different factors assessed in the test, the instruction of the teacher, the number of different strategies employed by test-takers, and whether and how the test scores are corrected for guessing. In this study we evaluate the accuracy of estimated true scores for different methods to correct for guessing. We hereby focus on MC test in which incorrect answers are not directly penalized, meaning that no points are deducted for filling out the wrong answer. In this context, test takers' optimal and most common strategy is to guess instead of omit answers. Here, guessing is defined as the probability of answering an item correct when a student has an infinite negative ability for this item.

Psychometrically, guessing poses a problem as it interferes with estimating true scores using observed item responses on MC items. The problem is that it remains a question whether correct answers are due to knowledge or lucky guesses (Bar-Hillel, Budescu, & Attali, 2005; Budescu & Bar-Hillel, 1993). Without correcting for guessing, using the so-called *number right scoring* rule, the number of correct answers are simply summed to obtain a total test score. Under this scoring rule, guesses that result in correct answers will always lead to overestimates of the true scores. In order to make this estimation more accurate, some kind of correction seems appropriate.

Whether MC test scores should be corrected for guessing or not has been a topic for debate since the introduction of correction methods (e.g., Angoff & Schrader, 1984; Lord, 1975). Recently, there has been a shift from using a classical correction for guessing method, such as *formula scoring*, to a number right scoring method (i.e., not correcting for guessing), as was done for the new Scholastic Aptitude Test (SAT) by the College Board (Guo, 2017). However, the context of large-scale standardized tests such as the SAT is different from the small-scale tests used in higher education and consequently different challenges exist in estimating students' true scores here. In large-scale standardized tests item properties are often known, making it possible to

obtain more accurate true score estimates. For higher education tests, the samples are mostly too small and tests are often designed for individual courses by different academics throughout cohorts, making it harder to obtain information on the functioning of items to use in estimating students' true scores.

The higher education context also differs from the context of large-scale tests such as the SAT in terms of the guessing process. Guo (2017) evaluated both number right and formula scoring in the context of large-scale standardized tests and denoted omitted answers are often observed under formula scoring. While this might be true for large-scale standardized tests, omitted answers are rarely observed in higher education course tests where a correction for guessing is applied. This difference in occurrence of omitted answers might be caused by different factors. The stakes are often high for tests such as the SAT as these are used in the college admittance process. While for individual courses in a higher education curriculum the stakes are generally low. The scope of a higher education course might be more confined than a large-scale standardized test, which might result in less random guesses and relatively more partial knowledge in higher education tests. When we compare the optimal strategy for test-takers, the optimal strategy under number right scoring is to guess instead of omitting answers, as omissions are not penalized. Under formula scoring however, the optimal strategy would be to omit answers only when they would be pure guesses. For all other guesses, the certainty of guessing would be higher than the penalty given for an omission (Budescu & Bar-Hillel, 1993). So, if students in higher education more often possess partial knowledge and have fewer pure guesses, their optimal strategy might be not to omit answers.

Consequently, rarely observing omissions in higher education MC test responses, the question remains how to correct for guessing. Here, it remains a problem that the underlying process of guessing is unobserved and may differ for individuals as they may differ in risk aversion or calibration accuracy (Budescu & Bar-Hillel, 2005; Espinosa & Gardeazabal, 2010). Facing these individual differences and latent processes, it is hard to decide which guessing method leads to the most accurate true score estimate. Especially, as different correction methods differ in the (implicit) assumptions that are made about the underlying guessing process. Overall, the goal of

this simulation study is to evaluate the accuracy of estimated true scores using different correction for guessing methods in non-standardized small-scaled higher education MC tests where the predominant strategy of students is to guess instead of omit answers. This study hereby provides a review of the different correction for guessing methods available and fills a gap in the literature by focusing on guessing methods in higher education classroom assessment as opposed to large-scale testing.

Study 1

Correction for Guessing Methods

Classical correction method. Classically, subtracting a score from the number correct score is used to correct for guessing. The first known publication to suggest a method to correct for guessing in classroom testing was written by McCall (1920). Opposed to using the number of correct responses as an estimate of the true knowledge score (here referred to as the *number right scoring* or *no correction* method), McCall (1920) proposed a correction method (widely known as an example of *formula scoring*; Lord, 1963) in which the number of correct responses, R , is adjusted for guessing using

$$(1) S = R - \frac{k-R}{a-1}.$$

Where S is the adjusted number correct score, k is the total number of items and a refers to the number of answer alternatives. Consequently, estimated true knowledge proportion, \widehat{tkp}_i , for a number correct R is obtained using:

$$(2) \widehat{tkp}_i = S/k.$$

The implicit assumption about the latent guessing process here is that examinees either know the answer to an item or else guess among the alternative response options at random (Lord, 1975). Consequently, it is assumed that all incorrect responses are guessed wrong and correct responses are obtained either by knowledge or guessing ($k - R$; Diamond & Evans, 1973) with a probability of $(a-1)/a$ to choose an incorrect option and $1/a$ to choose the correct option.

Psychometrically, there are at least two important problems in the estimation of the test-taker's true score. First, responses that were not guessed are nevertheless

corrected for it, lowering the total correct score. This results in an adjusted score that is an *underestimation* of the true knowledge score. Second, when an examinee guesses he or she will not consider all answer options most of the time because one or more options can be ruled out by some partial knowledge. This might result in an adjusted score that is an *overestimation* of the true knowledge score.

Extended classical correction method. Another issue with the classical correction method is that the adjusted number correct score S is not a real *estimate* of the true knowledge score as S has no error. That is, although the relation between R and S can be expressed as a linear model $S = b_0 + b_1R$, with $b_0 = \frac{-k}{a-1}$ and $b_1 = \frac{a}{a-1}$, b_0 and b_1 are not estimated parameter values that vary per sample but fixed values defined by constants k and a . Calandra (1941) and Hamilton (1950, eq. 15) transformed the classical correction into a true linear regression equation by incorporating sample properties. They used the mean number correct score, \bar{R} , and variance of the number correct score s_R^2 to derive the estimated adjusted number correct score \hat{S} :

$$(3) \hat{S} = \frac{\bar{R}(k-\bar{R})-ks_R^2}{(a-1)s_R^2} + \frac{(as_R^2-k-\bar{R})R}{(a-1)s_R^2} \text{ (Hamilton, 1950).}$$

By substituting \hat{S} (Equation 3) for S in Equation 2, estimated true knowledge proportions are derived.

Beta binomial correction method. Although an improvement, this *extended classical correction* method does not solve the two issues that lead to under- and overestimation of true scores due to correcting of answers that were not guessed and answers where partial knowledge was used. Moreover, it might be suboptimal as it only uses the sample mean and variance of the number correct scores in the sample to estimate \hat{S} . A scoring formula in which the entire distribution of the number correct score is taken into account might lead to a better estimate of the true knowledge score. This brings us to the extensively discussed (Lord, 1959; Morrison & Brockway, 1979) mixed binomial model as a method to obtain accurate knowledge estimates. The mixed binomial model is based on the binomial model in which the probability of a certain number correct score R in a test of k items is defined as:

$$(4) b(R, k, p) = \binom{k}{R} p^R (1-p)^{k-R}.$$

Where p is a fixed probability of a correct response to an item j , $j = 1, \dots, k$. Note that in this binomial distribution it is assumed that examinees have the same true knowledge score and all items have the same difficulty level (Keats & Lord, 1962). To allow for differences in individuals' ability level, the mixed binomial distribution was proposed (Lord, 1959). See Appendix A for an elaborate description of the mixed binomial distribution method and the calculation of the true knowledge proportions.

Extended beta binomial correction method with three parameters. Because the beta binomial model takes the distributional properties of the number correct score R into account, it may be an improvement over the *classical* and *extended classical* correction for guessing method. Nevertheless, some problems remain. First, in the beta binomial model it is implicitly assumed that all items are equally difficult. For many tests this is not only unrealistic to assume, it even would be undesirable to have a test with equally difficult items. Instead, a test developer wishes to vary the item difficulty level to optimally discriminate between students among the entire range of true knowledge levels. Secondly, the probability p covers the full $0 < p < 1$ range in the mixed binomial distribution. For multiple choice items, however, the lower bound probability, π_0 , might be larger as guessing gives a probability of at least $1/a$ for a correct answer (Morrison & Brockway, 1979). As Carlin and Rubin (1991) mention, this is evident in observed score distributions for MC tests which often exhibit a lack of low scores. Lord (1965) proposed a modified beta binomial distribution with range $0 \leq a < p < b \leq 1$ to solve this issue.

Extended beta binomial correction method two parameters. Morrison and Brockway (1979) derived at the same extended beta binomial method as Carlin and Rubin (1991), see Equation 17 in Appendix A. Except, Morrison and Brockway (1979) their model is somewhat simpler as π_0 is not estimated but fixed at $1/a$. The model is simpler, however, a being fixed might give problems similar to those mentioned for the classical correction method where the guessing probability is also fixed at $1/a$.

Three-parameter logistic model (3PLM) correction method. Although taking into account more properties of the distribution of knowledge scores than the correction methods discussed so far, the extended beta binomial model does not provide a solution to the unrealistic assumption that all items are equally difficult. Item

response theory (IRT) models take individual item parameters into account and may therefore be more adequate in estimating the true knowledge proportion. Whereas IRT models are generally applied to large-scale standardized tests, sample sizes in higher education tests are often too small to apply IRT. However, we are still interested to see how the performance of an IRT correction for guessing compares to other methods. As explained, each method makes assumptions about the underlying guessing process. Given that this process is unobserved it is unsure which method results in the most accurate true scores on average, even if the context in which the method is applied is suboptimal. In the three-parameter logistic model (3PLM; Birnbaum, 1968) a separate parameter is estimated not only for each item's difficulty, δ_j but also for the item's discriminability, α_j , and the lower bound probability, π_{0j} . Moreover, individuals may vary with respect to their true knowledge level, indicated by a latent trait level θ_i which, unlike the binomial mixture models, may vary between individuals with the same number correct score R . The 3PLM model is defined as:

$$(5) P(X_j = 1 | \hat{\theta}_i, \hat{\delta}_j, \hat{\alpha}_j, \hat{\pi}_{0j}) = \hat{\pi}_{0j} + (1 - \hat{\pi}_{0j}) \frac{\exp(\hat{\alpha}_j(\hat{\theta}_i - \hat{\delta}_j))}{1 + \exp(\hat{\alpha}_j(\hat{\theta}_i - \hat{\delta}_j))}.$$

See Equation 18 in Appendix A for an overview of how $\hat{\theta}_i$ (latent trait scores) are consequently calculated using these probabilities.

To summarize the developments in the correction for guessing methods, different types of methods can be distinguished. Overall, the methods differ with respect to the extent in which they take the sample characteristics into account. Neither the *no correction*, nor the *classical correction* method takes any sample characteristic into account in estimating true knowledge scores from item responses. The *extended classical correction* method takes into account the mean and variance of the sample. The methods using the *beta binomial distribution* take the distribution of total scores R into account by assuming all items have a similar difficulty and the *extended beta binomial* methods also take into account differences in true knowledge. Finally, the *3PLM* integrates variability in students' ability and in items.

Method

A simulation study was performed using R (R core Team, 2016) to evaluate the different correction methods. The aim of our study is to evaluate the different correction methods in estimating accurate true scores in a range of realistic higher education test settings. In practice however, we only have the observed responses as the true scores are unknown by definition. Therefore, in order to evaluate the correction methods we performed a simulation study in which true scores were simulated throughout a range of realistic higher education settings. In order to define this range we used estimates from observational data from education settings, such as, for example, sample size, the number of items, and item information (see the next paragraph for more detailed information on the parameters that were varied). Additionally, many factors might influence the accuracy of the correction methods, by performing a simulation study we were able to evaluate the influence of the factors in isolation.

First, for each student i , $i = 1, \dots, n$, a true knowledge score τ_i , was simulated from a truncated normal distribution $N \sim (\mu_\tau, \sigma_\tau)$. Unlike a uniform distribution in which each true knowledge score would have an equal probability of occurring, a normal distribution was assumed as most students have an average true knowledge score and probabilities to score higher or lower decline. Here, μ_τ refers to the average true knowledge score of the sample and σ_τ indicates the variation in the true knowledge score across students. The true knowledge scores were sampled from a truncated distribution with $\tau_{min} = -3$ and $\tau_{max} = 3$ as these were subsequently transformed to a *true knowledge proportion (tkp)* which is bounded between 0 and 1. Transformation was done using the linear equation:

$$(6) \text{tkp}_i = b_0 + b_1 \tau_i.$$

Here, b_0 was fixed at .5, assuming that an average student has a general knowledge proportion of .5 when $\bar{\tau} = 0$ and $b_1 = \frac{0.5}{3} = 0.167$ to obtain a *tkp* between 0 and 1. Next, item responses were simulated for each student, based on their simulated true knowledge score. To do so, the 3PLM (Equation 5) was applied to calculate the probability to answer an item correct for each item for each student. Using the ltm

package in R (Rizopoulos, 2006) these were subsequently converted into correct ($X_j = 1$) and incorrect ($X_j = 0$) item scores using the probability to answer an item correct of .5 as the cut-score.

Simulated conditions. To evaluate the bias of the correction methods within different realistic higher educational contexts, true knowledge scores were simulated under different conditions. Datasets were simulated varying different variables: the average true knowledge score, the variance in the true knowledge score, the sample size, the number of items, the number of item response options, the average difficulty of the items, and the average discriminability of the items. As shown in Table 1, the manipulated variables can be divided into those related to the sample and those related to the test.

Table 1: Overview Simulated Variables and Values

	Variable	Notation	Value(s)
Sample Properties	Average true knowledge score	μ_τ	-1, 0, 1
	Variance true knowledge score	σ_τ	1, 1.8
	Size sample	n	20, 200, 400
Test Properties	Number of items	k	20, 40, 50, 60
	Number of item response options	a	2, 3, 4
	Average item difficulty	μ_δ	-.8, 0, .8
	Variability in item difficulty	σ_δ	.5
	Average item discriminability	μ_α	.3, .5
	Variability in item discriminability	σ_α	.3
	Average guessing probability	μ_γ	.25
Variability in guessing probability	σ_γ	.05	

Table 2: Settings for Simulated Datasets

Setting	μ_τ	σ_τ	n	k	a	μ_δ	μ_α
1	0	1	400	40	4	0	.5
2	-1	1	400	40	4	0	.5
3	1	1	400	40	4	0	.5
4	0	1.8	400	40	4	0	.5
5	0	1	200	40	4	0	.5
6	0	1	20	40	4	0	.5
7	0	1	400	20	2	0	.5
8	0	1	400	20	3	0	.5
9	0	1	400	20	4	0	.5
10	0	1	400	50	3	0	.5
11	0	1	400	60	2	0	.5
12	0	1	400	60	3	0	.5
13	0	1	400	60	4	0	.5
14	0	1	400	40	4	.8	.5
15	0	1	400	40	4	-.8	.5
16	0	1	400	40	4	0	.3

Overall 16 settings (see Table 2) were chosen in which datasets were simulated. These specific values and combinations of values were chosen for their relevance to practical settings. In European higher education popular study programs have about 400 students while more specialized study programs might only have 20 students in a cohort. Furthermore, most MC tests consist of 40 items with each having four item response categories. If more items are included in a test, they generally have fewer item response categories. The item difficulty and item discriminability values were based on data of first year psychology students. As most tests consist of four item response categories the average guessing probability was set to .25 with a standard deviation of .05 throughout items. In this way, the guessing probability could be lower for items where, for example, students choose the distractor answer options and the guessing probability could be higher for items where students, for example, only guessed among three answer options due to partial knowledge. For each scenario 1000 datasets were simulated. The R code is available upon request.

Outcome measure. Finally, to evaluate the accuracy of each correction method, two measures were used to compare the estimated true knowledge proportion, \widehat{tkp}_i , to the true knowledge proportions, tkp_i . The mean sum of squared error (MSE) was calculated using:

$$(7) \text{MSE} = \sqrt{\frac{\sum_{i=1}^n (\widehat{tkp}_i - tkp_i)^2}{n}}.$$

The MSE is informative as it shows the absolute differences between the estimated and true knowledge scores. However, the MSE does not give information about the ranking of the estimated true scores. Therefore, the correlation between the estimated and true knowledge scores was calculated as a second measure of bias. Notably, the correlation is only interesting to compare the no correction and the classical correction methods on one hand and the remaining methods on the other hand, as each of these former methods are a linear transformation of one another.

Results

To keep a clear overview, the results of the comparison of different methods are discussed separately for each manipulated variable. To handle the large number of results only main effects are discussed.

True knowledge scores. Table 3 shows the results for varying levels of the average true knowledge scores in a sample for seven different correction methods. In an average cohort bias was lowest for the extended classical correction method, followed by the classical correction method. In a cohort where the average true knowledge score was low, the extended classical correction had the lowest bias. In a cohort where the average true knowledge score was high, no correction clearly resulted in the lowest bias. The correlation values in Table 3 show that overall, independent of the cohort's average true knowledge, no correction and the classical correction methods resulted in the highest correlation.

Table 3: Effect of Changes in True Knowledge Scores

Correction Method	Bias						Correlation					
	Average		Low Mean		High Mean		Average		Low Mean		High Mean	
	Cohort	True Knowledge	Cohort	True Knowledge	Cohort	True Knowledge	Cohort	True Knowledge	Cohort	True Knowledge	Cohort	True Knowledge
1 None	.1677 (.0029)	.2346 (.0057)	.1177 (.0015)	.7362 (.0385)	.6924 (.0401)	.7237 (.0406)	.1677 (.0029)	.2346 (.0057)	.1177 (.0015)	.7362 (.0385)	.6924 (.0401)	.7237 (.0406)
2 Classical	.1147 (.0015)	.1309 (.0026)	.1244 (.0025)	.7363 (.0385)	.6925 (.0402)	.7240 (.0406)	.1147 (.0015)	.1309 (.0026)	.1244 (.0025)	.7363 (.0385)	.6925 (.0402)	.7240 (.0406)
3 Extended classical	.1118 (.0015)	.1227 (.0024)	.1261 (.0027)	.7362 (.0385)	.6924 (.0401)	.7237 (.0406)	.1118 (.0015)	.1227 (.0024)	.1261 (.0027)	.7362 (.0385)	.6924 (.0401)	.7237 (.0406)
4 Beta Binomial	.2078 (.0040)	.2847 (.0057)	.2422 (.0043)	.7272 (.0390)	.6784 (.0407)	.7150 (.0405)	.2078 (.0040)	.2847 (.0057)	.2422 (.0043)	.7272 (.0390)	.6784 (.0407)	.7150 (.0405)
5 Extended BB ¹ 3 par. ²	.3369 (.0958)	.2972 (.0309)	.4213 (.1541)	.6792 (.1208)	.6322 (.1221)	.6689 (.1023)	.3369 (.0958)	.2972 (.0309)	.4213 (.1541)	.6792 (.1208)	.6322 (.1221)	.6689 (.1023)
6 Extended B2 par.	.2898 (.0072)	.2185 (.0035)	.3930 (.0116)	.7003 (.0400)	.6689 (.0412)	.6740 (.0426)	.2898 (.0072)	.2185 (.0035)	.3930 (.0116)	.7003 (.0400)	.6689 (.0412)	.6740 (.0426)
7 3PL ³ model	.1550 (.0220)	.2147 (.0252)	.1564 (.0390)	.6186 (.4018)	.5355 (.3963)	.6347 (.3669)	.1550 (.0220)	.2147 (.0252)	.1564 (.0390)	.6186 (.4018)	.5355 (.3963)	.6347 (.3669)
1 Rasch model	.1516 (.0093)	.2286 (.0119)	.1098 (.0063)	.7395 (.0376)	.6906 (.0417)	.7223 (.0388)	.1516 (.0093)	.2286 (.0119)	.1098 (.0063)	.7395 (.0376)	.6906 (.0417)	.7223 (.0388)
2 Weighted item difficulty	.1616 (.0029)	.2241 (.0057)	.1181 (.0015)	.7314 (.0388)	.7030 (.0372)	.6859 (.0489)	.1616 (.0029)	.2241 (.0057)	.1181 (.0015)	.7314 (.0388)	.7030 (.0372)	.6859 (.0489)

Note: Standard deviations over simulations given between brackets. ¹BB = Beta Binomial. ² par. = parameters. ³PL = three parameter logistic.

In addition, the variability of the true knowledge score distribution in a cohort was varied. As shown in Table 4, the mean bias results show that in both an average and variable cohort the extended classical correction method performed best. The same holds for the correlation results, independent of the variability in true knowledge scores, no correction and the classical methods were best. Notably, the correlations were less stable in a variable cohort than in a less variable cohort.

Table 4: Effect of Changes in Variability in True Knowledge Scores

Correction Method	Bias				Correlation			
	Average Cohort		Variable Cohort		Average Cohort		Variable Cohort	
1 None	.1677	(.0029)	.1846	(.0330)	.7362	(.0385)	.5579	(.0683)
2 Classical	.1147	(.0015)	.1396	(.0024)	.7363	(.0385)	.5579	(.0683)
3 Extended classical	.1118	(.0015)	.1366	(.0021)	.7362	(.0385)	.5579	(.0683)
4 Beta Binomial	.2078	(.0040)	.1645	(.0038)	.7272	(.0390)	.8370	(.0269)
5 Extended BB ¹ 3 par. ²	.3369	(.0958)	.3199	(.1102)	.6792	(.1208)	.7924	(.1139)
6 Extended BB 2 par.	.2898	(.0072)	.2457	(.0076)	.7003	(.0400)	.8132	(.0289)
7 3PL ³ model	.1550	(.0220)	.1595	(.0297)	.6186	(.4018)	.7687	(.3639)
1 Rasch model	.1516	(.0093)	.1767	(.0099)	.7395	(.0376)	.8365	(.0259)
2 Weighted item difficulty	.1616	(.0029)	.1870	(.0040)	.7314	(.0388)	.8323	(.0276)

Note: Standard deviations over simulations given between brackets. ¹BB = Beta Binomial. ² par. = parameters. ³3PL = three parameter logistic.

Item Difficulty. As shown in Table 5, varying the item difficulty results influenced the most optimal correction method. For a test with more difficult items the mean bias was lowest when no correction for guessing was applied. Alternatively, for a test with less difficult items the extended classical correction method performed best. The correlation results show that no correction and the classical correction methods resulted in the highest correlation.

Table 5: Effect of Changes in Item Difficulty

Correction Method	Bias					
	Average Cohort	Higher Item Difficulty (.8)	Lower Item Difficulty (-.8)	Average Cohort	Higher Item Difficulty (.8)	Lower Item Difficulty (-.8)
1 None	.1677 (.0029)	.1306 (.0024)	.2215 (.0039)	.7362 (.0385)	.7109 (.0388)	.7338 (.0386)
2 Classical	.1147 (.0015)	.1479 (.0025)	.1436 (.0023)	.7363 (.0385)	.7109 (.0389)	.7338 (.0386)
3 Extended classical	.1118 (.0015)	.1502 (.0027)	.1373 (.0022)	.7362 (.0385)	.7109 (.0388)	.7338 (.0386)
4 Beta Binomial	.2078 (.0040)	.2124 (.0041)	.2091 (.0040)	.7272 (.0390)	.6996 (.0394)	.7252 (.0390)
5 Extended BB ¹ 3 par. ²	.3369 (.0958)	.3415 (.0937)	.3223 (.0888)	.6792 (.1208)	.6415 (.1346)	.6832 (.1008)
6 Extended BB 2 par.	.2898 (.0072)	.3166 (.0068)	.2714 (.0070)	.7003 (.0400)	.6754 (.0401)	.6969 (.0417)
7 3PL ³ model	.1550 (.0220)	.1523 (.0208)	.1835 (.0216)	.6186 (.4018)	.5784 (.3826)	.6217 (.4094)
1 Rasch model	.1516 (.0093)	.1267 (.0085)	.1927 (.0092)	.7395 (.0376)	.7089 (.0377)	.7362 (.0387)
2 Weighted item difficulty	.1616 (.0029)	.1250 (.0021)	.2126 (.0037)	.7314 (.0388)	.7189 (.0359)	.7039 (.0454)

Note: Standard deviations over simulations given between brackets. ¹BB = Beta Binomial. ² par. = parameters. ³PL = three parameter logistic.

Item Discrimination. The results of varying the item discrimination levels are shown in Table 6. The mean bias results show that the extended classical correction method worked best, independent of the item discrimination level. Similarly, no correction and the classical correction methods resulted in the highest correlation. Overall, the correlation was considerably lower when item discrimination was lower, indicating the need for well discriminating items when estimating true scores. Also, the correlations in a cohort with lower item discrimination levels were less stable than that of an average cohort.

Table 6: Effect of Changes in Item Discrimination

Correction Method	Bias				Correlation			
	Average Cohort		Lower Item Discrimination .3		Average Cohort		Lower Item Discrimination .3	
1 None	.1677	(.0029)	.1846	(.0330)	.7362	(.0385)	.5579	(.0683)
2 Classical	.1147	(.0015)	.1396	(.0024)	.7363	(.0385)	.5579	(.0683)
3 Extended classical	.1118	(.0015)	.1366	(.0021)	.7362	(.0385)	.5579	(.0683)
4 Binomial	.2078	(.0040)	.2466	(.0151)	.7272	(.0390)	.5507	(.0670)
5 Extended BB ¹ 3 par. ²	.3369	(.0958)	.3597	(.0910)	.6792	(.1208)	.5038	(.1185)
6 Extended BB 2 par.	.2898	(.0072)	.3317	(.0141)	.7003	(.0400)	.5211	(.0675)
7 3PL ³ model	.1550	(.0220)	.2032	(.0280)	.6186	(.4018)	.3128	(.4910)
1 Rasch model	.1516	(.0093)	.1792	(.0104)	.7395	(.0376)	.5575	(.0688)
2 Weighted item difficulty	.1616	(.0029)	.1806	(.0033)	.7314	(.0388)	.5542	(.0688)

Note: Standard deviations over simulations given between brackets. ¹BB = Beta Binomial. ² par. = parameters. ³3PL = three parameter logistic.

Number of items. Table 7 shows the results for varying the number of items and number of response categories for the different correction methods. As the results show these variables influenced the optimal method. Where the extended classical correction method was best for tests with 40 items that had four response categories, the classical correction was best for tests of 50 items with 3 response categories. When a test included 60 items the 3PLM model resulted in the lowest bias. The 3PLM, however, produced relatively unstable results compared to no correction, which was the second-best method. For tests with 20 items no correction for guessing was best, followed by the 3PLM model. The correlation results were more consistent. Overall, no correction and the classical correction methods resulted in the highest correlation.

Notably, the correlation became smaller as a test only included 20 items having three of four response categories.

Number of students. The number of students in a cohort does not seem to influence which method is optimal as shown in Table 8. Overall, the extended classic correction resulted in the lowest bias. Similarly, no correction, the classical correction methods, and the lowest percentile correction resulted in the highest correlation.

Table 7: Effect of Changes in Number of Items

Method	Bias																																																																							
	40 items				50 items				60 items				20 items																																																											
	4 options ¹	3 options	2 options	3 options	4 options	3 options	2 options	3 options	4 options	3 options	2 options	3 options	4 options	3 options	2 options	3 options	4 options																																																							
1 ^a	.1677 (.0029)	.1647 (.0027)	.1617 (.0025)	.1624 (.0079)	.1620 (.0080)	.1833 (.0040)	.1836 (.0110)	.1841 (.0180)	.1147 (.0015)	.1259 (.0020)	.2646 (.0064)	.2641 (.0126)	.2648 (.0127)	.2896 (.0095)	.2897 (.0165)	.2889 (.0162)	.1118 (.0015)	.1275 (.0023)	.2850 (.0079)	.2840 (.0146)	.2849 (.0145)	.3296 (.0157)	.3289 (.0235)	.3281 (.0230)	.2078 (.0040)	.1994 (.0035)	.1918 (.0031)	.1923 (.0078)	.1923 (.0078)	.2375 (.0058)	.2379 (.0120)	.2375 (.0121)	.3369 (.0958)	.3235 (.0940)	.3416 (.1006)	.3379 (.1411)	.3307 (.1388)	.3360 (.0873)	.3372 (.1142)	.3374 (.1140)	.2898 (.0072)	.3489 (.0092)	.4525 (.0111)	.4530 (.0119)	.4531 (.0118)	.3888 (.0103)	.3897 (.0132)	.3886 (.0130)	.1550 (.0220)	.1416 (.0203)	.1335 (.0196)	.1316 (.0466)	.1335 (.0491)	.2016 (.0246)	.2031 (.0491)	.1990 (.0452)	.1516 (.0093)	.1458 (.0084)	.1416 (.0079)	.1417 (.0077)	.1420 (.0077)	.1747 (.0113)	.1755 (.0116)	.1744 (.0116)	.1616 (.0029)	.1584 (.0027)	.1550 (.0024)	.1558 (.0082)	.1554 (.0082)	.1786 (.0040)	.1790 (.0114)	.1794 (.0112)

¹Average Cohort. Note: Standard deviations over simulations given between brackets. ^aNo Correction, ^bClassical Correction, ^cExtended classical correction, ^dBeta Binomial correction, ^eExtended Beta Binomial correction three parameters, ^fExtended Beta Binomial correction two parameters, ^gThree parameter logistic model, ^hRasch model, ⁱWeighted item difficulty correction.

Table 7: Effect of Changes in Number of Items

Method	Correlation															
	50 items				60 items				20 items							
	40 items	40 items ¹	3 options	3 options	2 options	2 options	3 options	3 options	4 options	4 options	2 options	2 options	3 options	3 options	4 options	4 options
1 ^a	.7362 (.0385)	.7716 (.0331)	.7362 (.0385)	.7994 (.0275)	.8007 (.0271)	.7716 (.0331)	.6098 (.0581)	.6115 (.0572)	.7362 (.0385)	.7994 (.0275)	.8007 (.0271)	.7716 (.0331)	.6098 (.0581)	.6115 (.0572)	.7362 (.0385)	.7994 (.0275)
2 ^b	.7363 (.0385)	.7718 (.0332)	.7363 (.0385)	.7916 (.0260)	.7925 (.0260)	.7718 (.0332)	.5958 (.0571)	.5984 (.0563)	.7363 (.0385)	.7916 (.0260)	.7925 (.0260)	.7718 (.0332)	.5958 (.0571)	.5984 (.0563)	.7363 (.0385)	.7916 (.0260)
3 ^c	.7362 (.0385)	.7716 (.0331)	.7362 (.0385)	.7994 (.0275)	.8007 (.0271)	.7716 (.0331)	.6098 (.0581)	.6115 (.0572)	.7362 (.0385)	.7994 (.0275)	.8007 (.0271)	.7716 (.0331)	.6098 (.0581)	.6115 (.0572)	.7362 (.0385)	.7994 (.0275)
4 ^d	.7272 (.0390)	.7616 (.0332)	.7272 (.0390)	.7893 (.0274)	.7900 (.0273)	.7616 (.0332)	.6018 (.0583)	.6039 (.0576)	.7272 (.0390)	.7893 (.0274)	.7900 (.0273)	.7616 (.0332)	.6018 (.0583)	.6039 (.0576)	.7272 (.0390)	.7893 (.0274)
5 ^e	.6792 (.1208)	.7100 (.1234)	.6792 (.1208)	.7194 (.1455)	.7262 (.1423)	.7100 (.1234)	.5555 (.1311)	.5511 (.1350)	.6792 (.1208)	.7194 (.1455)	.7262 (.1423)	.7100 (.1234)	.5555 (.1311)	.5511 (.1350)	.6792 (.1208)	.7194 (.1455)
6 ^f	.7003 (.0400)	.6949 (.0375)	.7003 (.0400)	.5848 (.0406)	.5856 (.0405)	.6949 (.0375)	.5269 (.0591)	.5301 (.0580)	.7003 (.0400)	.5848 (.0406)	.5856 (.0405)	.6949 (.0375)	.5269 (.0591)	.5301 (.0580)	.7003 (.0400)	.5848 (.0406)
7 ^g	.6186 (.4018)	.6641 (.4019)	.6186 (.4018)	.6989 (.3996)	.6881 (.4019)	.6641 (.4019)	.4708 (.3527)	.4937 (.3263)	.6186 (.4018)	.6989 (.3996)	.6881 (.4019)	.6641 (.4019)	.4708 (.3527)	.4937 (.3263)	.6186 (.4018)	.6989 (.3996)
1 ^h	.7395 (.0376)	.7753 (.0321)	.8028 (.0274)	.8015 (.0259)	.8017 (.0268)	.6118 (.0558)	.6080 (.0595)	.6115 (.0575)	.7395 (.0376)	.7753 (.0321)	.8028 (.0274)	.8015 (.0259)	.8017 (.0268)	.6118 (.0558)	.6080 (.0595)	.6115 (.0575)
2 ⁱ	.7314 (.0388)	.7668 (.0334)	.7314 (.0388)	.7314 (.0280)	.7965 (.0274)	.7668 (.0334)	.6040 (.0588)	.6061 (.0572)	.7314 (.0388)	.7314 (.0280)	.7965 (.0274)	.7668 (.0334)	.6040 (.0588)	.6061 (.0572)	.7314 (.0388)	.7314 (.0280)

¹Average Cohort. Note: Standard deviations over simulations given between brackets. ^aNo Correction, ^bClassical Correction, ^cExtended classical correction, ^dBeta Binomial correction, ^eExtended Beta Binomial correction three parameters, ^fExtended Beta Binomial correction two parameters, ^gThree parameter logistic model, ^hRasch model, ⁱWeighted item difficulty correction.

Table 8: Effect of Changes in Number of Students

Correction Method	Bias			Correlation		
	400 Students ¹	200 Students	20 Students	400 students ¹	200 Students	20 Students
1 None	.1677 (.0029)	.1673 (.0034)	.1679 (.0227)	.7362 (.0385)	.7370 (.0420)	.7282 (.1144)
2 Classical	.1147 (.0015)	.1147 (.0016)	.1146 (.0189)	.7363 (.0385)	.7370 (.0420)	.7282 (.1144)
3 Extended classical	.1118 (.0015)	.1119 (.0017)	.1123 (.0189)	.7362 (.0385)	.7370 (.0420)	.7282 (.1144)
4 Beta Binomial	.2078 (.0040)	.2079 (.0044)	.2240 (.0560)	.7272 (.0390)	.7280 (.0423)	.7253 (.1097)
5 Extended BB 3 par.	.3369 (.0958)	.3402 (.0976)	.3891 (.1483)	.6792 (.1208)	.6772 (.1215)	.6668 (.1623)
6 Extended BB 2 par.	.2898 (.0072)	.2900 (.0082)	.3014 (.0498)	.7003 (.0400)	.7002 (.0456)	.6978 (.1198)
7 3PL model	.1550 (.0220)	.1761 (.0258)	.2403 () ²	.6186 (.4018)	.5334 (.4505)	.1682 (.4859)
1 Rasch model	.1516 (.0093)	.1523 (.0109)	.1542 (.0249)	.7395 (.0376)	.7360 (.0446)	.2379 (.4911)
2 Weighted item difficulty	.1616 (.0029)	.1604 (.0033)	.1441 (.0220)	.7314 (.0388)	.7311 (.4025)	.7135 (.1202)

¹Average Cohort. ² Too little data for 3PLM to provide parameter estimates. Note: Standard deviations over simulations given between brackets. ¹BB = Beta Binomial. ² par. = parameters. ³3PL = three parameter logistic.

Discussion

In Study 1 we compared the accuracy of estimated students' true scores for different correction methods: not correcting for guessing, the (extended) classical correction, the beta binomial correction, and the three-parameter logistic correction method. Overall, the extended classical method resulted in the lowest bias and the highest correlation. However, the results showed that the best method to correct for guessing did depend on the average true knowledge proportion in a cohort, the item difficulty, and the number of items and its number of response categories. In a cohort in which the true knowledge score was high and for tests with difficult items, no correction for guessing resulted in the least bias. This might indicate that no correction for guessing is better at estimating students' true knowledge scores when a student's ability and the difficulty of a test do not align. For example, students who have a low true knowledge score will only answer easy items on a difficult test correct. If you consequently would correct for guessing this might introduce bias as no guessing occurred (i.e., students knew the answer to the easy items). Alternatively, not correcting for guessing might be better in these situations. Furthermore, for tests with 60 items, the three-parameter logistic model performed best on average. However, high standard deviations in the bias and correlation estimates showed these were unstable. Alternatively, no correction for guessing was second best for these tests and performed more stable. For tests with 20 items, no correction for guessing outperformed the other methods. Despite these minor exceptions, the overall results indicate that in most higher education test settings a correction for guessing is necessary to estimate students' true scores accurately.

Furthermore, as the results show, the beta binomial models performed worse than the other methods throughout all settings. It seems an advantage that the beta binomial models use the complete distribution of total scores instead of only using some properties of the distribution when estimating the true knowledge scores. However, by approaching the complete distribution, some stringent assumptions are made: items are assumed to be parallel with respect to difficulty, discrimination, and lower bound probability. The results of our simulation show that these assumptions may not be realistic and as a consequence the complete distribution does not reflect the true

distribution well. Instead, the 3PLM does a much better job of estimating accurate true knowledge scores by estimating the item parameters as well as an individual latent theta score for each student. This model, however, has the problem that the number of parameters is very large, requiring a large sample size to obtain stable true score estimates. In contrast to large-scale standardized tests in which item parameters are known, sufficiently large sample sizes to estimate item parameters accurately often lack in higher educational contexts. Consequently, the 3PLM is not an optimal method to estimate true score knowledge proportions.

Still, the results of Study 1 suggest that using item information might result in more accurate true knowledge score estimation in these contexts. Consequently, in Study 2 we attempt to improve the estimation of true scores by evaluating the fit of the Rasch model and a self-defined weighted item difficulty model, which both include item information but are less complex than the 3PLM and as such may require smaller sample sizes for obtaining stable results.

Study 2

Correction Methods Using Item Difficulty Information

Rasch model. The one-parameter logistic or Rasch model (Rasch, 1960) is a simpler IRT model than the 3PLM as only one parameter is estimated for the item's difficulty, $\hat{\delta}_j$. As in the 3PLM, students' ability is taken into account as individuals may vary with respect to their true knowledge score, indicated by θ_i . The Rasch model is defined by:

$$(8) P(X_j = 1 | \hat{\theta}_i, \hat{\delta}_j) = \frac{\exp(\hat{\theta}_i - \hat{\delta}_j)}{1 + \exp(\hat{\theta}_i - \hat{\delta}_j)}.$$

Subsequently, true knowledge scores, $\hat{\theta}_i$ (latent trait scores) can be calculated using the vector of responses on items, $j = 1, \dots, k$, of participant i and the estimated item parameters, using Equation 18 in Appendix A.

Weighted item difficulty method. In addition, a *weighted item difficulty correction* method was developed to incorporate item difficulty information. Here, credits students receive for correct item responses are weighted by the difficulty of the item. The rationale behind this method is that students who answer difficult items correctly

are assumed to have a higher true knowledge proportion than students who answer easy items correctly. Item difficulty is defined as the proportion correct and the true knowledge proportion of an item as the 1- the proportion correct. So, the easier an item and the higher the proportion correct for that item, the less weight it has in the estimated true knowledge proportion. The estimated true knowledge proportion is the sum of all item true knowledge proportions divided by the maximum score that can be obtained when all items are answered correctly:

$$(9) \widehat{tkp}_i = \frac{\sum_{j=1}^k X_{ij}(1-\bar{X}_j)}{\sum_{j=1}^k (1-\bar{X}_j)}.$$

By incorporating item difficulty information using sample information without the need to estimate population estimates or students' theta values, this method is less complex than the Rasch model. Note that our heading might be somewhat misleading as the Rasch and weighted item difficulty methods are more general methods to estimate true knowledge proportions and do not necessarily include a guessing parameter. Still, all included methods share the common goal to estimate true knowledge proportions. While most methods do this by including a guessing parameter, the Rasch and weighted item difficulty method do so differently. To evaluate how these two methods perform under different realistic conditions, a second simulation study was performed.

Method

Study 2 employed the same methods as Study 1 with the exception of evaluating the Rasch correction method and a weighted item difficulty correction method.

Results

For comparison, the results of the methods added in Study 2 are compared to the classical correction, the extended classical correction, and the 3PLM correction method from Study 1.

True knowledge scores. Table 3 shows the results for varying levels of the average true knowledge scores in a sample. As the results show, the extended classical correction resulted in the lowest bias in a sample with an average or low average true

knowledge score. In a sample where the average true knowledge score was high, the Rasch model resulted in the lowest bias followed by the weighted item difficulty correction method. Notably, the standard deviation for the latter was smaller than the variability in the Rasch model bias estimates. The correlation values show that for an average cohort the Rasch model the correlation was highest. In a cohort with low true knowledge the weighted item difficulty correction method performed best while in a cohort with a high average true knowledge score the classical correction methods performed best. Here, the standard deviation of the Rasch model correction method was smallest.

The results of varying the variability of the true knowledge score distribution in a cohort are shown in Table 4. The results did not depend on the variability in true knowledge scores in a cohort. While the bias values were lowest for the extended classical correction, the correlation was highest for the Rasch model correction. As the results show, the bias and correlation values varied more strongly across the correction methods in a variable cohort.

Item Discrimination. The results of varying the item discrimination levels are shown in Table 5. The mean bias results show that the extended classical correction method worked best, independent of the item discrimination level. Furthermore, the classical correction methods resulted in the highest correlation when item discrimination was low. In an average cohort the Rasch model resulted in the highest correlation. Here, the classical correction methods followed closely.

Item Difficulty. As shown in Table 6, the weighted item difficulty correction method had the lowest bias and highest correlation when item difficulty was high. When item difficulty was low, the extended classical correction method resulted in the lowest bias and the Rasch model in the highest correlation, closely followed by the classical correction methods.

Number of items. Table 7 shows the results for varying the number of items and number of response categories for the different correction methods. For a 20 item tests with 2 response categories, the Rasch model had the lowest bias. Here, the weighted item difficulty correction performed second best. Notably, the standard

deviation of the bias estimates was smallest for the weighted item difficulty method. For tests with more items the 3PLM method had the lowest bias. However, the variability in the estimated bias values was smaller for the Rasch and weighted item difficulty method. For 40 and 50 item tests the classical correction methods had the lowest bias. The correlation results show that the Rasch model resulted in the largest correlation for almost all tests. Only when a test had 20 items with 2 or 3 response categories, the classical correction methods performed best.

Number of students. The results for varying the number of students in a cohort are displayed in Table 8. The extended classical correction method resulted in the lowest bias. Also, the classical correction methods had the largest correlation for cohorts with 200 or 20 students. For the average cohort of 400 students the Rasch model correction had the highest correlations.

Discussion

In Study 2, we extended the comparison of different correction for guessing methods in higher education by including two additional correction methods. These alternative methods used item difficulty information without becoming as complex as the 3PLM method. Comparing the alternative methods to the 3PLM results showed that overall the standard deviation in the bias and the standard deviation in the correlation estimates for the Rasch and weighted item difficulty method were lower. In addition to this increased stability, the Rasch and weighted item difficulty method performed better than the 3PLM model in most simulated settings. Exceptions existed when the true knowledge in a cohort was low, the variability in true scores was high, items had low difficulty, and when tests had 50 or 60 items. In these settings, the average bias estimate was lower for the 3PLM method. For the correlation results, the alternative methods outperformed the 3PLM method throughout all settings. This can be explained by the complexity of the models. When many items are included, the complex 3PLM has more information to estimate θ scores adequately. However, when less information is provided the simpler Rasch model results in more accurate θ estimation. Note, however, that the 3PLM bias results are quite unstable and differences between the 3PLM and Rasch are consequently very small and negligible

in some cases. Overall, the results when varying the cohort size show that the 3PLM and Rasch model are not useful in practical settings with small sample sizes.

Compared to the classical correction methods that generally performed best in Study 1, the results of Study 2 indicate that methods using item information in some settings result in more accurate estimated knowledge scores. For the bias of the correction methods, this holds for cohorts in which the average true knowledge was high. Here, both the weighted item difficulty seems to produce the most stable lowest bias estimates. Also, when test have items with high difficulty, using the weighted item difficulty correction resulted in less biased estimated true knowledge scores.

Furthermore, for tests with 20 or 60 items methods using item information had lower bias than the classical correction methods. The correlation results show that the values were quite close to those of the classical correction methods overall. In cohorts with high variability in the true knowledge scores, the item difficulty methods clearly outperformed the classical correction methods.

Overall, from the correction methods that incorporate item difficulty, the weighted item difficulty correction method showed to be best and most stable.

General Discussion

The aim of this study was to compare the accuracy of different correction for guessing methods within realistic higher educational settings (i.e., small-scale non-standardized tests) in which students' predominant strategy is to guess instead of omit answers. Datasets were simulated in different settings, varying the average true knowledge score, the variance in the true knowledge score, the sample size, the number of items and the number of item response options, the mean difficulty of the items, and the mean discriminability of the items. Consequently, the performance of each method was evaluated using the mean sum of squared error (MSE) as an indication of the bias and by using the correlation between the estimated true scores and the simulated true knowledge scores.

Overall, taking the results from Study 1 and 2 together, the results indicate that a correction for guessing results in lower bias than no correction in almost all situations. The only exception existed for a cohort in which the mean true knowledge score was

high. Here, the bias was lowest when no correction was used which is expected given that students with high true knowledge scores will guess less often. In general, the results show that a correction for guessing is preferable in small-scale non-standardized MC tests. The results showed that on average, taking all sample characteristics into account, the optimal method in terms of lowest bias and highest correlation was the extended classical method. Independent of the specific context in terms of the item discrimination and the number of students in a cohort, the extended classical correction method resulted in the most accurate estimated true knowledge scores.

However, in some settings other correction methods resulted in the most accurate estimated true scores. Which method performed best depended on the average true knowledge score, the variability in true knowledge, the item difficulty, and the number of items in a test and the number of response categories. In a cohort with a high average true knowledge score, the Rasch method performed best. This may be explained by the effect that people with high true scores may hardly guess and therefore no guessing parameter, such as included in the 3PLM, is required. Similarly, when the variability in true scores in a cohort was high, the Rasch and weighted item difficulty method resulted in the highest correlation. Furthermore, for tests with higher item difficulty the Rasch and weighted item difficulty resulted in the most accurate estimated true scores. Also, correction methods that incorporated item difficulty information resulted in lower bias for 20 and 60 item tests. Here the results show that estimated true scores are overall less accurate with 20 items compared to 60 items, having less information to accurately estimate true scores.

The purpose of this study was to compare the accuracy of the different correction methods. When we order the methods based on the amount of information they take into account we can categorize them accordingly: the classical correction method which does not take sample information into account, the extended classical correction method which incorporates the sample's mean and variance, the beta binomial correction methods which take the distribution of total scores into account, and the methods that take into account information on item difficulty. The results showed that including item information is beneficial in correcting for guessing as long

as the methods do not become too complex. Although increased complexity might be useful in large-scale standardized tests, our results show that these methods give instable estimated true knowledge scores in the setting of small scaled non standardized tests, commonly used in higher education. Similarly, the beta binomial methods showed to be too complex and consequently did not outperform other correction methods. Finally, the results showed that incorporating some sample information as done by the extended classical correction method resulted in more accurate estimated true scores than the classical correction method. Only in the specific situation where tests did not have 40 items with four response categories, the classical correction method performed better than the extended classical correction methods. Interestingly, both classical methods performed suboptimal compared to the correction methods that include item difficulty information for these tests.

Overall, our findings apply to the specific context in which our study was placed; small-scale non-standardized MC tests. Consequently, they cannot be generalized to other situations such as large-scale standardized tests. For such tests the process of true score estimating is significantly different and the extent to which item information is available allows for more accurate true score estimation. For example, where omissions are observed in large-scale standardized tests, omissions may be imputed using answers on other items in the estimation. In the context of higher education tests, this is not possible and true score estimation is limited to the use of a correction for guessing. Furthermore, the accuracy of estimated true scores depends on various factors, such as, for example, the instruction of the teacher and the different strategies employed by the test-taker. In this study we focused on tests in which students' predominant strategy is to guess instead of omit answers. As a consequence, we did not study different guessing strategies under different correction methods or for different instructions in this simulation study, which is outside the scope of this study. Similarly, the focus was on the accuracy of estimated true knowledge scores, other reasons to correct for guessing than this accuracy, such as discouraging students' guessing behavior, were neither evaluated.

Based on the current results, it would be recommended to apply an extended classical correction method as developed by Calandra (1941) and Hamilton (1950) to correct

for guessing in higher educational MC tests. Although some situations existed for which another method, such as the Rasch model, resulted in more accurate estimated true scores (i.e., fewer bias or higher correlations), the extended classical correction method performed best in most situations. Furthermore, even though other methods outperformed the extended classical correction method in some specific situations, its performance still was not bad in these situations. Given this and the fact that in practice it is difficult to specify the specific situation one is in, the extended classical correction method is preferred overall. Specifying the specific situation one is in is difficult as cohort characteristics or test characteristics are often unknown and aspects such as the mean true score and exam difficulty are confounded in a particular test. Overall, when the extended classical correction method outperformed the other methods, the difference in bias with the classical method was small. Still, the extended performed consistently better in these situations. As pointed out by anonymous reviewers, the classical method might consequently be preferred over the extended classical in terms of easiness in explaining the method to students. However, to our opinion, this should not be a factor in determining which method to apply. In the end, the fairest method is not the one that is understood by students but the one that most accurately measures their true knowledge.

Both the extended classical correction method and the classical correction method, which performed well in our study, make use of the guessing parameter (α) in correcting for guessing. Their performance suggests that correction methods might become even more accurate when the α probability is more accurately estimated. In these methods α is defined as one divided by the number of response categories. For future studies, it would be interesting to see if alternative ways of estimating α lead to more accurate estimated true knowledge scores. For example, now it is assumed that students guess among all answer response categories. This, however, does not take into account any partial knowledge that students use to rule out one or more response categories when guessing. A proposed adjustment of α is to estimate α based on the number of response categories that were chosen by students in a test. In this case, when none of the students would choose alternative D on an item with four response categories, the guessing probability for this item would increase to 1/3. Another

proposition is to estimate α based on the lowest scoring students, who serve as a reference. Hereby, it could be assumed that the, for example, ten percent lowest scoring students guessed on all items. Consequently, the average of the mean proportion correct for these students could serve as an estimate of the guessing probability α . Future studies might focus on these kinds of adjustments of α by simulating response categories instead of correct/incorrect scores.

Throughout the years there have been several developments in methods to correct for guessing. However, in practice these developments are rarely applied to higher education tests. Contrary to the commonly used classical correction method, our simulation study showed that the extended classical correction method developed by Calandra (1941) and Hamilton (1950) or our proposed weighted item difficulty correction might result in more accurate estimated true knowledge scores in a practical higher education setting. In practice using the extended classical correction method might require taking additional steps to correct for guessing, this can be easily done by using a web application that is under development.

Appendix A. Method Elaboration

In this appendix a more detailed description of some the correction for guessing methods is provided.

Beta Binomial Correction Method

In the mixed binomial distribution heterogeneity among individuals is captured by assuming p has a (prior) Beta distribution with density $w(p)$ across the population (Morrison & Brockway, 1979). In the mixed binomial distribution, which allows for differences in individual's ability level, the probability that an individual with true score probability p has a number correct score R is then defined as:

$$(10) b_w(R, k) = \int_0^1 b(R, k, p) w(p) dp.$$

In order to get the probability of a number correct score R , given probability density $w(p)$, the aggregated mixed binomial distribution can be approached by a beta binomial distribution (Keats & Lord, 1962; Kendall & Stuart, 1969) which is defined as:

$$(11) P(R | \alpha, \beta) = \frac{1}{k+1} \frac{B(\alpha+R, \beta+k-R)}{B(R+1, k-R+1)B(\alpha, \beta)}.$$

Here, α and β are estimates of the first and second moments of the beta binomial distribution defined as:

$$(12) \hat{\alpha} = \frac{\bar{R}}{k} d$$

and

$$(13) \hat{\beta} = (1 - \frac{\bar{R}}{k}) d.$$

Where d equals:

$$(14) d = \frac{k(1 - \hat{r}_{xx'}) - 1}{\hat{r}_{xx'}}.$$

Here, the sample reliability, $\hat{r}_{xx'}$, is estimated by:

$$(15) \hat{r}_{xx'} = 1 - \frac{\bar{R}(1 - \frac{\bar{R}}{k})}{s_R^2}.$$

Having obtained the probability of a number correct score given probability p , true knowledge proportions are estimated using the cumulative distribution of p . For all possible total scores, R , the probabilities are summed.

$$(16) \widehat{tkp}_i = \int_0^R P(R) dR.$$

Extended Beta Binomial Correction Method with Three Parameters

Carlin and Rubin (1991) extended the beta binomial model as a special case of the one developed by Lord (1965) by only having the lower limit of the beta as a parameter and restricting the probability of a correct item response to be greater than some constant π_0 :

$$(17) P(R|\alpha, \beta, \pi_0) = \binom{k}{R} B(\alpha, \beta)^{-1} \sum_{j=0}^R \binom{R}{j} B(\alpha + j, \beta + k - R) \pi_0^{R-j} (1 - \pi_0)^{k-R+j}.$$

Here, π_0 is approached by the third factorial moment of the beta binomial distribution (see Carlin & Rubin, 1991; eq. 24). True knowledge proportions are calculated using Equation 16.

Three Parameter Logistic Model (3PLM) Correction Method

Having obtained the probabilities to answer an item correct, conditional on the true score and item parameters, true knowledge scores, $\hat{\theta}_i$ (latent trait scores) can be calculated using the vector of responses on items, $j = 1, \dots, k$, of participant i and the estimated item parameters. Subsequently, estimated true knowledge scores are obtained using:

$$(18) \widehat{tkp}_i = -\frac{\hat{\theta}_{min}}{\hat{\theta}_{max} - \hat{\theta}_{min}} + \left(\frac{1}{\hat{\theta}_{max} - \hat{\theta}_{min}} * \hat{\theta}_i \right).$$

Here, $\hat{\theta}_{min}$ and $\hat{\theta}_{max}$ refer to the minimum and maximum possible theta values given the response patterns. A response pattern with all incorrect items (assuming positive item discrimination parameters) corresponds to $\hat{\theta}_{min}$ and a pattern with all correct items (given positive item discrimination parameters) corresponds to $\hat{\theta}_{max}$. By taking both variability in students' response patterns and variability in item properties into account, the 3PLM will, at least asymptotically, result in the most accurate estimated true knowledge scores.

5

Comparing the Validity of Different Cut-Score Methods for Dutch Higher Education

This chapter is submitted as:

Yocarini, I. E., Bouwmeester, S., & Jongerling, J. (submitted). Comparing the validity of different cut-score methods in higher education.

Abstract

In Dutch higher education, budget and time constraints often limit the use of expert panels to set the standard. The aim of this study is to compare the accuracy of three cut-score (also referred to as standard setting) methods that are tenable for small-scaled, non-standardized Dutch higher education tests. In the classical absolute method, the cut-score is set at a specific percentage of test items. Alternatively, in compromise methods a relative aspect is introduced and sample information is taken into account. Hereby, the Cohen method uses information from the best performing students and the Hofstee method specifies the highest and lowest scorings students. Simulations were performed to obtain students' true and estimated grades and to obtain realistic higher education contexts by varying sample size, test difficulty, test discrimination, test length, and the number of response options. Both, the accuracy of the estimated grades and the pass/fail classification accuracy were evaluated. Generally, results show that the classical method mostly underestimates students' ability, while the Cohen method sometimes overestimates ability. Consequently, for higher education tests, taking into account some sample information in terms of the best performing students might be beneficial in estimating students' grades.

Keywords: assessment, cut-score methods, grade estimation, higher education, standard setting.

Introduction

In Dutch higher education, large-scale standardized tests are often unavailable and most tests are designed in-house. For these tests, the pass/fail cut-score (i.e., caesura) is often determined by individual academics (Sadler, 2014) because expert panels are most of the time not available due to time and budget constraints or because experts might be poor at judging item difficulty (e.g., Clauser, Clauser, & Hambleton, 2013; Clauser, Mee, Baldwin, Margolis, & Dillon, 2009; Impara & Plake, 1998; Van de Watering & Van der Rijt, 2006). As Kane (2017) explains, students with scores above the cut-score are assumed to have achieved an appropriate performance standard and those with scores below the cut-score have not. Consequently, the performance standard that is set at a student's ability level is translated into test scores to be able to make a pass/fail decision or to assign grades to test scores (a process also referred to as standard setting; Beuk, 1984; Reckase, 2006).

It is important that the chosen cut-score is valid, meaning that students who pass meet the performance requirements as specified in the curriculum and students who fail do not (Taylor, 2011). Instead of using expert panels to determine the performance standard on the level of students' ability, the cut-score in Dutch higher education tests is often based on a certain pre-fixed percentage of test items to be answered correctly (Cohen- Schotanus & Van der Vleuten, 2010). In this way, the percentage of test items to be answered correctly is viewed as a proxy for the percentage of knowledge a student has (i.e., the performance level). Unfortunately, little attention seems to be paid to the question of whether the percentage correct is a valid proxy for the required performance level of students in Dutch higher education. Where much literature is contributed to determining performance standards and setting cut-scores using panels (see e.g., Blömeke & Gustafsson, 2017; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Norcini, 2003; Reckase, 2006), little research has focused on how to determine cut-scores in this restricted higher education setting where panels are not used. In this study we evaluate the accuracy of different cut-score methods that are tenable in realistic Dutch higher educational contexts (i.e., non-standardized, small-scaled tests).

Overall, cut-score methods can be divided into three categories: absolute or criterion-referenced methods, relative or norm-referenced methods, and combinations of these two (Cohen-Schotanus & Van der Vleuten, 2010). While the cut-score in the absolute method is based on a specific ability level and in practice set at a specific percentage of items to answer correctly, the cut-score in the relative method is set at an ability level in comparison to other test-takers, operationalized as, for example, a certain percentile score. A compromise method combines the two. In choosing among cut-score methods, it is important that the standards set are consistent with the purpose of the test (Norcini, 2003) as different type of standards fit different purposes. For example, when using a test for the purpose of selecting a certain number of examinees, such as selecting the best job candidates, a relative standard clearly fits best. Contrary, for tests designed to assess competence, such as the diagnostic skills required by a clinical psychologist, absolute standards suit best.

In higher educational contexts, in which the purpose of tests is to determine whether an examinee has sufficient knowledge or skills in a particular domain, absolute cut-scores seem most appropriate (Norcini, 2003). In contrast, a relative method would only provide information on the ranking of students. Unfortunately, in Dutch higher educational courses the absolute cut-score is hardly ever related to the performance standards set in the curriculum (e.g., learning goals on individual courses or end qualification requirements for the entire curriculum). As such, the cut-score is often unrelated to the content of the items and the performance standard. In most of the methods applied in higher education (as explained below), for example, a pre-fixed percentage of items to answer correct is used to determine the cut-score. In European higher education this percentage is often set at 55% or 60% of test items, whereas in the UK the cut-score is often set at 50% (Cohen-Schotanus & Van der Vleuten, 2010). It is hereby implied that this percentage can be interpreted as a certain knowledge level that is required to pass the exam. Unfortunately, however, the link between the required knowledge level and the number of items answered correct to pass the test is hardly ever explicitly motivated, rendering the percentage rather arbitrary.

The situation is even more complicated as tests in practice differ in difficulty and may be quite unreliable as they are designed in-house by individual academics. As research has shown, experts (such as those designing the test) are not very good at estimating item characteristics for a given sample (e.g., Clauser et al., 2009; Clauser et al., 2013; Impara & Plake, 1998; Van de Watering & Van der Rijt, 2006), which might result in unreliable tests. If the required percentage of correct items to pass would be explicitly linked to the specific content of the items, varying test difficulty would be less problematic. As this is hardly done and tests in practice might be quite unreliable, the arbitrary percentage used to set the cut-score might be problematic in practice. To illustrate, a difficult test might result in many false negatives, students who fail but truly meet the study requirements. Whereas an absolute cut-score for a very easy test might result in many false positives, students who pass but do not truly meet the performance criteria. Consequently, when many students fail a test, it is not clear whether the test was too difficult or the true ability of a cohort was low and whether the cut-score should be adapted accordingly. In an attempt to tackle this issue, some study programs in higher education apply a compromise method. Whereas a fully relative method does not fit the purpose of educational tests, compromise methods incorporate only some information on students' performance on the test in the cut-score.

When the cut-score is determined using an arbitrary percentage of items that should be answered correct, whether in an absolute or compromise method, the question remains which cut-score method is best at calculating students' grades as a representation of students' ability level. Assuming students have a specific true grade on a test (i.e., the grade someone would obtain on a perfectly reliable test), the best cut-score method is the method that results in estimated grades on a test that are closest to the specific true grade the student has for that test. Several studies have evaluated the effects of using different cut-score methods (e.g., Cohen-Schotanus & Van der Vleuten, 2010; Dochy, Kyndt, Baeten, Pottier, & Veestraeten, 2009). These studies have compared different methods relative to each other by looking at aspects such as the fluctuations in cut-scores or failure rates across different methods. Although this provides information that might be useful to assess the practical

consequences of implementing a specific cut-score method, it does not provide information on the accuracy of each individual method. To get this information, true grades are required as a reference to compare to the estimated grades using the different cut-score methods. Since these true grades are not available in practice, simulations are performed to obtain them in this study. In addition, previous studies evaluated the dichotomous pass/fail decision instead of looking at grade accuracy over the entire range of grades. Where pass/fail scores are the main focus at the institutional level, the exact grade obtained on a test is most important at the level of the individual student. For example, specific grade requirements often apply for graduating with honors, entering a specific master program, applying for a specific job, or when compensation between (a cluster of) courses is allowed. Consequently, in comparing different cut-score methods in this study, we also focus on the accuracy of the estimated grades. Furthermore, this study looks at multiple choice (MC) tests.

Method

Methods to Calculate Grades in Higher Education Tests

Three cut-score methods that are tenable in a Dutch higher education context are included in this study: the classical cut-score method, the Cohen method, and the Hofstee method. In light of the arbitrariness of the chosen percentages, grades were also calculated without the use of a cut-score, as a fourth possibility to estimate students' ability level using tests scores.

Classical cut-score method. Classically, the absolute cut-score for a Dutch higher education test is set by determining a pre-fixed percentage of items a student has to answer correct in order to pass, after correcting for random guessing:

$$\text{cut score} = cN + x(N - cN)$$

Where N refers to the total number of items in a test, c to the proportion of items answered correct due to guessing in the test, and x to the pre-fixed percentage of items answered correct at a pass. To explain this method, let's take the Dutch higher education as an illustration and suppose the grading scale runs from 1 to 10 and the cut-score is set at 55% of the total items after correcting for guessing, which results in a passing grade of 5.5. Further, assume that we have a test with 40 items that each

have 4 answer categories. Here, 30 items would remain after we correct for guessing ($\frac{1}{4} * 40 = 10$), of which 55% corresponds to 16.5 items. Adding the 10 items that should be answered correctly when students would have purely guessed, this results in a cut-score of 26.5 ($10 + 16.5 = 26.5$), for which students would receive a grade of 5.5:

$$\text{cut score} = \frac{1}{4} * 40 + 0.55(40 - \frac{1}{4} * 40) = 26.5$$

When we plot grades (on the y-axis) against the number of items answered correctly on a test (on the x-axis) a passing grade corresponds to the coordinate of (26.5, 5.5). A perfect 10 is scored when all items are answered correct, making the coordinate of the maximum possible grade achievable (40, 10) in our example. Connecting these two coordinates by a line (i.e., interpolating) determines the grades assigned to different total test scores. This is illustrated in Figure 1 by the solid line. As shown, grades below 1 are typically set to 1. Note that this is only one example, in the simulations the percentages were varied (lowering to 50% and increasing it to 60%) resulting in different cut-scores.

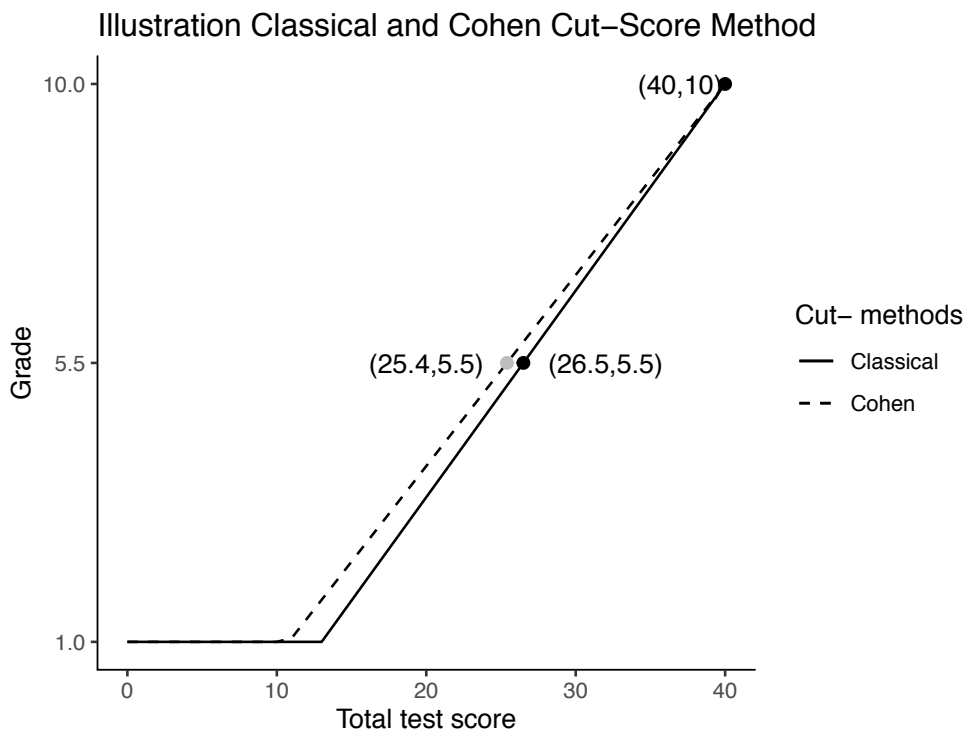


Figure 1. Illustration of the grades assigned to total test scores using a classical or Cohen cut-score method.

Cohen method. A second approach to consider in setting the cut-score in higher education tests is the compromise method proposed by Cohen-Schotanus & Van der Vleuten (2010). In the Cohen method, the cut-score is determined by using a pre-fixed percentage of items answered correctly for a passing grade as well. Additionally, the best performing students are used as a point of reference for the difficulty of the test, using the test score of a specific percentile score to adjust the pre-fixed percentage:

$$\text{cut score} = cN + x(N^* - cN).$$

Here, N^* refers to the score of the n^{th} percentile score (Cohen-Schotanus & Van der Vleuten, 2010). To illustrate, if the 95th percentile score would, for example, be 38 on a 40-item four-response-options test with items having a probability of random guessing of 1/4, the resulting cut-score would be 25.4:

$$\text{cut score} = \frac{1}{4} * 40 + 0.55 \left(38 - \frac{1}{4} * 40 \right) = 25.4.$$

As illustrated by the dashed line in Figure 1, the coordinates of the maximum grade (40, 10) and the cut-score (25.4, 5.5) are connected by a line to transform the total test scores into grades. In this method, more information on the sample characteristics is taken into account in setting the cut-score. However, note that when the top students score the maximum score of 40, the cut-score is similar to that of the classical method in which a similar pre-fixed percentage is chosen. Also, if the maximum test score is lower than 40, the maximum grade is still set at (40, 10). Hereby, this method assumes that the top percentage of students is a stable and therefore reliable group to use as a proxy for the difficulty of a test. Furthermore, Cohen-Schotanus and Van der Vleuten (2010) choose to set the pre-fixed percentage at 60 percent, while a percentage of 55 is often used in the classical cut-score method as done in our illustration. In the simulations in this study, the percentage was varied, as well as the percentile used as a reference for test difficulty.

Hofstee method. Another compromise method included is the Hofstee method (Hofstee, 1983). In this method, the minimum and maximum acceptable cut-scores are specified (k_{min} and k_{max}), as well as the minimum and maximum acceptable failure rates (f_{min} and f_{max}). Using these values, the cut-score is determined at the point where the line between the coordinates (f_{min}, k_{max}) and (f_{max}, k_{min}) intersects

the distribution of the observed test scores (for a detailed description of this method, see Hofstee, 1983). Consequently, using the cut-score as a reference, a line is drawn from the maximum obtainable grade through the reference point, similar to the way in which test scores are transformed to grades as shown in Figure 1. Whereas the Hofstee method requires panelist to determine the acceptable cut-scores and failure rates, in Dutch higher education these percentages may in practice be chosen by individual instructors themselves. Consequently, a set of percentages considered relevant was used in this study to be able to include the method as a comparison. From the three approaches, the Hofstee method is the most relative approach as both the minimum and maximum allowed failure rates are specified.

No cut-score. As mentioned, the percentage of 55% or 60% of items to answer correct in order to pass a test is rather arbitrary. Consequently, one could choose to remove this arbitrary point and determine students' grades without the cut-score as a point of reference. In this situation, a line connects two coordinates (here: (1, 10) and (10, 40) for the minimum and maximum points, respectively) corresponding to the number of correct items for both the minimum (taking random guessing into account) and maximum obtainable grade. Using this method, there is no cut-score to use as a reference point in advance. Continuing our example, the test score corresponding to a grade of 5.5 would be 25 ($25 = \frac{5.5-1.0}{\frac{10-1}{40-10}} + 10$).

In this study we were especially interested in evaluating the classical absolute standard (i.e., using a pre-fixed percentage of correct items) that is often used as a standard in Dutch higher education. Furthermore, we were interested to assess the influence of the arbitrariness of the percentage used in the cut-score method on the accuracy of the estimated grades. In addition, cohort and test characteristics were varied to mimic realistic higher educational settings. Specifically, we varied the cohort sizes, number of items on a test, test difficulty, test discrimination, the average true grade in a cohort, and the variance of the true grades in a cohort.

Simulation Procedure

To evaluate the different cut-score methods, simulations were performed using R (R core Team, 2016) similar to the procedure used in Chapter 4. First, students'

underlying ability levels (i.e., thetas) were simulated, where the theta scores form a continuous normal distribution: $N \sim (\mu_\theta, \sigma_\theta)$. Here μ_θ refers to the latent ability score of the sample, which was assumed to be 0 following the standard normal distribution, and σ_θ refers to the variation in the latent ability scores across students, which was assumed to be 1. A standard normal distribution was assumed because academic performance is the sum of multiple subparts. As stated by the central limit theorem of Lindeberg and Lévy (Billingsley, 1961), the sum of a large number of independent and equally distributed stochastic variables with finite variance approximates a normal distribution.

Second, the theta scores were transformed to true grades. These true grades were used to evaluate the accuracy of the estimated grades across the different cut-score methods. A linear transformation was performed:

$$\text{true grade}_i = b_0 + b_1 \theta_i.$$

Here, b_0 is the grade at the average theta score of zero (θ_0), this true grade average was set to 6.0 (on a scale of 1.0 to 10.0, b_{min} and b_{max} , respectively), following our previous illustration of Dutch higher education. Given prior education requirements, we assumed the probability that a student accidentally gets into higher education (i.e., is a false positive) with a true grade of 1.0 or smaller to be extremely small, about 1 in 1000 students. Assuming about 0.1% of students to have a minimum true grade of 1.0, $b_1 = \frac{b_0 - b_{min}}{\theta_0 - \theta_{min}} = \frac{6.0 - 1.0}{0 - -3.09}$. Here, -3.09 (θ_{min}), refers to the standard theta value for which the probability to obtain this theta or lower equals 0.1%. Note that for comparison of the true grades to the estimated grades, true grades smaller than 1.0 were rounded to 1.0 and scores larger than 10.0 were rounded to 10. In this way both true and estimated grades were on the same scale. Taking these aspects into account, there were relatively more true grades close to the maximum score of 10 than to the minimum score of 1.0, which is a realistic assumption.

The choices of the average true grade and the proportion of minimum grades in transforming theta scores to true grades is of course arbitrary and is hard to account for with practical evidence. To evaluate whether the results depended on these arbitrary choices, a sensitivity analysis was performed in which different theta to true

grade transformations were incorporated. In this sensitivity analysis, we used two alternative assumptions about the average true grade at a theta of zero (b_0), lowering it to 5.5 and increasing it to 6.5. Additionally, we varied the percentage of minimum grades in the population, lowering it to 0.05% ($\theta_{min} = -3.29$), and increasing it to 0.5% ($\theta_{min} = -2.58$), 1% ($\theta_{min} = -2.33$), and 2.5% ($\theta_{min} = -1.96$). In these alternative transformations, $b_1 = \frac{6.0-1.0}{0--3.29}$, $b_1 = \frac{6.0-1.0}{0--2.58}$, $b_1 = \frac{6.0-1.0}{0--2.33}$, and $b_1 = \frac{6.0-1.0}{0--1.96}$, respectively. The results of the different transformations also provided information on the performance of the different cut-score methods in different samples, in terms of the average ability level and variation thereof.

Third, item responses were simulated for each student i for each item j , using the students' theta scores. To do so, the three-parameter logistic model (3PLM; Birnbaum, 1968) from item response theory (IRT) was applied. By using this IRT model the probability of giving a correct response on a test item was simulated conditional on the specific theta value of a student and the characteristics of the test item in terms of item difficulty, item discrimination, and the probability of guessing on the item:

$$P(X_j = 1 | \hat{\theta}_i, \hat{\delta}_j, \hat{\alpha}_j, \hat{\pi}_{0j}) = \hat{\pi}_{0j} + (1 - \hat{\pi}_{0j}) \frac{\exp(\hat{\alpha}_j(\hat{\theta}_i - \hat{\delta}_j))}{1 + \exp(\hat{\alpha}_j(\hat{\theta}_i - \hat{\delta}_j))}$$

Here $\hat{\theta}_i$ refers to the student i 's true score, $\hat{\delta}_j$ to the item j 's difficulty, $\hat{\alpha}_j$ to the item j 's discrimination, and $\hat{\pi}_{0j}$ to the item j 's guessing probability. Table 1 shows the values that were used for these parameters. The item difficulty and item discriminability values were based on data of first year Psychology students. As most tests consist of four item response categories in Dutch higher education, the average guessing probability was set to .25 with a standard deviation of .05. This standard deviation was implemented to allow for less or more guessing as we believe students differ in their willingness to guess on items (see e.g., Budescu & Bo, 2015) and differ in the amount of partial knowledge they have which might let them eliminate one or two item response categories. The ltm package in R (Rizopoulos, 2006) was used to calculate the probability to answer an item correctly given the item parameters and a student's true score. These probabilities were then converted into correct ($X_j = 1$) and

incorrect ($X_j = 0$) item scores using the probability to answer an item correct of .5 as the cut-score. See Appendix A for an example of the code that was used.

Included Variables and Values

Table 1 shows variables that were varied to evaluate the accuracy of the different cut-score methods within realistic higher educational contexts, in addition to the average true grade and the proportion of minimum true grades. The sample sizes varied such that smaller classes of more specialized study programs were represented as well. Similarly, the number of items and response categories varied to assess the performance of cut-score methods for different types of realistic tests. The specific values were chosen according to the observation that most Dutch MC tests consist of 40 items with each having four item response categories. If more items are included in a test, they generally have fewer item response categories. In the three different cut-score methods described in the introduction, different arbitrary choices are made in practice, such as the percentage of items answered correctly required to acquire a passing grade. In this study, we were not only interested in which method was most accurate but also to what extent these arbitrary choices influenced the accuracy. In the classical and Cohen method the percentage of items to have answered correct at the cut-score was varied. Additionally, in the Cohen method the size of the group of best performing students used to determine the cut-score was varied.

Table 1: Overview simulated variables and values

	Variable	Notation	Value(s) ¹	
Sample properties	Average true grade	$\mu_{true\ grade}$	5.5, 6.0 , 6.5	
	Proportion of minimum true grades	$p_{true\ grade\ 1.0}$	0.05, 0.1 , 0.5, 1, 2.5	
	Size sample	n	20, 200, 400	
Test properties	Number of items	k	20, 40 , 50, 60	
	Number of item response options	a	2, 3, 4	
	Average item difficulty	μ_{δ}	-.8, 0 , .8	
	Variability in item difficulty	σ_{δ}	.5	
	Average item discriminability	μ_{α}	.3, .5, 1.5, 2	
	Variability in item discriminability	σ_{α}	.3	
	Average guessing probability	μ_{γ}	.25	
	Variability in guessing probability	σ_{γ}	.05	
Method properties	Percentage of items to answer correctly		50%, 55% , 60%	
	Percentile reference group Cohen method		90 th , 95th	
	Hofstee	Min. acceptable cut-score	k_{min}	.50
		Max. acceptable cut-score	k_{max}	.60
		Min. acceptable failure rate	f_{min}	0
		Max. acceptable failure rate	f_{max}	1

¹Values in bold illustrate the standard value used when the specific variable was not manipulated.

In total all these variations resulted in 18 scenarios (see Table 2) in which datasets were simulated. Note that only combinations that were considered relevant and realistic for Dutch higher education settings were simulated. For each of these scenarios, all possible theta to true grade transformations were evaluated and for each scenario 1000 datasets were simulated to obtain stable results.

Table 2: Scenarios for Simulated Datasets

Scenario	n	μ_δ	μ_α	k	a	Percentage Best Students in Cohen Method	Pre-fixed Percentage of Items to Answer Correctly
1	400	0	.5	40	4	.95	.55
2	200	0	.5	40	4	.95	.55
3	20	0	.5	40	4	.95	.55
4	400	.8	.5	40	4	.95	.55
5	400	-.8	.5	40	4	.95	.55
6	400	0	.3	40	4	.95	.55
7	400	0	1	40	4	.95	.55
8	400	0	1.5	40	4	.95	.55
9	400	0	2	40	4	.95	.55
10	400	0	.5	20	2	.95	.55
11	400	0	.5	20	3	.95	.55
12	400	0	.5	20	4	.95	.55
13	400	0	.5	60	2	.95	.55
14	400	0	.5	60	3	.95	.55
15	400	0	.5	60	4	.95	.55
16	400	0	.5	40	4	.90	.55
17	400	0	.5	40	4	.95	.50
18	400	0	.5	40	4	.95	.60

Note: cells in grey indicate the manipulated variable in comparison to the first scenario.

Outcome Measures

The accuracy of the different standard setting methods was evaluated by looking at different measures. First, by evaluating the square root of the mean sum of squared error (MSE):

$$MSE = \sqrt{\frac{\sum_{i=1}^n (\widehat{grade}_i - true\ grade_i)^2}{n}}$$

Here, \widehat{grade}_i refers to the estimated grade of student i after applying one of the cut-score methods to the simulated test score. In addition to the overall MSE measure, the MSE was also constructed for different grade windows to assess whether the cut-score methods differ in the location of (in)accuracy: grade 1-4 (low grades), grade 4-6 (average grades), and grade 6-10 (high grades). As a descriptive, the average true and estimated grade were evaluated to assess the extent of the (in)accuracy between the

grades and see if there is an over- or underestimation. Here, the true and estimated grades were evaluated across the separate windows as well. Furthermore, the correlation between the true and estimated grades was evaluated for all grades, as well as for the grades in the specific windows.

Given the pass/fail nature of cut-scores, classification rates were also evaluated. Here, the focus is not on the specific grade but on the dichotomous pass/fail decision that motivates the use of a cut-score. Specifically, we focused on the sensitivity, the specificity, the total proportion of misclassifications, and the positive predictive value. Sensitivity refers to the proportion of correctly passed students from all students that should have passed based on their true grade, that is: $Sensitivity = \frac{TP}{TP+FN}$, where TP denotes the true positives and FN the false negatives. Specificity refers to the proportion of correctly identified failed students, given all students that should have failed based on their true grade, that is, $Specificity = \frac{TN}{TN+FP}$, where TN denotes the true negatives and FP the false positives. The proportion of misclassifications is the proportion of all misclassified students given the entire cohort, $Total\ miss = \frac{FP+FN}{TN+TP+FP+FN}$. The positive predictive value (PPV) shows the proportion of students that had a true grade above the cut-score from all the students that passed: $= \frac{TP}{TP+FP}$.

Results

Given the vast amount of results, results for the first scenario are discussed in detail for each outcome measure while tables with the outcome measures for scenario 2 to 18 as shown in Table 2 can be found at our Open Science Framework (OSF) following the link: <https://osf.io/jgsx2/>. For scenario 1, the grade accuracy and classification accuracy are discussed in different sections. In addition to the inspection of the grade accuracy for separate windows of grades (below 4, between or equal to 4 and 6, and above 6), the classification accuracy provides an indication of the accuracy of each method in two windows (below and above the cut-score). As these classification rates already provide us with information on two windows, the grade accuracy per window is only discussed in terms of relevant patterns. For the detailed results per window visit our OSF page to find these tables. Additionally, the most important deviations

from the results of scenario 1 will be discussed in the final section for each of the manipulated variables.

Overall, the results showed that not using a cut-score in estimating grades resulted in the least accurate grades and classification rates in almost all scenarios and transformations. Therefore, this method is hereafter not taken into account in the Results section (except for the results portrayed in the tables). In discussing the results, we are mainly interested in which method performs best, followed by the influence of the specific transformation (i.e., the results of our sensitivity analysis of the assumptions on the average true grade and proportion of minimum 1.0 true grades in the cohort).

Accuracy of Grades

MSE. As shown in bold in Table 3, the classical cut-score method resulted in the highest MSE values, except for cohorts with a low average true grade of 5.5. Here, the Hofstee method resulted in the largest MSE. Notably, the differences in MSE values across the methods increased considerably (from about a 0.25 difference to a 0.92 difference) as the average true grade increased and the differences slightly decreased (by 0.3 for cohorts with a high true grade and only 0.07 for cohorts with a low true grade) as the proportion of the true grades of 1.0 (i.e., the minimum true grade) increased. Overall, the MSE values became larger as the average true grade increased, where the increase was larger for the classical and no cut-score method (about 1.1) compared to the Cohen and Hofstee method (having a 0.75 increase). Also, MSE values increased as the proportion of minimum true grades increased (by .3-.4 for the classical and no cut-score method, and by .65 for the Cohen and Hofstee method). Evaluating the MSE per window shows the Hofstee and Cohen method have the highest MSE for grades below 4, while the classical method results in the highest MSE values for grades above 6.

Table 3: MSE and Correlation for Different True Grade Transformations for Scenario 1

True Grade Transformation		MSE				Correlation			
Mean	Prop. 1.0 true grades	Class ¹	Cohen	Hof	None	Class	Cohen	Hof	None
5.5	.0005	1.15	1.08	1.10	1.39	0.74	0.74	0.74	0.74
5.5	.001	1.17	1.12	1.14	1.42	0.74	0.74	0.74	0.74
5.5	.005	1.28	1.29	1.30	1.53	0.74	0.74	0.74	0.74
5.5	.01	1.37	1.40	1.42	1.62	0.74	0.74	0.74	0.74
5.5	.025	1.56	1.63	1.65	1.80	0.74	0.74	0.74	0.74
6.0	.0005	1.47	1.02	1.02	1.82	0.74	0.74	0.74	0.74
6.0	.001	1.50	1.09	1.09	1.85	0.74	0.74	0.74	0.74
6.0	.005	1.61	1.30	1.30	1.96	0.74	0.74	0.74	0.74
6.0	.01	1.70	1.44	1.44	2.04	0.74	0.74	0.74	0.74
6.0	.025	1.88	1.69	1.70	2.19	0.74	0.74	0.74	0.74
6.5	.0005	1.87	1.20	1.18	2.27	0.74	0.74	0.74	0.74
6.5	.001	1.90	1.26	1.25	2.30	0.74	0.74	0.74	0.74
6.5	.005	2.00	1.47	1.46	2.39	0.74	0.74	0.74	0.74
6.5	.01	2.07	1.60	1.59	2.45	0.74	0.74	0.74	0.74
6.5	.025	2.21	1.84	1.84	2.57	0.73	0.73	0.73	0.73

¹Class refers to the classical method. Note: in scenario 1 $N = 400$, mean difficulty was set at 0, mean discrimination 0.5, tests had 40 items with 4 answer alternatives, the Cohen percentile was set at .95 and the pre-fixed percentage of items to answer correct at 55% in the classical and Cohen method.

Grades. Overall, the average estimated grade using the classical method was always lower than the average true grade, resulting in an underestimation of the true grades as shown in Table 4. The Cohen and Hofstee method resulted in quite similar average estimated grades, which were an overestimation of the true grades for the cohorts with a low average grade and an underestimation for cohorts with a high average true grade. Overall, for cohorts with a low true grade, the overestimation was largest for the Hofstee method. For cohorts with a higher true grade, the underestimation was largest for the classical method estimates. The results per window showed that the specific transformation only mattered slightly for the under- or overestimation of the cut-score methods. For grades below 4 the proportion of minimum true grades slightly influenced the classical methods' accuracy as it underestimated true grades when few minimum grades occurred and overestimated true grades in cohorts with more

minimum grades. For grades above 6, the Cohen and Hofstee method underestimated grades except when the average grade and proportion of true grades were both low.

Table 4: Mean and SD for Different True Grade Transformations over all Grades for Scenario 1

Theta to True Grade Transformation		True Grade	
Mean	Proportion of 1.0 True Grades	Mean	SD
5.5	.0005	5.50	1.37
5.5	.001	5.50	1.46
5.5	.005	5.50	1.73
5.5	.01	5.49	1.90
5.5	.025	5.49	2.20
6.0	.0005	5.99	1.52
6.0	.001	5.99	1.61
6.0	.005	5.98	1.90
6.0	.01	5.98	2.07
6.0	.025	5.95	2.37
6.5	.0005	6.48	1.65
6.5	.001	6.48	1.74
6.5	.005	6.45	2.03
6.5	.01	6.43	2.20
6.5	.025	6.38	2.49
Methods to calculate grades	Classical	5.00	1.49
	Cohen	6.03	1.19
	Hofstee	6.07	1.17
	No cut-score	4.50	1.34

Correlation. As can be seen in Table 3, the correlation between the true and estimated grades using all grades was about .74 for all cut-score methods, across all transformations. For the different windows, a similar pattern was observed, except for grades above 6. Here, the classical cut-score method resulted in lower correlation values compared to the other three methods.

Classification accuracy

Sensitivity. In general, the sensitivity (i.e., the proportion of students that passed, from all those that should have passed) was lowest for grades estimated with the classical cut-score method, as can be seen in Table 5. The difference in sensitivity between the methods only slightly increased as the average true grade increased. Overall, sensitivity decreased a bit as the average grade increased. Only for cohorts with higher average true grades the sensitivity slightly increased as the proportion of true grades of 1.0 increased. So, the classical cut-score method resulted in the highest false negative rate.

Specificity. The specificity (i.e., the proportion of students that failed, from all those that should have failed) was lowest for the Hofstee method, followed by the Cohen method as can be seen in Table 5. The difference in specificity across the methods decreased as the average true grade increased and slightly increased as the proportion of true grades of 1.0 increased in a cohort with an average true grade of 6.0 or 6.5. Whereas the specificity values for the grades estimated using the classical method were not influenced much by the transformation applied, the specificity of the Cohen and Hofstee method increased as the average true grade increased and slightly decreased as the proportion of true grades of 1.0 increased for cohorts with higher average true grades. So, the false positive rate was highest for the Cohen and Hofstee methods.

Total proportion of misclassifications. As shown in Table 5, the total proportion of misclassifications was highest for the Hofstee and Cohen method when the average true grade was low. When the average true grade was higher the classical correction method resulted in the highest proportion of misclassifications. In general, the differences between the methods increased as the average true grade increased as well. Also, the difference between the methods decreased as the proportion of true grades of 1.0 increased for these cohorts with a higher average. Overall, the proportion of misclassifications increased a bit for grades estimated using the classical method as the true grade average increased and decreased slightly as the proportion of true grades of 1.0 increased in cohorts with an average of 6.0 or 6.5.

Table 5: Sensitivity, Specificity, Total Misclassifications, and Positive Predictive Value for Different True Grade Transformations for Scenario 1

True grade transformation	Sensitivity				Specificity				Total Misclassification				Positive Predictive Value			
	Class ¹	Cohen	Hofstee	None	Class	Cohen	Hofstee	None	Class	Cohen	Hofstee	None	Class	Cohen	Hofstee	None
Mean																
Prop. 1.0																
grades ²																
5.5	.64	.92	.93	.42	.88	.54	.51	.96	.24	.27	.28	.31	.84	.67	.66	.91
5.5	.64	.92	.93	.42	.88	.54	.51	.96	.24	.27	.28	.31	.84	.67	.66	.91
5.5	.64	.92	.93	.42	.88	.54	.51	.96	.24	.27	.28	.31	.84	.67	.66	.91
5.5	.64	.92	.93	.42	.88	.54	.51	.96	.24	.27	.28	.31	.84	.67	.66	.91
5.5	.64	.92	.93	.42	.88	.54	.51	.96	.24	.27	.28	.31	.84	.67	.66	.91
6.0	.56	.87	.89	.35	.92	.63	.60	.98	.31	.22	.22	.42	.92	.80	.79	.97
6.0	.56	.88	.89	.35	.92	.62	.59	.98	.30	.22	.22	.41	.92	.79	.78	.96
6.0	.58	.89	.90	.36	.91	.61	.58	.98	.29	.22	.23	.39	.91	.77	.76	.96
6.0	.58	.89	.90	.37	.91	.60	.57	.97	.28	.23	.23	.38	.90	.76	.75	.95
6.0	.59	.89	.91	.38	.91	.59	.56	.97	.28	.23	.24	.37	.89	.75	.74	.95
6.5	.51	.83	.85	.31	.95	.70	.67	.99	.37	.20	.20	.50	.96	.88	.87	.99
6.5	.51	.84	.86	.32	.95	.69	.66	.99	.36	.20	.20	.49	.96	.87	.86	.98
6.5	.53	.85	.87	.33	.94	.67	.64	.98	.34	.21	.20	.46	.95	.84	.83	.98
6.5	.54	.86	.88	.34	.93	.65	.62	.98	.33	.21	.21	.45	.94	.83	.82	.97
6.5	.55	.87	.89	.35	.93	.64	.61	.98	.31	.21	.21	.42	.93	.81	.80	.97

¹Class refers to the classical method. ²Proportion 1.0 true grades.

For the Cohen and Hofstee methods this was the other way around and the proportion of misclassifications decreased slightly as the average true grade increased. Overall, for most transformations, the Cohen method resulted in the lowest proportion of misclassifications.

Positive predictive value. The positive predictive value (i.e., all students that should have passed from those that passed) was lowest for grades estimated using the Hofstee and Cohen method, as shown in Table 5. The difference in the positive predictive value between the methods decreased somewhat as the true average grade increased and the difference between the methods became slightly larger as the proportion of true grades of 1.0 increased in cohorts where the average was 6.0 or 6.5. Overall, the positive predictive values increased as the average true grade increased, with the increase being somewhat larger for the Hofstee and Cohen method than the classical method. The positive predictive values slightly decreased as the proportion of minimum true grades increased for cohorts with an average true grade of 6.0 or 6.5. Overall, the classical cut-score method had the highest positive predictive value.

Results per Variable

Sample size. For smaller samples, the patterns in MSE were similar to those observed in scenario 1 in which 400 students were included. Only, the Cohen method showed differences, where it resulted in the highest MSE for some transformations in very small cohorts of 20 students. As sample sizes were smaller, the average estimated grade for the Cohen method became higher. Additionally, the Cohen method resulted in the largest proportion of misclassifications for cohorts with an average true grade. Overall, the specificity and positive predictive value were slightly lower for the Cohen method, thereby slightly increasing the differences between the methods for these classification rates. Thus, the Cohen method seems less accurate for small sample sizes.

Test difficulty. Varying test difficulty showed different results for the MSE values as illustrated in Figure 2. With increasing test difficulty, the Cohen and Hofstee method performed best for most cohorts, while for easier tests, the classical method most

often resulted in the lowest MSE values. As shown, MSE values, as well as the difference in them across methods increased as test difficulty increased. Similarly, the average estimated grades decreased as test difficulty increased. Whereas the degree of over- and underestimation for the Cohen and Hofstee method decreased as tests were more difficult (depending on the specific grade window), the degree of underestimation increased for the classical method to a larger extent (throughout all windows). Furthermore, correlation values slightly decreased as test difficulty increased. Additionally, when test difficulty increased, sensitivity values decreased and more so for the classical method, thereby increasing the differences in sensitivity between the methods. For easier tests, sensitivity values increased, decreasing the differences in methods. The specificity values increased for the Cohen and Hofstee method as difficulty increases (decreasing the differences), and the values decreased for all methods as tests were easier. When test difficulty increased, the proportion of misclassifications was highest for the classical method, regardless of the transformation. Consequently, the differences between the methods increased for more difficult tests. For easier tests, the Hofstee method mostly resulted in the highest proportion of misclassifications. Finally, the positive predictive value increased as tests were more difficult and decreased as tests were easier. To summarize, for difficult tests differences between the methods become more pronounced, where the Cohen method outperformed the other methods, whereas the classical method was superior for easier tests.

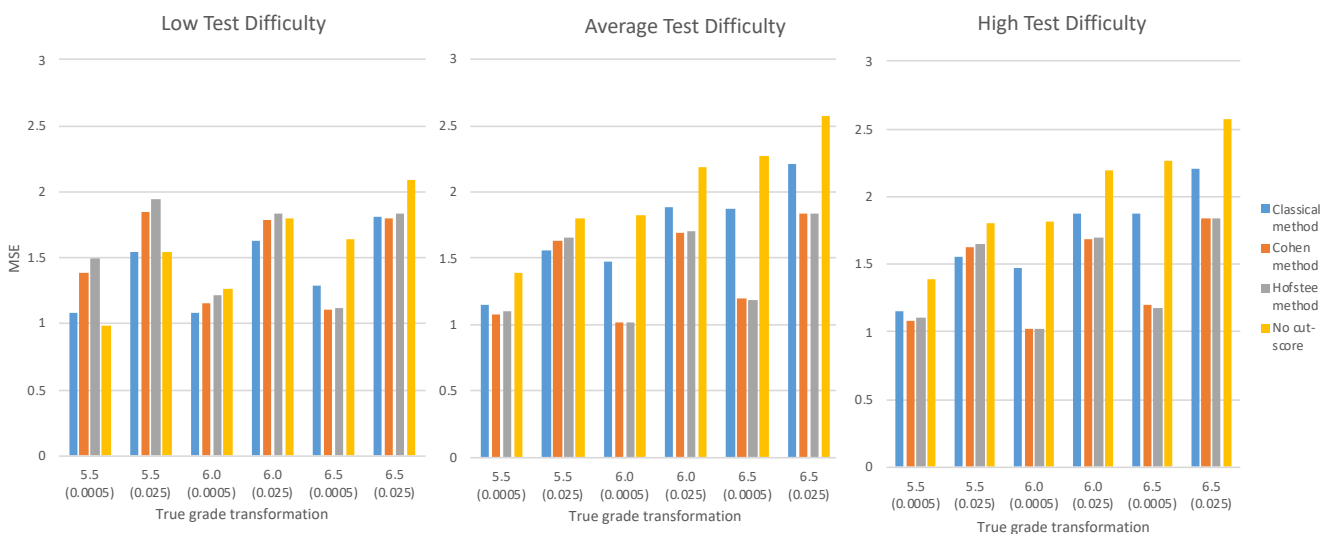


Figure 2. MSE values for varying test difficulty from -0.8, through 0, to 0.8.

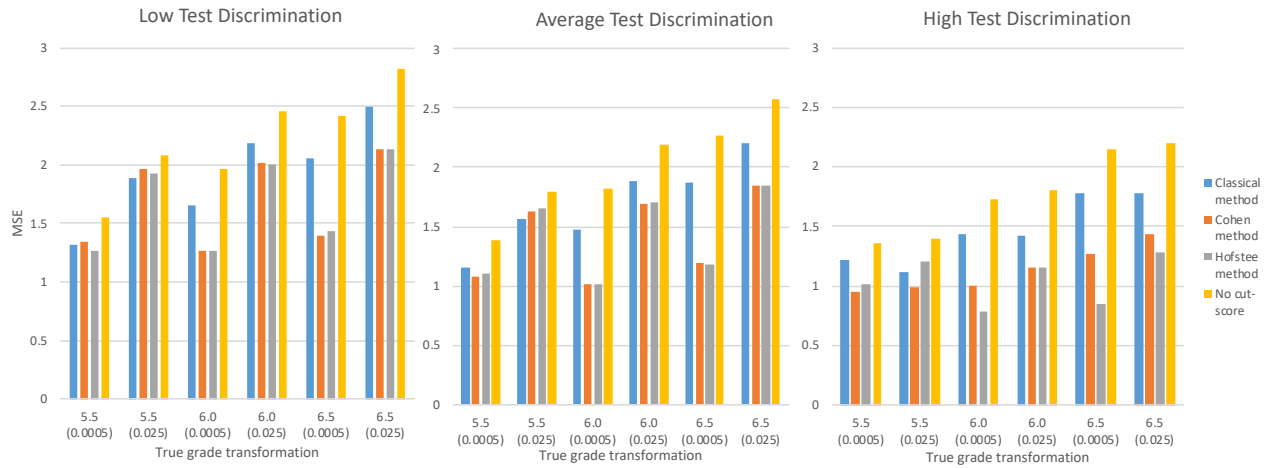


Figure 3. MSE values for varying test discrimination from 0.30, 0.50 to 1.0.

Item discrimination. The MSE values for varying item discrimination are shown in Figure 3 for three of the simulated scenarios. As can be seen, the Cohen method resulted in the highest MSE when item discrimination was low, for cohorts with a low average true grade. As item discrimination increased, the classical method resulted in the largest MSE values regardless of the transformation. Overall, MSE values increased as test discrimination decreased and MSE decreased as test discrimination increased (here, the differences between the methods became slightly more pronounced). With decreased item discrimination, the average estimated grade for the Cohen method increased, while that of the others remained the same estimate. This increased the overestimation of the Cohen method for grades below 6. As item discrimination increased, the average estimated grade slightly increased for the classical and Hofstee method, while that of the Cohen method decreased to a relatively larger extent. As a consequence, the Cohen method resulted in an underestimation and overestimation similar to that of the classical method as test discrimination increased. Overall, as discrimination decreased, the correlation values decreased as well, and vice versa. Furthermore, the classification accuracy decreased as discrimination decreased. For the proportion of misclassifications, the Cohen method had the largest proportions for cohorts with a low average true grade. Overall, the differences in the classification measures became larger as discrimination decreased. As item discrimination increased, the classification accuracy increased and the differences in sensitivity, specificity, and positive predictive value between the methods decreased. Thus, differences between methods were smaller for tests with

high discrimination, and for tests with lower discrimination especially the Cohen method was affected, resulting in weaker performance for cohorts with low average true grades.

Test length and the number of response options. The results for different simulated tests showed different patterns compared to the first scenario where tests had 40 items with 4 response options. As shown in Figure 4, test length was of influence as the classical method resulted in the highest MSE values for shorter tests, regardless of the transformation applied. For these short tests, the average estimated grade of the Cohen method decreased, while that of the Hofstee method slightly increased. For longer tests, the Cohen method resulted in the largest MSE values for cohorts with an average true grade of 5.5, as the MSE values of the classical method decreased. Here, the average estimated grade of the Cohen method slightly increased. Furthermore, the correlation values decreased for shorter tests and increased for longer tests. Overall, the classification accuracy decreased for shorter tests. The sensitivity and specificity values slightly decreased, the positive predictive value decreased only for the classical method, and the proportion of misclassifications increased. Overall, this showed that longer tests resulted in higher accuracy than shorter tests.

Second, the number of response options showed to be of considerable influence, as shown by the MSE values in Figure 5 in which 60 item tests with different number of response options are portrayed (note that the 20 item tests showed similar results, yet slightly more pronounced). Overall, the MSE values increased as the number of response options decreased, thereby increasing the differences in MSE among the methods where the classical method had the highest MSE values. With fewer response options, the average estimated grades decreased for the classical and Cohen method (with a stronger decrease for the first). Furthermore, correlation values showed differences among the methods as the correlation values of the classical method were lower with fewer response options. Also, with fewer response options, the sensitivity values decreased considerably, while the specificity, proportion of misclassifications, and positive predictive value increased (relatively more for the Cohen method but not for the Hofstee method). Overall, the classical method had the largest proportion of misclassifications in tests with only few response options. So, in general more

response options results in higher accuracy, with this effect being most pronounced for the classical method.

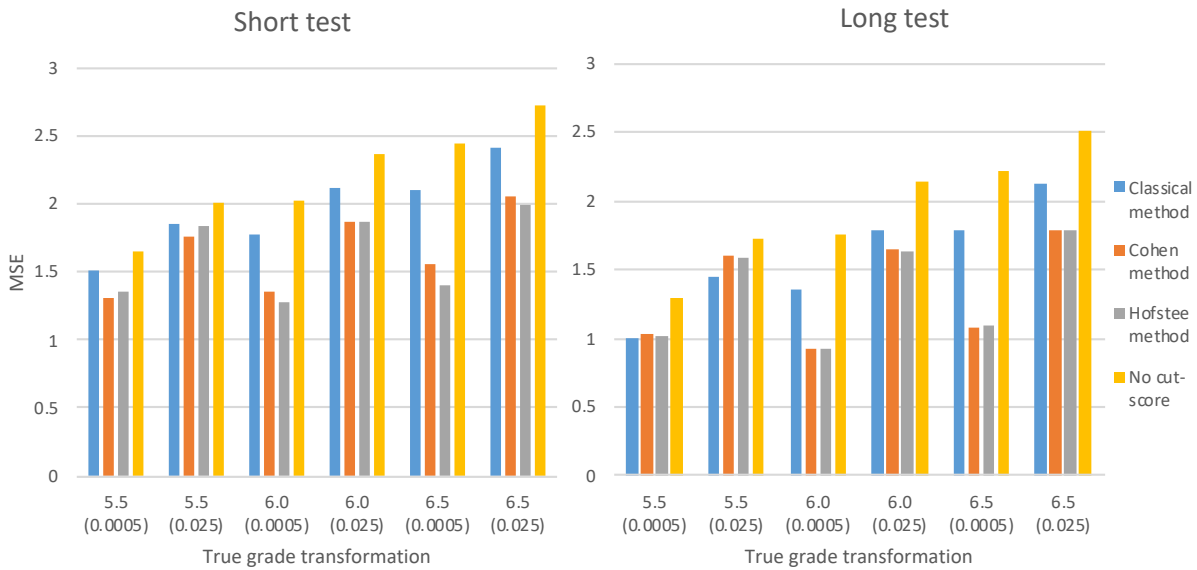


Figure 4. MSE values for a short test of 20 items and a long test of 60 items.

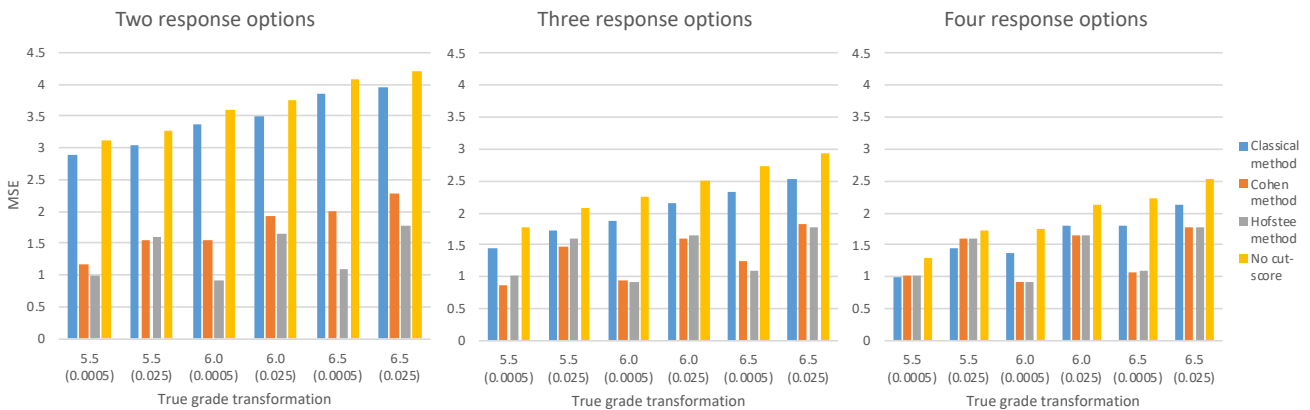


Figure 5. MSE values for varying the number of response options in a 60 item test.

Cohen percentage. Increasing the percentage of best scoring students used as a reference for test difficulty in the Cohen method (from 5% to 10%) only resulted in a slight increase in the MSE for the Cohen method. As a result, it resulted in the (second) largest MSE values for grades below 6. This is because increasing the percentage resulted in a higher average estimated grade, increasing the overestimation of the Cohen method for these windows. Furthermore, the sensitivity, specificity, and positive predictive value for the Cohen method decreased, while the proportion of misclassifications slightly increased for cohorts with an average true grade of 5.5. These results show that using a smaller reference group in the Cohen method leads to more accurate results.

Cut-score percentage. Varying the percentage of items to set the cut-score in the classical and Cohen method also influenced the results. For cohorts with a low average true grade, the Cohen method had the highest MSE when the percentage was lowered. Alternatively, increasing the percentage resulted in the classical method having the highest MSE values regardless of the transformation. When the percentage was lower, the estimated grade of the classical method increased most. Also, for these situations, sensitivity values increased, while the specificity and positive predictive values of the classical method decreased (with the Cohen scoring the lowest positive predictive values). Finally, the Cohen method resulted in the largest proportion of misclassifications for cohorts with an average true grade of 5.5 and 6.0. For a higher cut-score percentage, the average estimated grades decreased (more so for the classical method). Furthermore, the sensitivity values for the classical method became larger (and slightly increased for the Cohen method) as the percentage was lowered. As the percentage increased, the sensitivity values decreased, while the specificity increased (relatively more for the classical method), as did the positive predictive value (more so for the Cohen method). Overall, the proportion of misclassifications decreased for the Cohen method and increased for the classical method, as the percentage was stricter. To conclude, at higher percentages the Cohen method was superior, whereas at lower percentages the classical method performed best.

Discussion

The aim of this study was to compare the accuracy of different cut-score methods that are tenable in a Dutch higher education context in which small-scale non-standardized tests are used that are designed in-house. In this study we compared the performance of an absolute method (i.e., the classical cut-score method) and compromise methods (the Cohen and Hofstee method). These methods can be arranged according to the degree of sample information taken into account: the classical does not, the Cohen only takes information of the best performing students into account, and the Hofstee method specifies both the highest and lowest scoring students. To evaluate the performance of these methods, simulations were performed to obtain students' true and estimated grades, thereby creating realistic higher education contexts by varying sample size, test difficulty, test discrimination, test

length and the number of response options, as well as the specific percentages used in the cut-score methods. Additionally, to prevent information from getting lost, performance was not only evaluated using the pass/ fail classification rates but also by evaluating the accuracy of the estimated grades throughout the entire grading scale.

In general, the results showed that not using a cut-score method in estimating grades did not result in accurate grades as these severely underestimated the students' true grades. Similarly, the classical method did not perform well as it underestimated the true grades in most simulated scenarios (to a larger extent when students' true grades increased). Only for cohorts in which the true ability of students was low, this underestimation was less problematic and the differences with the other methods were smaller. In general, the Cohen and Hofstee method resulted in quite similar estimated grades, that were either an under- or overestimation, depending on the students' true grades. Specifically, for students with low true grades, the Cohen and Hofstee method overestimated the true grades (the overestimation of the Hofstee method being slightly larger), while for students with high true grades, the Cohen and Hofstee method underestimated the true grades (yet to a smaller extent than the classical method). Furthermore, the classification rates showed that the classical method did not perform well at classifying the proportion of students that passed from those that should have passed (i.e., sensitivity), resulting in the highest false negative rate. Contrary, the Hofstee and Cohen method were bad at separating the proportion of students that failed from those that should have failed. In other words, using a more relative approach seems to result in higher false positive rates. This was also evident by the positive predictive values (i.e., all students that should have passed from those that passed) which were higher in the Hofstee and Cohen method compared to the classical method. Overall, the proportion of misclassification was highest in the Hofstee and Cohen method when the average true grade was low (having more misclassifications due to the overestimation of grades), and highest for the classical method when average true grade increased (having more misclassifications due to the underestimation of grades).

Besides these results, varying the sample size, test difficulty, test discrimination, test length, and number of response options in MC items influenced the performance of

the cut-score methods to different degrees. With very small sample sizes, the Cohen method performed worse than the Hofstee method when students' true grades were low, resulting in a higher MSE and more misclassifications. Given the relative nature of the Cohen method, selecting the best performing students in such small cohorts of low performing students, seems to result in an overestimation that results in worse performance than the classical method. An aspect that influenced the results to a larger extent was the test difficulty, as tests became more difficult the Cohen method performed better and the classical method worse. Here, the misclassification proportions were higher for the classical method than the Cohen method, regardless of the true grade in the cohort. For easier tests, the Cohen and Hofstee method performed worse (though differences in performance between the methods became smaller). Similarly, the item discrimination influenced the results. With lower item discrimination Cohen's performance decreased, and in general the differences in classification rates between the methods became larger. On the other hand, when item discrimination increased, the classical method's performance decreased. Finally, results for different test lengths showed that shorter tests resulted in decreased performance of the classical method. Longer tests resulted in decreased performance of the Cohen method when true grades were low. For cohorts with higher true grades, Cohen's proportion of misclassification slightly decreased while that of the classical method increased. More so than the other variables, the number of response options was of influence. As fewer response options were included, performance declined (mostly so for the classical method), showing that tests with only two response options are preferably avoided.

Additionally, varying the percentages to set the cut-score provided information on how to best use the Cohen and classical method. Overall, increasing the percentage of students to use as a reference in the Cohen method decreased its performance. In line with previous studies (e.g., Cohen-Schotanus & Van der Vleuten, 2010; Taylor; 2011), it is therefore advised to use the best five instead of ten percent as a reference group. Furthermore, evaluating the pre-fixed percentage of items to answer correctly in the classical and Cohen method, showed that lowering the percentage results in decreased performance of the Cohen method. Alternatively, increasing the percentage harmed

the performance of the classical method. This shows that when using the classical method, the percentage set should not be too strict, while for the Cohen method the percentage should not be too lenient, as in accordance with Cohen-Schotanus and Van der Vleuten (2010) who used 60%.

Providing recommendations on the use of cut-score methods in higher education tests is complicated by the finding of our sensitivity analysis that showed that the specific theta to true grade transformation mattered in the comparison of the cut-score methods. As theta and consequently true grades are not available in practice, it is difficult to determine the specific situation in practice. For example, for non-standardized tests that vary in difficulty such as those used in higher education, a cohort's ability level is often unknown. Overall, though, the results show the strengths and weaknesses of each method that may be evaluated in light of the goal of higher education testing. For example, it might be preferred to have false positives in an educational setting, assuming that students might fail on a consequent test (or other subsequent academic hurdles), than to have false negatives where a student might not recover from an undeserved fail. From this perspective, one might prefer to apply the Cohen method, sometimes overestimating grades, than having a classical method in which student' true grades are mostly underestimated. As a next step, a decision tree will be developed to provide a detailed and easily accessible overview of the results that were obtained. This may function as an aid in choosing the most appropriate cut-score method in a specific higher education setting. Overall, the sensitivity analysis is important to perform when doing a simulation study in order to assess the tenability of the assumptions and assess how results depend on the researchers' choices. While these assumptions may be considered a disadvantage of using simulations as a research method, it does have the advantage of forcing the researcher to make these choices explicit. Simultaneously, these are easily published and made transparent, thereby increasing the reproducibility of research.

Importantly, as shortly touched upon in the introduction of this study, using a percentage of correctly answered items to set the cut-score as a proxy for students' ability levels is questionable. Ideally, the cut-score should be set based on the performance standard, or student ability level, that is required for a test such that the

pass or fail decision implies that students meet these performance standards or not. This should be done at the level of students' underlying ability (e.g., true grades). By only using a (arbitrary) percentage of test items to answer correctly to set the cut-score, the pass/fail decision risks losing its meaning across different forms of the test (e.g., tests with varying difficulties) as the meaning of passing or failing would vary based on the properties of the test. This is a problematic aspect of setting the cut-score at a specific percentage of test items answered correct that has not received much attention in Dutch higher education and is an important aspect for policy makers and instructors to be aware of.

Currently, our study mainly focused on MC knowledge tests for which we included a guessing parameter in the simulation of students' response patterns. However, the cut-score method that was consequently applied to the response patterns only used students' total scores. In that sense, our results might also apply to open ended items. As the results of varying the type of test showed, increasing the guessing probability by decreasing response options resulted in less accuracy. Consequently, open ended items where guessing might be low, might result in higher accuracy levels overall, not taking into account any effects of subjectivity in assessing open item responses. Furthermore, the Dutch grading scale was used as an example and reference in this simulation study. This is, however, just an example of possible transformations. As long as the assumption of the normal standard distribution of theta and the equal intervals in the grading scale applies, the transformation of theta to true grades might mimic other grading scales where the results still apply. Notably, this may not be the case for situations in which instructors determine grades by strictly using normative information. Given the aim of higher education tests, that is, to test for a specific knowledge or ability level, this might also not be an appropriate method for transforming test scores to grades in higher education. In addition, this study evaluated the Cohen method in which the 5% best performing students were used as a reference for test difficulty. Hereby, it is assumed that his group is stable. For future studies, it would be interesting to assess the accuracy when using other (possibly more stable) groups as reference for test difficulty, such as a group of students between the 25th and 75th percentile.

Overall, the aim of our study was to evaluate the cut-score methods in small-scale non-standardized Dutch higher education tests that are often designed in-house. With this, we aim to make seemingly arbitrary choices more conscious choices. In general, our results show that the classical method mostly underestimates students' ability, while the Cohen method sometimes overestimates students' ability. This shows that, as a whole, taking into account some sample information in terms of the performance of the best scoring students might be beneficial in estimating students' grades.

Appendix A. Detailed Outline Simulation Procedure

In this appendix, the simulation is discussed in detail.

Simulating Student Grades

First, student theta scores were generated using the R function `theta_sim`. A sample of n true grades was generated from a standard normal distribution with a mean m and standard deviation sd using the `rnorm` function:

```
> theta_sim <- function(n, m, sd) {  
+   theta <- rnorm(n, mint, maxt, m, sd) #sampling from normal  
+                                       distribution  
+   return(theta)}  

```

Second, theta scores were transformed to true grades using a linear transformation in which it was assumed that the average true grade corresponded to an average theta score of zero. For this, the R function `TrueGrade` was applied. This function takes the true grades, the average true grade (`mTrGr`), the minimum true grade (`mingrade`), the maximum true grade (`maxgrade`), and the theta score at the minimum true grade of 1.0 that indicates a specific proportion of 1.0 true grades (`mintheta`):

```
> TrueGrade <- function(theta, mingrade, maxgrade, mintheta){  
+   TrGr <- c()  
+   for (i in 1:length(theta)){ #transformation theta to true grade  
+     TrGr[i] <- meanTrGr + ((mTrGr-mingrade)/(0-  
+mintheta)*theta[i])}  
+   TrGr[TrGr<1.0] <- mingrade #round grades  
+   TrGr[TrGr>10.0] <- maxgrade  
+   return(TrGr)}  

```

Next, correct and incorrect scores were simulated for each item based on the students' theta using item response theory (IRT). Specifically, the three-parameter logistic model (3PLM; Birnbaum, 1968) was applied, having a parameter for each item's difficulty, δ_j , the item's discriminability, α_j , and the lower bound probability, π_{0j} :

$$(1) P(X_j = 1 | \hat{\theta}_i, \hat{\delta}_j, \hat{\alpha}_j, \hat{\pi}_{0j}) = \hat{\pi}_{0j} + (1 - \hat{\pi}_{0j}) \frac{\exp(\hat{\alpha}_j(\hat{\theta}_i - \hat{\delta}_j))}{1 + \exp(\hat{\alpha}_j(\hat{\theta}_i - \hat{\delta}_j))}$$

In the function `simcorincor` the item score patterns are simulated using the `ltm` package (Rizopoulos, 2006). This function takes the k number of items in a test as an argument, as well as the n number of students, and the mean and standard deviation

of the normal distributions from which item discrimination, item difficulty, and item guessing parameters were sampled.

```
> simcorincor <- function(k, n, meandif, sddif, meandisc, sddisc,
+                          meanguess, sdguess, Theta){
+   beta0 <- rnorm(k, mean = meandif, sd = sddif) #simulating item
+     difficulty parameters (delta)
+   beta1 <- rnorm(k, mean = meandisc, sd = sddisc) #simulating item
+     discrimination parameters (alpha)
+   beta2 <- rnorm(k, mean = meanguess, sd = sdguess) #simulating
+     item guessing parameters (alpha)
+   thetas <- as.matrix(cbind(beta0,beta1,beta2))
+   corincorsim <- rmvlogis(n, thetas = thetas, IRT = TRUE, link =
+     c("logit"), distr = c("normal"), z.vals = theta)
+   return(corincor)}
```

Cut-Score Methods

After obtaining students' true scores and true knowledge proportions (which are linearly related) as well as students 'observed' item scores, different cut-score methods were applied to the test scores to convert them into grades.

For the classical cut-score method the function `ClassicalMethod` was written. Its arguments are: the `k` number of items, the percentage of items to answer correct after correction for guessing at the cut-score (`caesura_perc`), the grade obtained at the cut-score (`csgrade`), the correct incorrect scores for each student (`corincorsim`) and the minimum (`mingrade`) and maximum grade that can be obtained (`maxgrade`).

```
> ClassicalMethod <- function(k, caesura_perc, csgrade, itemdata,
+                             corincor, mingrade, maxgrade){
+   NrCorrect <- apply(corincor, 1, sum) #total number of correct
+     items for everyone
+   Guess <- (itemdata[1] / 2) + (itemdata[2] / 3) + (itemdata[3]/4)
+     #determine the guessing probability based on the number of items
+     with 2 (in [1]), 3 (in [2]) or 4 (in[3]) answer alternatives
+   Caesura <- (Guess * k) + caesura_perc * (k - (guess * k))
+     #determine cut-score taking into account correction for guessing
+   b1 <- (maxgrade - csgrade) / (k - Caesura) #slope calculating
+     grades from total scores
+   Grade <- round(csgrade + (NrCorrect - Caesura)*b1, 1)
+   Grade[Grade < minTrGr] <- minTrGr #truncate grade
+   result <- list(caesura = Caesura, grade = Grade)
+   return(result)}
```

The Cohen method was applied using the `CohenMethod`. In addition to arguments described before, this function takes the size of the group that is selected as the best performing students (`perc_trim_cohen`) as an additional argument.

```

> CohenMethod <- function(k, perc_trim_cohen, caesura_perc, csgrade,
+   itemdata, corincor, mingrade, maxgrade){
+   NrCorrect <- apply(corincor, 1, sum) #determine the total number
+     of correct items for each individual
+   Knowlprop <- NrCorrect/k
+   Trimknow <- Knowlprop [trunc(rank(Knowlprop))/length(Knowlprop)
+     >= perc_trim_cohen] #rankorder individuals based on
+     their knowledge proportion, determine percentile
+     rank and remove those with rank beneath percentile
+     we want to use for caesura determination.
+   Guess <- (itemdata[1] / 2) + (itemdata[2] / 3) + (itemdata[3]/4)
+     #determine the guessing probability based on the number of items
+     with 2 (in [1]), 3 (in [2]) or 4 (in[3]) answer alternatives
+   NrCorrectTrim <- Trimknow * k #estimate the number of correct
+     Items for the selected individuals
+   Caesura <- (guess * k) + caesura_perc * (mean(NrCorrectTrim) -
+     (guess * k)) #determine cut-score taking into account
+     correction for guessing
+   b1 <- (maxgrade - csgrade) / (k - Caesura) #slope
+   Grade <- round(csgrade + (NrCorrect - Caesura)*b1, 1)
+   Grade[Grade < mingrade] <- mingrade #truncate grades
+   result <- list(caesura= Caesura, grade = Grade
+   return(result)}

```

For the Hofstee method the function `HofsteeMethod` was written. This function takes the following arguments not used before: the minimum acceptable cut-score (K_{min}), the maximum acceptable cut-score (K_{max}), the minimum acceptable failure rate (F_{min}), and the maximum acceptable failure rate (F_{max}).

```

> HofsteeMethod <- function(Kmin, Kmax, Fmax, Fmin, csgrade,
+ corincor,k) {
+   NrCorrect <- apply(corincor, 1, sum)#determine the total number
+     of correct items for each individual
+   Knowlprop <- NrCorrect / k
+   hofsteefunctions <- function(knwldgrange) {
+   cdf <- ecdf(Knowlprop) #create the cumulative density function
+     of the estimated knowledge proportion
+   prop_failed_cdf <- cdf(knwldgrange) #determine the cumulative
+     probability of knowledge proportions tussen Kmin and Kmax. This
+     gives the probability of individuals that score lower than cut
+     off indicated by Kmin and Kmax (who fail the test)
+   b1 = (Fmax - Fmin) / (Kmin - Kmax)
+   b0 = -Kmin * b1 + Fmax
+   prop_failed_lin = b0 + b1 * knwldgrange #caesura function
+   return(matrix(c(prop_failed_lin, prop_failed_cdf), 1, 2))} #this
+     function returns the cut-score proportion and the observed
+     proportion fails for knowledge proportion.

#The following lines of code are used to create a range of knowledge
+ proportions for which we want to determine the cumulative
+ probabilities
+   Kmin100 = Kmin * 100
+   Kmax100 = Kmax * 100
+   knwldgrange <- matrix(seq(from=Kmin100, to=Kmax100, by=.1),,

```

```

1)/100
+ funct <- cbind(knwdlgrange, t(apply(knwdlgrange, 1,
      hofsteefunctions))) #create a matrix with the
range of cut-scores between Kmin and Kmax, the corresponding y-
coordinates on the caesura-line and the y-coordinates (the
cumulative probabilities) based on the cdf of the knowledge
proportions
+ if (length(funcnt [funcnt [,3] <= Fmax & funcnt [,3] >= Fmin])== 0)
  {restricfuncnt <- matrix(NA, nrow = dim(funcnt)[1], ncol = 3)}
+ else {
+   restricfuncnt <- matrix(funcnt[funcnt[, 3] <= Fmax &
      funcnt[, 3] >= Fmin],, 3)} #create a matrix
for all observations that fall between minimum and maximum
acceptable fail rates (Fmin and Fmax) with range of cut-scores
between Kmin and Kmax, the corresponding y-coordinates on the
caesura-line and the y-coordinates (the cumulative
probabilities) based on the cdf of the knowledge proportions
+ restricfuncnt_dif <- cbind(restricfuncnt,
abs(restricfuncnt[,2] - restricfuncnt[, 3])) #calculate
difference between y-coordinate on caesura line and
corresponding y-coordinate based on cdf of knowledge proportions
+ colnames(restricfuncnt_dif) <- c("knowledgeprop",
  "prop_failed_lin", "prop_failed_cdf", "dif")
#create two separate dataframes with 1) the range of
knowledge proportions between Kmin and Kmax and 2) The
proportion of individuals that fail based on either the caesura
line or the estimated cdf.
+ prop_failed_lin <- data.frame(x = restricfuncnt_dif[, 1], y =
  restricfuncnt_dif[, 2])
+ prop_failed_cdf <- data.frame(x = restricfuncnt_dif[, 1], y =
  restricfuncnt_dif[, 3]) #select rows
belonging to the knowledge proportion for which the difference
between the caesura line and the estimated cdf are minimal
+ min_dif <- matrix(c(restricfuncnt_dif[restricfuncnt_dif[,4] ==
  min(restricfuncnt_dif[, 4])],, 4) #selects knowledge
levels for which difference is minimal.
#If more than one row has the same minimal difference between the
caesura line and the estimated cdf, the average knowledge proportion
of these rows is determined
+ caesur a_perc <- mean(min_dif[, 1])
+ b1 <- (maxgrade - csgrade) / (mingrade - caesura_perc)
+ Grade <- round(csgrade + b1 * (Knowlprop - caesura_perc), 1)
+ Grade [Grade < mingrade] <- mingrade
+ Grade [Grade > maxgrade] <- maxgrade
+ result <- list(grade = Grade, caesura = caesura_perc)
+ return(result)}

```

Finally, the no cut-score method was applied using the `NoCaesuraMethod` function. This function uses similar arguments as mentioned before.

```
> NoCaesuraMethod <- function(k, corincor, itemdata, maxgrade,
                             mingrade,) {
+   NrCorrect <- apply(corincor, 1, sum) #determine the true
                             number of correct
+   items for each individual
+   Guess <- (itemdata[1] / 2) + (itemdata[2] / 3) + (itemdata[3] /
+ 4)
+   #determine the guessing probability based on the number of items
+   with 2 (in [1]), 3 (in [2]) or 4 (in[3]) answer alternatives
+   b1 <- (maxgrade - mingrade) / (k) #slope
+   Grade <- round((NrCorrect - (Guess*k)) * b1, 1) #grade
+   Grade[Grade < mingrade] <- mingrade #truncate grades
+   result <- list(caesura= Caesura, grade = Grade)
+   return(result)}
```

Note that these methods are now described such that `corincor` scores are used as input. However, to compare outcomes to true scores these methods can also be adjusted, such that they use the true knowledge proportion times the `k` items as the total score as input for the cut-score methods.

6

General Discussion

In higher education curricula, students' performance is continuously evaluated by administering tests. With these tests, students' performance is estimated, based on which different decisions are made. On the level of the curriculum, tests are combined to inform decisions to determine whether students are allowed to continue their studies or whether students meet the requirements to receive their diploma. Additionally, on the level of individual courses, students' performance can be evaluated using individual tests for which decisions are made such as whether students meet the requirements to pass the test. The aim of this dissertation was to evaluate the decisions made in higher education about students' performance, both on the curriculum level in which multiple tests are combined in Chapter 2 and 3, and on the level of individual tests in Chapter 4 and 5. To preserve the educational quality of a study program's diploma, such that students who receive the diploma meet the requirements set by the institution, these decisions on students' performance should be valid.

General Discussion

As the example in the introduction of this dissertation illustrates, different stakeholders in higher education curricula make different decisions about students' performance. Whereas, as described, objectives may differ depending on one's perspective, both course instructors like Mark and policy makers like Carol wish to make accurate decisions about students' performance. As available resources, such as time and budget, are limited in higher education curricula, the quality of tests in higher education may be limited, which consequently may result in inaccurate decisions. In Chapter 1 and Chapter 2 the accuracy and consequences of allowing compensation when combining multiple tests was evaluated for decisions such as the binding study advice (BSA; that is, the Dutch academic dismissal policy) decision made by Carol. As both studies show, some of the motivations to implement course compensation may be questioned. For example, the results of Chapter 2 showed that using the average grade to make decisions does not necessarily result in more accurate decisions than a traditional testing system in which course credits are assigned to individual courses. Instead, the accuracy of compensatory decision rules relative to conjunctive rules depends on the degree of compensation allowed (i.e., the

specific GPA and minimum grade that is required), as well as the context in which the decision is made, in terms of the average test reliability, average test correlation, and the number of reexaminations.

Furthermore, the results of Chapter 3 showed that in curricula in which course compensation is allowed, three groups of students may be identified for which the relation between a first-year precursor course and second-year sequel course is positive. Most relevant for this dissertation, one of these classes is characterized by an overall low performance in the first-year on average (that is, the Psychology students had a low first-year average and a high number of compensations and retakes on average). For this group, the average grade on the precursor course was below the required average grade, while the average grade on the sequel course was insufficient considering the Dutch grading scale (i.e., a cut-score of 5.5 on a 1-10-point scale). These results show that when performance on a precursor course is low, performance on a sequel course is low as well and suggest that knowledge accumulation for this group of students might not be sufficient when a precursor course could be compensated. At the same time, the precursor course was also compensated (i.e., students performed low) by students whose overall first-year performance was moderate to high and in these groups of students the grade on the sequel course was on average sufficient. This seems to suggest that some of the knowledge and skills required to score well on a sequel course could be accumulated in other courses. In this sense, course compensation might be undesirable in curricula where the content of sequel courses, where performance may be low for a group of students whose overall first-year performance is low, is critical in the end qualifications of the curriculum.

Given the discussion on whether to allow course compensation or not in a higher education curriculum, the results of Chapter 2 include a direct comparison of the accuracy of compensatory and conjunctive decision rules. The results of Chapter 3, however, only apply to students in a specific compensatory testing system. Whether the observed patterns generalize to students in a conjunctive testing system is unclear. However, if the learning processes would be similar across testing systems, the results might indicate that the group of students for whom performance on the precursor and

sequel course is both low would be larger in a testing system in which the minimum required grade on individual courses is lower (as is generally true for compensatory testing systems). Overall, the discussion on whether to allow for compensation and to what extent, mostly seems to be a discussion of favoring false negative over false positive misclassifications or the other way around. As Albers, Vermue, de Wolff, and Beldhuis (2018) conclude their study on the BSA decision and the requirements set within this decision, deciding on acceptable false positive and false negative rates is the primary role of policy makers in higher educational institutions. The results of Chapter 2 thereby provide a guideline in showing how different decision rules result in different type of misclassifications in what situations. Importantly, however, policy makers should design a testing system that fits the nature of their decision and the end qualification norms of a study program. That is, if compensation is allowed, this should be in line with the end qualification norms. For example, Psychology students would still be trained psychologist when they receive their bachelor degree, even though they were allowed to compensate courses in their first year. Furthermore, as Smits, Kelderman, and Hoeksma (2015) argue, course compensation should not be implemented to correct for imprecise individual measurements or, as the results in Chapter 2 show, should not be implemented because it is assumed that the average is more reliable. As described in the introduction of this dissertation, the context of the decision should thereby be considered. For example, for students to receive their master degree in which a high level of expertise is considered, each course should be passed to ensure they are experts as is prescribed by the degree. For receiving a bachelor diploma however, a student might still meet the end qualification requirements when a first-year course was compensated.

Overall, the results from Chapter 2 show the importance of the quality of the individual tests for the accuracy of a decision based on the combination of tests. However, as mentioned, the quality of instructor-made multiple choice (MC) tests in higher education is often low. Focus in Chapter 4 and 5 is specifically on MC items, as this is a commonly used item format in higher education because of the possibility to efficiently assess a broad range of material and to easily score the items (Brown & Abdalnabi, 2017). Furthermore, choices made in transforming test scores to grades

often seem arbitrary and in Chapter 4 and Chapter 5 the choices for methods to correct for guessing and to determine the cut-score were evaluated to see if these could be improved. Whereas the results from Chapter 5 showed that the most accurate cut-score method varied across contexts, the results from Chapter 4 showed that in general the extended classical correction for guessing method was more accurate than the classical method that is often applied in Dutch higher education. Consequently, students' ability levels could be estimated more accurately using test scores in most higher education tests and situations, by using some sample information in the correction for guessing and in setting the cut-score. When information on the sample is included in the correction for guessing or cut-score method, a more relative instead of absolute approach is taken. This might seem to contradict with the aim of higher education tests which is to measure students' knowledge and skills on a specific course, instead of ranking students based on their relative performance. In higher education testing, however, many aspects vary throughout courses, such as the ability levels of students or the difficulty of tests. To account for these differences to some extent, some sample information could result in more accurate decisions. Consequently, the accuracy of the decision should also be taken into account in designing higher education policies. To ensure the method does not hinge on the relative approach too much, conditions could be introduced such that taking sample information into account is only allowed in certain situations (e.g., the minimum required sample size).

Research in this dissertation underlines the importance of the way in which decisions on students' performance are made in higher education curricula. Instead of using arbitrary cut-scores or classical methods for the sake of tradition, the research presented shows that these decisions should be considered carefully and preferably substantiated by (scientific) arguments. Overall, the large number of decisions on students' performance made in higher education, and the differences between tests in higher education and those often studied in the educational measurement literature, show the relevance of studying tests in higher education within this field as well.

Scientific Contributions

The aim of this dissertation was to evaluate decisions on students' performance in higher education. To assess the accuracy of these decisions, simulations were performed. Whereas simulations are a common research method in the field of psychometrics, it is yet quite unknown in the field of educational sciences. When, however, we wish to assess the accuracy of decisions in the context of higher education, simulations are required to obtain true scores, as we need to compare the decision based on students' true scores with the decision based on observed test scores to get information on the accuracy of the decision. The advantage of performing simulations is that one can model different contexts and easily perform sensitivity analyses by varying these aspects of these contexts across simulations.

Simultaneously, researchers who perform simulations have to make their underlying data production processes explicit and can easily share this code, increasing the transparency of their research. Contrary, when using empirical data one is limited to one specific context. This aspect is especially difficult when studying policy changes in higher education as many factors influence students' performance in such a specific context. Here, cross validation and increasing focus on predicting future data instead of explaining the dataset at hand (see e.g., Yarkoni & Westfall, 2017) might improve research in this field.

On the other hand, a limitation of using simulations is that the model used to set up simulations is only an abstract representation of reality and might consequently not capture the whole truth. For example, assumptions about the distribution of students' grades, students' guessing behavior, or their choices of what courses to retake are made, which may not perfectly reflect reality. Still, by performing simulations a researcher makes its data production model explicit and can make its code easily available, increasing the transparency of research. If one does not agree with the model, one can easily adapt it and replicate the research using an adapted model. Overall, comparing a conjunctive and compensatory testing system empirically is difficult due to the many possible influences on student behavior and the lack of possibilities for randomized controlled trials in higher education research. In this

light, simulations allow for a good stepping stone for future (quasi)experimental studies.

Directions for Future Studies and Practice

In this dissertation, the accuracy of the decisions on students' performance that are made in higher education was the main focus. However, as was also touched upon, additional motivations underlie the implementation of educational policy that were not addressed. For example, a compensatory decision rule was implemented at the Erasmus University Rotterdam to, among other motivations, direct study behavior such that students' procrastination behavior would be reduced. As students' study behavior was targeted by the new educational policy, not only student success should be evaluated in evaluating the effectiveness of the new policy but changes or differences in student (procrastination) behavior should also be assessed (Boevé et al., 2017). For this purpose, new developments in data collection such as the use of mobile-devices for conducting diary studies (known as ecological momentary assessment or experience-based sampling) might be useful to assess students' study behavior by means of their study time allocation.

Furthermore, in this dissertation simulations studies were performed in which students' reactive behavior was not incorporated. Instead of the view of a passive and naïve student who studies at the best of his or her abilities regardless of the testing system that is employed, students could be viewed as decision makers who strategically prepare for a test. Van Naerssen (1970) was one of the first to develop an economic decision-making model of examinations in which he describes the student as an agent that wishes to optimize his or her learning process such that the total effort for the student is minimized. This model was further developed by Wilbrink (1995). Although it is not clear (neither given) whether decision accuracy would vary when student behavior is modelled, future studies might evaluate this by extending the simulation model by incorporating student behavior. Budescu and Bar-Hillel (1993) and Budescu and Bo (2015) designed models for test taking behavior in which decision theory is combined with psychometric theory. Together, these approaches might serve as a good starting point for setting up a model of student strategic behavior at the level of the curriculum in which multiple tests are combined.

In addition, it is important for educational institutions to consider the evaluation of newly implemented educational policy or changes therein. In Chapter 3, data was obtained from the Erasmus Educational Research (EER) database, in which students' test scores throughout different study programs and schools are collected. Overall, the collection of data across different schools for this database was challenged by differences in reporting students' performance and use of definitions (aspects also experienced by Nakabo-Ssewanyana, 1999). To overcome such problems, the evaluation of new educational policy (i.e., its effectiveness) should be implemented or planned simultaneously. For this purpose, it is important to collect data that is informative and consistent and plan its collection in advance. Adopting a university-wide policy for the reporting and collecting of students' performance would not only improve comparisons across study programs for management purposes but for the scientific evaluation of educational policy as well. As empirical research and comparative research in higher education is complicated by the many influencing factors, additional challenges such as inconsistent data should be resolved when possible.

Finally, this dissertation showed that the quality of the decisions to be taken are highly dependent on the quality of the tests. It would therefore be useful to invest in improving this quality. Apart from actions institutions could take themselves, such as, for example, training of staff, the quality of tests in higher education could be improved by combining forces across higher education institutions as well. For the most popular bachelor programs, offered at multiple (international) universities, it would be valuable for instructors to come together to design items collectively. In this way, coordinators of similar courses could collectively build a test item bank in which data on students' performance could be collected and saved. By combining forces, instructors might have more time available to increase the quality of their test items and hence tests. Ultimately, nation-wide tests might even be constructed for specific study programs to safeguard the quality of these tests nation-wide.

S

Summary

S

In higher education curricula, students' performance is continuously evaluated by administering tests. With these tests, students' performance is estimated, based on which different decisions are made. On the level of the curriculum, tests are combined to inform decisions to determine whether students are allowed to continue their studies or whether students meet the requirements to receive their diploma. Additionally, on the level of individual courses, students' performance can be evaluated using individual tests for which decisions are made such as whether students meet the requirements to pass the test. The aim of this dissertation is to evaluate the decisions made in higher education about students' performance, both on the curriculum level in which multiple tests are combined (Chapter 2 and 3) and on the level of individual tests (Chapter 4 and 5). To preserve the educational quality of a study program's diploma, such that students who receive the diploma meet the requirements set by the institution, these decisions on students' performance should be valid.

To evaluate the accuracy of the decision made in the academic dismissal (AD) policy (known as the binding study advice, BSA, in Dutch higher education) at the end of the first year of the bachelor, in which the decision whether a student is allowed to continue their studies is made, a simulation study was performed in **Chapter 2**. By performing real-data-guided simulations the accuracy of this BSA decision using different complex compensatory and conjunctive decision rules was evaluated. Additionally, simulations were performed to mimic several realistic higher education contexts. Overall, the results show that the accuracy depends on the degree of compensation allowed; on the required average, the minimum grade, as well as their combination. In general, within compensatory decision rules the false negative rate (i.e. those students who truly meet the requirements yet were not allowed to continue their studies) was lower and the false positive rate (i.e., students who do not truly meet the requirements yet are allowed to continue their studies) higher compared to conjunctive decision rules. Furthermore, the results showed that which rule is more accurate also depends on the average test reliability, the average correlation between tests, and the number of retakes. Together, these results show that the reason for allowing course compensation in higher education, namely that the average grade is

more reliable, does not generally hold in all situations but that the accuracy of the complex compensatory decision rule depends on its context.

In **Chapter 3** the consequences of allowing compensation in the first-year BSA decision on performance on a second-year sequel course was evaluated using data from a Psychology bachelor and a Law undergraduate program. In particular, sequel courses that build on material from precursor courses were evaluated to assess possible consequences of allowing course compensation in knowledge accumulation. Extending on previous research, students' performance on sequel courses was evaluated for different groups of students by applying a latent class regression. Student groups were distinguished who portrayed different unobserved study processes by forming the latent classes based on similar patterns in first-year averages, variability in first-year grades, and similar average number of compensated and retaken tests. Across the two study programs, three classes of students were identified. The results showed that average performance on the first-year precursor course was under the required average grade (<6.0 on the Dutch 1-10 grading scale) while the second-year sequel course was on average a failing grade (<5.5) for students who were in the lowest performing class. This seems to suggest that compensating a precursor course might on average have negative consequences on the knowledge accumulation for students in the class with overall low first-year performance, such that performance on later courses is not sufficient. However, the results also show that the precursor course is compensated by students in each of the three classes, yet performance on the sequel course for students in the other two classes is on average not insufficient. This seems to suggest that students with higher first-year performance might not experience negative consequences in knowledge accumulation when they compensate a precursor course.

The evaluation of the BSA decision in higher education in which multiple tests are combined showed the importance of the quality of individual tests. Consequently, a shift in focus was made to decisions about students' performance and students' true score estimation on individual tests in higher education. In **Chapter 4** the accuracy of different methods to correct for guessing in estimating true scores in higher education were evaluated. Specifically, the focus was on multiple choice (MC) tests in which

incorrect answers are not directly penalized and students' optimal and most common strategy therefore is to guess (as is also common in Dutch higher education). Classically, a correction for guessing is made using formula scoring. Alternative correction methods, such as the extended classical method, (extended) beta binomial models, and models from item response theory, incorporate sample characteristics. Performing simulations, the accuracy of the estimated true knowledge of students was evaluated for the different correction methods in different realistic higher education settings. Overall, the results showed that the estimation of true scores in MC tests might be improved for most contexts in Dutch higher education, by using the extended classical correction method proposed by Calandra (1941) and Hamilton (1950) or by using a method, such as our proposed weighted item difficulty correction, that incorporates item characteristics in the true score estimation.

Finally, in **Chapter 5** the decision of assigning grades to students' test scores as well as the decision to give students a pass or fail in Dutch higher education study programs were evaluated. The accuracy of three standard setting methods (the classical absolute method, and the Cohen and Hofstee compromise methods; Cohen-Schotanus & van der Vleuten, 2010; Hofstee, 1983) that are tenable in small-scaled, non-standardized tests were assessed. Again, simulations were performed to obtain students true and estimated grades and to evaluate realistic higher education contexts. Overall, the results showed that the classical absolute method underestimates students' true ability in almost all simulated situations, while the Cohen and Hofstee methods overestimate ability in only some situations. Taken together, therefore, it might generally be beneficial to take into account some sample information in terms of the best scoring students.

R

References

R

- Albers, C., Vermue, C., Wolff, T., de, Beldhuis, H. (2018). Model-based academic dismissal policies: A case-study from the Netherlands. doi: 10.31234/osf.io/6a9cz
- Arnold, I. J. M. (2011). Compensatorische toetsing en kwaliteit [Compensatory testing and quality]. *Tijdschrift voor Hoger Onderwijs*, 29, 31-40.
- Arnold, I.J.M. & Van den Brink, W. (2009). De invloed van compensatie op studie-uitval en doorstroom [The influence of compensation on study dropout and progress]. *Tijdschrift voor Hoger Onderwijs & Management*, 16(3), 11-15.
- Angoff, W. H., & Schrader, W. B. (1984). A study of hypotheses basic to the use of rights and formula scores. *Journal of Educational Measurement*, 21, 1-17. doi: 10.1111/j.1745-3984.1984.tb00217.x
- Bakker, M. (2012, 30 January). Vijven, en toch een UvA-diploma [Insufficient grades, and still a diploma]. *De Volkskrant*.
- Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society*, 4, 3-12. doi: 10.1007/s11299-005-0001-z
- Barefoot, B. O. (2004). Higher education's revolving door: Confronting the problem of student drop out in US colleges and universities. *Open Learning*, 19, 9-18. doi: 10.1080/0268051042000177818
- Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement*, 21, 147-152. doi: 10.1111/j.1745-3984.1984.tb00226.x
- Billingsley, P. (1961). The Lindeberg-Lévy theorem for martingales. *Proceedings of the American Mathematical Society*, 12, 788-792. doi: 10.2307/2034876
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 379-479). Reading, MA: Addison-Wesley.
- Black, P., & Wiliam, D. (2003). 'In praise of educational research': formative assessment. *British Educational Research Journal*, 29(5), 623-637. doi: 1.1080/0141192320133721
- Blömeke, S., & Gustafsson, J. E. (Eds). (2017). *Standard setting in education: The Nordic countries in an international perspective*. New York, NY: Springer. doi: 10.1007/978-3-319-50856-6_1

- Boer, H. de., Jongbloed, B., Benneworth, P., Cremonini, L., Kolster, R., ..., & Vossensteyn, H. (2015). Performance-based funding and performance agreements in fourteen higher education systems (Report for the Ministry of Education, Culture, and Science). Enschede, Netherlands: Center for Higher Education Policy Studies.
- Boevé, A. J., Meijer, R. R., Bosker, R. J., Vugteveen, J., Hoekstra, R., & Albers, C. J. (2017). Implementing the flipped classroom: An exploration of study behavior and student performance. *Higher Education*, 74, 1015-1032. doi: 10.1007/s10734-016-0104-y
- Brown, G. T. L., & Abdalnabi, H. H. A. (2017). Evaluating the quality of higher education instructor-constructed multiple choice tests: Impact on student grades. *Frontiers in Education*, 2, 1-12. doi: 10.3389/feduc.2017.00024
- Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement*, 30, 277-291. doi: 10.1111/j.1745-3984.1993.tb00427.x
- Budescu, D. V., & Bo, Y. (2015). Analyzing test-taking behavior: Decision theory meets psychometric theory. *Psychometrika*, 80, 1105-1122. doi: 10.1007/s11336-014-9425-x
- Calandra, A. (1941). Scoring problems and probability considerations. *Psychometrika*, 6, 1-9. doi: 10.1007/bf02288568
- Carlin, J. B., & Rubin, D. B. (1991). Summarizing multiple choice tests using three informative statistics. *Psychological Bulletin*, 110, 338-349. doi: 10.1037/0033-2909.112.338
- Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice*, 22, 32-41. doi:10.1111/j.1745-3992.2003.tb00126.x
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Clauser, J. C., Clauser, B. E., & Hambleton, R. K. (2013). Increasing the validity of Angoff standards through analysis of judge-level internal consistency. *Applied Measurement in Education*, 27, 19-30. doi: 10.1080/08957347.2013.853071
- Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement*, 46, 390-407. doi: 10.1111/j.1745-3984.209.00089.x

- Cohen-Schotanus, J. (1995). De praktijk van de compensation [Practice of compensation]. *Onderzoek van Onderwijs*, 24, 60-62.
- Cohen-Schotanus, J., & Van der Vleuten, C. P. (2010). A standard setting method with the best performing students as point of reference: Practical and affordable. *Medical Teacher*, 32, 154-160. doi: 10.3109/01421590903196979
- Cornelisz, I., Levels, M., Van der Velden, R., De Wolf, I., & Van Klaveren, C. (2018). The consequences of academic dismissal for academic success. ACLA working paper series obtained from: <http://www.acla.amsterdam/workingpapers-wp20181>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(8), 297-334. doi: 10.1007/BF02310555
- De Gruijter, D. N. M. (2008). Al dan geen compensatie in de propedeuse [Compensation in the first year of the bachelor].
- De Koning, B. B., Loyens, S. M. M., Rikers, R. M. J. P., Smeets, G., & van der Molen, H. T. (2014). Impact of binding study advice on study behavior and pre-university education qualification factors in a problem-based psychology bachelor program. *Studies in Higher Education*, 39, 835-847. doi: 10.1080/03075079.2012.754857
- Diamond, J. & Evans, W. (1973). The correction for guessing. *Review of Educational Research*, 43, 181-191. doi: 1.3102/00346543043002181
- DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, 2. doi: 10.5206/cjsotl-rcacae.2011.2.4
- Dochy, F., Kyndt, E., Baeten, M., Pottier, S., & Veestraeten, M. (2009). The effects of different standard setting methods and the composition of borderline groups: A study within a law curriculum. *Studies in Educational Evaluation*, 35, 174-182. doi: 10.1016/j.stueduc.2009.12.006
- Douglas, K. M. (2007). *A general method for estimating the classification reliability of complex decisions based on configural combinations of multiple assessment scores* (Unpublished doctoral dissertation). University of Maryland, Maryland.
- Douglas, K. M., & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics*, 35, 280-306. doi:10.3102/1076998609346969
- Draaijer, S. (2016). Supporting teachers in higher education in designing test items. [Amsterdam]: Vrije Universiteit Amsterdam.

- Espinosa, M. P., & Gardezabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, *54*, 415-425. doi: 10.1016/j.jmp.2010.06.001
- Guo, H. (2017). Predicting rights-only score distributions from data collected under formula score instructions. *Psychometrika*, *82*, 1-16. doi: 10.1007/s11336-016-9550-9
- Haladyna, T., & Hess, R. (1999). An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. *Educational Assessment*, *6*, 129-153. doi: 10.1207/S15326977EA0602_03
- Hamaker, E. L. (2012). Why researchers should think “within-person”: A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.). *Handbook of Research Methods for Studying Daily Life*, 43-61, New York, NY: Guilford.
- Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (pp 433-470). Westport, CT: Praeger.
- Hambleton, R. K., & Slater, S. C. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education*, *10*, 19-28. doi: 10.1207/s15324818ame1001_2
- Hamilton, C. H. (1950). Bias and error in multiple choice tests. *Psychometrika*, *15*, 151-168. doi: 1.1007/BF02289198
- He, Q. (2009). *Estimating the reliability of composite scores*. Coventry, UK: Ofqual. Retrieved from dera.ioe.ac.uk/1060/1/2010-02-01-composite-reliability.pdf.
- Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson, & J. S. Helmick (Eds.). *On educational testing* (pp. 109-127). San Fransisco: Jossey-Bass.
- Impara, J. C., & Plake, B. S. (1998). Teacher’s ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, *35*, 69-81. doi: 10.1111/j.1745-3984.1998.tb00528.x
- Kane, M. T. (2017). Using Empirical Results to Validate Performance Standards. In *Standard setting in Education: The Nordic countries in an international perspective* (pp. 11-29). Springer International Publishing. doi: 10.1007/978-3-319-50856-6_2
- Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika*, *27*, 59-72. doi: 1.1007/BF02289665
- Kendall, M. G., & Stuart, A. (1969). *The advanced theory of statistics* (3rd ed.). New York: Hafner

- Kickert, R., Stegers-Jager, K. M., Meeuwisse, M., Prinzie, P., Arends, L. R. (2017). The role of the assessment policy in the relation between learning and performance. *Medical Education*. doi: 10.1111/medu.13487
- Lee, W. C. (2010). Classification consistency and accuracy of complex assessments using item response theory. *Journal of Educational Measurement*, 47, 1-17. doi:10.1177/014662102237797
- Lee, W. C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, 26, 412-432. doi:10.1177/014662102237797
- Lord, F. M. (1959). An approach to mental test theory. *Psychometrika*, 24, 283-302. doi: 1.1007/BF02289812
- Lord, F. M. (1962). Cutting scores and errors of measurement. *Psychometrika*, 27, 19-30. doi:10.1002/j.2333-8504.1961.tb00102.x
- Lord, F. M. (1963). Formula scoring and validity. *Educational and Psychological Measurement*, 13, 663-672. doi: 1.1002/j.2333-8504.1962.tb0029.x
- Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, 30, 239-27. doi: 1.1002/j.2333-8504.1964.tb0096.x
- Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement*, 12, 7-11. doi: 1.1111/j.1745-3984.1975.tb01003.x
- Lord, F. M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley Publishing Company.
- McBee, M. T., Peters, S. J., & Waterman, C. (2014). Combining scores in multiple-criteria assessment systems: The impact of combination rule. *Gifted Child Quarterly*, 58, 69-89. doi:10.1177/0016986213513794
- McCall, W. A. (1920). A new kind of school examination. *The Journal of Educational Research*, 1, 33-46. doi: 1.1080/00220671.192.10879021
- McCutcheon, A. L. (2002). *Basis concepts and procedures in single- and multiple-group latent class analysis*. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 56- 85). Cambridge, UK: Cambridge University Press.
- Mehrens, W. A. (1990). Combining evaluation data from multiple scores. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 322-334). Newbury Park, CA: Sage. doi: 10.4135/9781412986250

- Morrison, D. G. & Brockway, G. (1979). A modified beta binomial model with applications to multiple choice and taste tests. *Psychometrika*, *44*, 427-442. doi: 1.1007/BF02296206
- Nakabo-Ssewanyana, S. (1999). Statistical data: The underestimated tool for higher education management. *Higher Education*, *37*, 259-279. doi: 10.1023/A:1003664331975
- Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education*, *37*, 464-469. doi: 10.1046/j.1365-2923.2003.01495.x
- Nuffic (2009). Grading systems in the Netherlands, the United States, and the United Kingdom. Den Haag: Nuffic. Retrieved from: <https://www.nuffic.nl/en/library/grading-systems-in-the-netherlands-the-united-states-and-the-united-kingdom.pdf>
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535-569. doi: 10.1080/10705510701575396
- R Core Team (2015, 2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research.
- Reckase, M. D. (2006). A conceptual framework for a psychometric theory of standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*, *25*, 4-18.
- Rekveld, I.J., & Starren, J. (1994). Een examenregeling zonder compensatie in het Nederlandse Hoger Onderwijs? Een vergelijking tussen compensatie en conjunctie [An examination rule without compensation in Dutch higher education? A comparison between compensation and conjunction]. *Tijdschrift voor Hoger onderwijs*, *12*, 210-219.
- Rijksoverheid [Dutch government]. (2018, 3 september). Van Engelshoven stelt paal en perk aan bindend studieadvies.
- Rizopoulos, D. (2006). Ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*, 1-25. doi: 1.18637/jss.v017.i05

- Roediger, H. L. III, & Karpicke, J. D. (2006). The power of testing memory. Basic research and implications for educational practice. *Perspectives on psychological science*, 1, 181-210. doi: 10.1111/j.1745-6916.2006.00012.x
- Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation*, 10(13), 1-4. <http://pareonline.net/pdf/v10n13.pdf>.
- Sadler, R. (2014). The futility of attempting to codify academic achievement standards. *Higher Education*, 67, 273-288. doi: 10.1007/s10734-013-9649-1
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74, 107-120. doi:10.1007/s11336-008-9101-0
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.
- Smits, N., Kelderman, H., & Hoeksma, J. B. (2015). Een vergelijking van compensatoir en conjunctief toetsen in het hoger onderwijs [A comparison of compensatory and conjunctive testing in higher education]. *Pedagogische Studiën*, 94, 275-285.
- Sneyers, E., & De Witte, K. (2018). Interventions in higher education and their effect on student success: A meta-analysis. *Educational Review*, 70, 208-228. doi: 10.1080/00131911.2017.1300874
- Taylor, C. A. (2011). Development of a modified Cohen method of standard setting. *Medical Teacher*, 33, e678-e682. doi: 10.3109/0142159X.2011.611192
- Van de Watering, G., & Van der Rijt, J. (2006). Teacher's and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review*, 1, 133-147. doi: 10.1016/j.edurev.2006.05.001
- Van Naerssen, R. F. (1970). *Over optimaal studeren en tentamens combineren* [oratie Universiteit van Amsterdam]. Amsterdam: Swets & Zeitlinger
- Van Rijn, P. W., Béguin, A. A., & Verstralen, H. H. F. M. (2009). Zakken of slagen? De nauwkeurigheid van examenuitslagen in het voortgezet onderwijs. *Pedagogische Studiën*, 86, 185-195.
- Van Rijn, P. W., Béguin, A. A., & Verstralen, H. H. F. M. (2012). Educational measurement issues and implications of high stakes decision making in final examinations in secondary education in the Netherlands. *Assessment in Education: Principles, Policy & Practice*, 19, 117-136. doi:10.1080/0969594X.2011.591289

- Vermeulen, L., Scheepers, A., Adriaans, M., Arends, L., Van den Bos, R. Bouwmeester, S., Van der Meer, F-B, Schaap, L., Smeets, G., Van der Molen, H., & Schmidt, H. (2012). Nominaal studeren in het eerste jaar. *Tijdschrift voor Hoger Onderwijs*, 30, 204 – 216.
- Vermunt, J. K. & Magidson, J. (2002). *Latent Class Cluster Analysis*. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89-106). Cambridge, UK: Cambridge University Press.
- Vermunt, J. K., & Magidson J. (2013a). *Technical guide for Latent GOLD 5.0: Basic, advanced and syntax*. Belmont, MA: Statistical Innovations Inc.
- Vermunt, J. K., & Magidson J. (2013b). *Latent GOLD 5.0 upgrade manual*. Belmont, MA: Statistical Innovations Inc.
- Vermunt, J. K., & Magidson J. (2015). *LG-Syntax users' guide: Manual for Latent GOLD 5.0 syntax module*. Belmont, MA: Statistical Innovations.
- Vossensteyn, J. J., Kottmann, A., Jongbloed, B. W. A., Kaiser, F., Cremonini, L., Stensaker, B., ... Wollscheid, S. (2015). *Dropout and completion in higher education in Europe: Main report*. European Union. doi: 10.2766/826962
- Wedel, M., & DeSarbo, W. A. (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (p. 352-388). Cambridge, MA: Blackwell.
- Wheadon, C., & Stockford, I. (2013). Estimation of composite score classification accuracy using compound probability distributions. *Psychological Test and Assessment Modeling*, 55, 162-180.
- Wilbrink, B. (1995). Studiestrategieën die voor studenten en docenten optimaal zijn: Het sturen van investeringen in de studie [Study strategies that are optimal for students and instructors: Guiding study investments]. In H. P. M. Creemers (ed.), *Onderwijsonderzoek in Nederland en Vlaanderen 1995: Proceedings van de Onderwijs Research Dagen 1995 te Groningen* (pp. 218-220). Groningen: Gronings Instituut voor Onderzoek van Onderwijs, Opvoeding en Ontwikkeling, Rijksuniversiteit Groningen
- Wilhelm, S., & Manjunath, B. G. (2014). Tmvtnorm: Truncated multivariate normal and student t distribution. R package version 1.4-9.
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12, 1100-1122. doi: 10.1177/174569161617693393

Yocarini, I. E., Bouwmeester, S., Smeets, G., & Arends, L. R. (2018). Systematic comparison of decision accuracy of complex decision rules combining multiple tests in a higher education context. *Educational Measurement: Issues and Practice*, 37, 24-39. doi: 10.1111/emip.12186

S

Samenvatting
(Summary in Dutch)

S

Door studenten binnen een curriculum herhaaldelijk te toetsen, wordt de prestatie van studenten in het hoger onderwijs continu geëvalueerd. Op basis van de scores op deze toetsen wordt een schatting gemaakt van de bekwaamheid van studenten, waar vervolgens verschillende beslissingen op worden gebaseerd. Op het niveau van het curriculum worden toetsen gecombineerd om te beslissen of studenten verder mogen met hun studie of dat studenten voldoen aan de eisen om hun diploma te ontvangen. Op het niveau van een individuele cursus beslist de docent of de prestatie van de student voldoet aan de eisen om een voldoende (cijfer) te krijgen voor de cursus. Het doel van dit proefschrift is om de beslissingen te evalueren die in het hoger onderwijs genomen worden op basis van de prestaties van studenten, beslissingen zowel op het niveau van het curriculum waarbij meerdere toetsen gecombineerd worden (Hoofdstuk 2 en 3), als op het niveau van een individuele toets (Hoofdstuk 4 en 5). Om de kwaliteit van het diploma van een studieprogramma te waarborgen, zullen deze beslissingen over studenten valide moeten zijn, in die zin dat studenten die het diploma ontvangen daadwerkelijk bekwaam zijn en aan de eindkwalificaties voldoen die de hoger onderwijsinstelling hieraan heeft verbonden.

In het Nederlands hoger onderwijs wordt op basis van het bindend studieadvies (BSA) aan het eind van het eerste jaar van de bachelor beslist of studenten verder mogen met hun studie. Om de accuraatheid van deze BSA-beslissing te evalueren onder verschillende complexe compensatoire en conjunctieve beslisregels in verschillende realistische hoger onderwijs curricula, wordt in Hoofdstuk 2 een simulatiestudie beschreven die is gebaseerd op empirische data. De resultaten tonen aan dat de accuraatheid van de BSA-beslissing afhankelijk is van de mate waarin compensatie is toegestaan. Zowel het vereiste gemiddelde cijfer, het vereiste minimum cijfer per toets, als de combinatie hiervan zijn hierbij van belang. Over het algemeen zijn er binnen een compensatoire beslisregel minder fout-negatieven (d.w.z. studenten die in werkelijkheid bekwaam zijn, maar op basis van hun toetsscores een negatief BSA krijgen en niet door mogen met hun studie) en meer fout-positieven (d.w.z. studenten die in werkelijkheid niet bekwaam zijn maar op basis van hun toetsscores een positief BSA krijgen en toch door mogen met hun studie), vergeleken met conjunctieve beslisregels. Ook laten de resultaten zien dat de meest accurate beslissing afhankelijk

is van de gemiddelde betrouwbaarheid van de toetsen, de gemiddelde correlatie tussen de toetsen, en het aantal herkansingen dat is toegestaan. Uit de resultaten blijkt dat één van de redenen om compensatie tussen cursussen toe te staan, namelijk dat het gemiddelde cijfer betrouwbaarder is, niet opgaat in alle situaties, maar dat de accuraatheid van de complexe compensatoire beslisregel afhankelijk is van de context.

In Hoofdstuk 3 zijn de consequenties van het toestaan van compensatie in de eerstejaars BSA-beslissing op de prestatie in een tweedejaars vervolgvak bekeken. Hiervoor zijn data van een Bacheloropleiding Psychologie en een Bacheloropleiding Rechten gebruikt. Specifieke vervolgvakken, waarin de cursusstof voortbouwt op materiaal uit een (eerstejaars) voorgaand vak, zijn hierbij interessant, omdat ze inzicht geven over de gevolgen van het toestaan van compensatie tussen cursussen op kennis-accumulatie. Onze studie ligt hierbij in het verlengde van eerdere studies, waarbij er in deze studie gekeken wordt naar prestaties op vervolgvakken voor verschillende groepen studenten met een latente klasse regressieanalyse. Hierbij werd onderscheid in de latente klasse gemaakt op basis van studenten die gelijke latente studieprocessen lieten zien in het gemiddelde cijfer in het eerste jaar, de spreiding in eerstejaars cijfers, het aantal gecompenseerde cursussen en het totale aantal herkansingen in het eerste jaar. In de twee bacheloropleidingen werd onderscheid gemaakt tussen drie klassen studenten. Voor de klasse met studenten wiens eerstejaars prestatie laag was, was het gemiddelde cijfer op het eerste vak onder het vereiste gemiddelde cijfer (< 6.0), terwijl het gemiddelde cijfer op het vervolgvak voor deze groep onvoldoende was (< 5.5). Deze resultaten lijken te suggereren dat het compenseren van een eerste vak voor studenten met slechte prestaties in het eerste jaar negatieve gevolgen zou kunnen hebben in hun kennis-accumulatie, waarbij de prestatie op een vervolgvak onvoldoende zou kunnen zijn. Echter, de resultaten laten ook zien dat het eerste vak gecompenseerd wordt door studenten in elk van de drie klassen, maar dat de gemiddelde prestatie op het vervolgvak voor studenten in de twee overige klassen niet onvoldoende is. Dit zou erop kunnen duiden dat studenten met hogere prestaties in het eerste jaar geen negatieve gevolgen in kennis-accumulatie ervaren wanneer zij een eerste vak compenseren.

De evaluatie van de BSA-beslissing in het hoger onderwijs, waarbij meerdere toetsen gecombineerd worden, laat zien hoe belangrijk de kwaliteit van de individuele toetsen is voor de accuraatheid van de beslissing. Derhalve hebben we de focus vervolgens verlegd naar beslissingen op basis van de prestaties van studenten en het schatten van studenten hun ware score op individuele toetsen in het hoger onderwijs. In Hoofdstuk 4 is de accuraatheid van verschillende correctiemethoden voor gokken bij meerkeuzetoetsen geëvalueerd. Hierbij lag de focus specifiek op meerkeuzetoetsen waarin incorrecte responses niet direct werden bestraft door minpunten en studenten hun optimale en meest gebruikelijke strategie daarom is om te gokken (zoals gewoonlijk in het Nederlands hoger onderwijs). Klassiek gezien wordt er gecorrigeerd voor gokken middels *formula scoring*. Alternatieve correctie-methoden, zoals de *extended* klassieke methode, de (*extended*) beta binomiale modellen en modellen vanuit de item response theorie maken gebruik van kenmerken uit de steekproef. Door middel van simulaties is de accuraatheid van de geschatte ware kennis van studenten beoordeeld voor de verschillende correctie-methoden in verschillende realistische hoger onderwijs curricula. De resultaten tonen aan dat het schatten van ware scores in meerkeuzetoetsen verbeterd zou kunnen worden door de *extended* klassieke correctie methode voorgedragen door Calandra (1941) en Hamilton (1950) te gebruiken, of door een methode te gebruiken die item kenmerken meeneemt in het schatten van ware scores, zoals de door ons voorgestelde gewogen correctie voor item moeilijkheid.

Ten slotte zijn in Hoofdstuk 5 de beslissingen onderzocht om cijfers toe te kennen aan de toetsscores (d.w.z. *standard setting*) van studenten en om op basis daarvan te beslissen of een student geslaagd of gezakt is. De accuraatheid van drie *standard setting* methoden (de klassieke absolute methode, en de Cohen en Hofstee compromis methoden; Cohen-Schotanus & van der Vleuten, 2010; Hofstee, 1983) die elk houdbaar zijn in kleinschalige, niet-gestandaardiseerde toetsen werd beoordeeld. Ook hiervoor is een simulatiestudie uitgevoerd om studenten hun ware en geschatte cijfers te verkrijgen en om realistische contexten in het hoger onderwijs na te bootsen. De resultaten laten zien dat de klassieke absolute methode de ware score van studenten in bijna alle gesimuleerde situaties onderschat, terwijl de Cohen en Hofstee methoden

de ware score in slechts enkele situaties overschat. De resultaten laten zien dat het gunstig zou kunnen zijn om informatie van de steekproef in de vorm van de best scorende studenten mee te nemen in het toewijzen van cijfers.

C

Curriculum Vitae

C

Curriculum Vitae

Iris Eleni Yocarini was born in Zeist, The Netherlands, on the 28th of September 1990. After completing her secondary education (bilingual gymnasium) at Het Christelijk Lyceum in Zeist, she started studying International Business Administration at the Rotterdam School of Management at the Erasmus University Rotterdam (EUR) in 2008 from which she obtained her diploma in 2011. During her final year, Iris enrolled in the Psychology bachelor at the Erasmus School of Behavioral and Social Sciences, following her interest in human decision-making. During this bachelor, Iris participated in the multidisciplinary (English) Erasmus Honors Program, worked as a student research assistant, and pursued her ambition to explore new territories by studying a semester abroad at the University of Western Ontario in London, ON, Canada. Iris obtained her Psychology bachelor diploma in 2013 and subsequently started her Psychology master Brain and Cognition at the EUR as well. During her master studies, Iris participated in the two-year Advanced Research Program, worked as an academic teacher in first-year statistics workgroups, worked as a research student assistant, and was a student member in the master educational committee. Following her interest in data analyses and statistics, Iris enrolled in the Psychology master Methods and Techniques at the School of Social Sciences at Leiden University, for which she did an internship at the department of Medical Psychology at the Erasmus Medical School in Rotterdam. In 2015, Iris received both her master diplomas (cum laude) and started working as a PhD student at the Department of Psychology, Education and Child Studies at the EUR. Whilst conducting her PhD research, Iris was a student member of the Interuniversity School of Psychometrics and Sociometrics (IOPS), presented her research at various national and international conferences, served as a consultant for data analyses, followed multiple courses, taught several psychological methods and statistics courses, served as a reviewer, and received the 'Graduate School Award for PhD Excellence: Best Poster' from the Erasmus Graduate School of Social Science and the Humanities. Furthermore, Iris organized the DPECS Graduate Research Day twice, set up an inter-faculty book club and the Methods and Techniques research meetings, organized colloquium meetings for which she invited guest speakers, and organized department drinks.

Publications

- Van Schie, K., Wanmakers, S., **Yocarini, I. E.**, & Bouwmeester, S. (2016). Psychometric qualities of the Thought Suppression Inventory- Revised in different age groups. *Personality and Individual Differences*, *91*, 89-97. doi: 10.1016/j.paid.2015.11.060.
- Yocarini, I.E.**, Bouwmeester, S., Smeets, G., & Arends, L.R. (2018). Systematic comparison of decision accuracy of complex compensatory decision rules combining multiple tests in a higher education context. *Educational Measurement Issues and Practice*, *37*, 24-39. doi: 10.1111/emip.12186

Papers

- Van Meggelen, M., Morina, N., Van der Heiden, C., Brinkman, W. P., **Yocarini, I.E.**, Tielman, M.L., Rodenburg, J., Van Ee-Blankers, E., Van Schie, K., Broekman, M.E., & Franken, I.H.A. (submitted). A randomized controlled trial on the efficacy of a computer-based intervention with elements of virtual reality and limited therapist assistance for the treatment of post-traumatic stress disorder.
- Yocarini, I.E.**, Bouwmeester, S., & Jongerling, J. (submitted). Correcting for guessing in estimating true scores in higher education tests.
- Yocarini, I.E.**, Bouwmeester, S., & Jongerling, J. (submitted). Comparing the validity of different cut-score methods in higher education.
- Yocarini, I.E.**, Bouwmeester, S., Smeets, G., & Arends, L.R. (submitted). Allowing course compensation in higher education: A latent class regression to evaluate second-year performance.
- Yocarini, I.E.**, Kickert, R., Jongerling, J., Meeuwisse, M., & Bouwmeester, S. (in preparation). Tracking higher education students' study time investment intensively.

Presentations

- Jul 2018 International Meeting Psychometric Society (IMPS) 2018, NY, USA
Presentation on a network study on students' study time allocation
- Jun 2018 Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS) summer conference 2018, University of Amsterdam, NL
Presentation on testing in higher education
- Nov 2017 DPECS Graduate Research Day 2017, Rotterdam, NL
Presentation on simulation studies as a research method
- Jul 2017 International Meeting Psychometric Society (IMPS) 2017, Zurich, Swiss
Presentation on validity of different cut-score methods in higher education
- May 2017 29th Annual Convention of the Association for Psychological Science, Boston, USA
Poster presentation on correction for guessing methods
- Oct 2017 Brain and Learning inaugural speech P. Verkoeijen, Avans Hogeschool, NL
Workshop on testing in higher education
- Jul 2016 International Meeting Psychometric Society (IMPS) 2016, Ashville, North Carolina, USA
Presentation on simulation study on the accuracy of different decision rules
- Jun 2016 Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS) summer conference 2016, University of Twente, Enschede, NL
Poster presentation on simulation study to assess accuracy of different decision rules
- May 2016 Education Research Days (ORD) 2016, Rotterdam, NL
Paper presentation on compensation and workshop with professionals in 'Kom over de brug' [bridging the gap to practice]
- Mar 2016 Center of Educational Learning, Erasmus University College, NL
Pitch on compensatory decision rules
- Nov 2015 Lunch seminar on educational quality, Erasmus University Rotterdam, NL
Presentation on compensatory decision rules
- Oct 2015 Educational quality meeting, University of Twente, Enschede, NL
Presentation on compensatory decision rules in higher education
- Sep 2015 Standard-setting Conference: International state of research in the Nordic countries, CEMO, University of Oslo, Norway
Presentation on compensatory decision rules (invited speaker)
- Jun 2015 Education Research Days (ORD) 2015, Leiden, NL
Poster on simulation study to assess compensatory decision rule

D

Dankwoord
[Acknowledgements]

D

And it's done, mijn proefschrift is af! En hoewel mijn naam mooi op de titelpagina prijkt, was dit mij nooit gelukt zonder de steun, aanwezigheid en inspiratie van veel anderen. Hier wil ik jullie allemaal graag voor bedanken, met in het bijzonder enkele van jullie.

Allereerst Samantha, bedankt voor zoveel! Zonder jouw support, loyaliteit en doorzettingsvermogen was dit proefschrift er niet geweest. Al voor mijn PhD project wist je mij te inspireren voor de methoden en technieken van de psychologie. Ik had mij geen betere dagelijks begeleider kunnen voorstellen en ben je ontzettend dankbaar voor alle tijd die je, ondanks je eigen situatie, tijdens het hele project in onze samenwerking hebt gestoken. Door het vertrouwen dat je mij vanaf de eerste dag gaf als onderzoeker en docent heb ik de vrijheid gehad om zelf te ontdekken hoe ik met plezier werk. Het sparren tijdens onze meetings zorgde er altijd voor dat ik, hoe ik ook binnenkwam, weer gemotiveerd verder kon. Ik kan je kritische blik erg waarderen en heb er veel van geleerd. Ik kijk met veel plezier terug naar onze samenwerking en vele congres tripjes; van Oslo tot New York.

Mijn promotoren, bedankt dat jullie mij de kans hebben gegeven om te promoveren en bedankt voor de tijd en feedback die jullie mij hebben gegeven. Ik heb hier veel van kunnen leren. Lidia, bedankt voor je steun en aanwezigheid bij al mijn presentaties door het gehele land. Guus, dankzij jou ben ik de link van ons onderzoek met de praktijk, die zo belangrijk is, niet uit het oog verloren.

Voor iemand die geen koffie drinkt, was ik de afgelopen jaren vaak bij het koffieapparaat te vinden. Joran (maar meteen ook Rob bedankt voor het laten kapen van jullie 3-tot-5 uur pauze(s)), bedankt dat ik je vaak tijdens je koffiepauze mocht storen met allerlei vragen. Zonder jouw enthousiasme, kennis en interesse in statistiek en R had ik mij waarschijnlijk een uitzonderlijke nerd gevoeld bij DPECS.

Voor ik mijn PhD project begon hebben verschillende onderzoekers mij geïnspireerd en gemotiveerd die ik ook graag wil bedanken. Katinka, bij jou kreeg ik als jouw student-assistent interesse in onderzoek en meteen hands-on ervaring met cognitief onderzoek, zo was ik al in mijn tweede jaar kind aan huis in het lab. Rolf, Peter, en Samantha, als jullie student-assistent op het replicatie project heb ik geleerd met een

kritische blik te kijken naar onderzoek in de psychologie. Diane, jouw opmerking dat niet veel studenten statistiek leuk vinden en ik daar wellicht wat mee zou moeten doen heeft geleid tot de keuze voor een tweede master, bedankt voor dit laatste zetje. Al tijdens mijn tijd als Psychologie student, hebben mijn mede M&T collega's mij al geïnspireerd met hun aanstekelijke passie voor het M&T onderwijs, het was erg fijn met jullie samen te werken.

Ik voelde mij al vanaf mijn derde dag op mijn gemak als PhD student dankzij mijn geweldige kamergenoten (dag 1 bestond vooral uit vragen beantwoorden en dag 2 belandde ik in een verjaardag viering zonder cadeau). Marieke, Denise, Lara, Milou, en Donna, bedankt voor alle support! Van perfect getimed netflix-tips, luisterende oren, wonderlijke woordspelingen, (free paper) hugs, GOT-nabeschouwingen zodat ik het ook nog begreep, (après-ski) verjaardagen, vergelijkingen van marathon schema's, reistips van Indonesië tot Japan, discussies over onze ideeën (en soms frustraties) over de wetenschap, tot onze roomie dinners. Zonder jullie, een v(h)eilige plek waar ik mij terug kon trekken om te knallen (in tijden van motivatie) of anderen van het werk kon houden (in die andere tijden), was ik niet elke dag met zoveel plezier naar werk gefietst.

Mijn paranimfen, Denise en Lara, bedankt dat jullie na 4 jaar gezelligheid en waar nodig steun, nog een keer naast mij staan. Denise, jouw creativiteit blijft mij verbazen, ik mis je hilarisch geknutselde insta posts. Lara, met een kamer in de vorm van een strandtent, bibliotheek, zonnig terras of hostel in Boston, was jij als een parttime kamergenoot in-another-room. Bedankt voor alle keren dat ik (even) mocht buurten om altijd weer weg te lopen met nieuw materiaal; van presentatie lay-outs tot interieuradviezen en van efficiënte mailbox managementmethoden tot nieuwe borrelafspraken.

Al had ik een kamer, na de grote PhD verhuizing naar de 16^e verdieping leken we soms een grote vissenkomp en werd de werkpret alleen maar groter. Beste PhD collega's, zonder jullie discussies over de lekkerste pindakaas en humus bij de lunches, discussions on cultural differences and tasty food during our second lunch breaks, constructieve feedback bij de pubgroup meetings, interesse in de colloquia, gezelligheid bij de borrels, fanatieke instelling op de sportdagen (maar ook tijdens het

rond-de-tafel pingpongen om half drie) en deelnames aan de GRDs was mijn PhD tijd maar leeg en saai geweest, thanks for all the fun!! In het bijzonder, nochtans in willekeurige volgorde: Willemijn, Keri, Nouran, Miranda, Sabrina, Anniek, Jacqueline, Loïs, en Işil. Ilse, hoe leuk dat jij na onze master ook je PhD onderzoek kwam doen en we zo ook samen konden blijven sporten en ontbijten, om weer meer te borrelen. Dankzij onze trip naar Lissabon kan ik ‘congres crashen’ mooi van mijn academische bucket list afstrepen. Julia, bedankt voor de gezelligheid en al je lieve woorden. Rob, jij motiveerde mij vaak met je interesse voor het onderwijsonderzoek en het is een gegeven dat ik zonder jouw enthousiasme en humor niet zoveel had gelachen tijdens werk.

Eerder op de 13^e verdieping, toen ik net begon en nog een PhD-er-zonder-M&T-sectie was, ben ik goed opgenomen door mijn fijne collega’s bij HLP, bedankt hiervoor, met in het bijzonder Marit, Gertjan, Vincent, Mario, Gerdien, Tamara, Wim, en Martine.

Ook ben ik mijn (PhD) collega’s bij het IOPS dankbaar. Als enige Rotterdammer heb ik met veel van jullie mee mogen liften (letterlijk zelfs, bedankt Nikky) op onze congressen, cursussen, workshops, en borrels, bedankt voor jullie open- en gezelligheid!

Naast werken moet er natuurlijk ook ontspannen en genoten worden, in sommige tijden was dat wat meer nodig dan in andere. Gelukkig maakte ik daar ook altijd tijd voor en kon ik zo alle (wellicht ietwat lange) verhalen over mijn leven als PhD student kwijt, stoom afblazen, op z’n tijd doen aan wat zelfreflectie, of simpelweg kaas eten; bedankt Suze, Annick, mijn IBA-vrienden, mijn oud-huisgenoten en my fellow psychos! Jullie verhalen, humor, en tips als niet-wetenschappers hielpen mij de boel altijd weer te relativieren.

Voor de eindsprint van dit proefschrift heb ik veel inspiratie opgedaan en mijn drive hervonden in mijn tweede thuisland Griekenland, bedank Artisa en Villa Mariëlle voor de goede verzorging.

Ten slotte gaat de meeste dank uit naar mijn familie, ik bof maar. Lieve zussen, bedankt dat jullie er altijd voor mij zijn en mij inspireren. Niki je hebt mij laten zien dat als ik het (onderzoek) ooit zat mocht zijn, het roer altijd om kan: no spang. Xenia, jouw drive om alles uit het leven te halen lijkt onuitputtelijk en is bijzonder: yolo. Alexi, jouw ontwikkeling is indrukwekkend, je laat zien dat de mogelijkheden eindeloos zijn, als je het maar probeert. Ik ben een hele trotse grote zus!

Lieve papa en mama, van het kiezen van witte muurverf tot het vertalen van mijn (soms chaotische) gedachten, bedankt dat jullie altijd luisteren en mij eindeloos steunen. Dankzij de vrijheid die jullie mij gaven om mijn eigen keuzes te maken en de wereld te ontdekken kon ik mijn nieuws-en leergierigheid achterna. Mama, zonder jouw doorzettingsvermogen en (nochtans bijna Rotterdamse) niet-lullen-maar-poetsen mentaliteit als voorbeeld was ik niet geslaagd als onderzoeker. Papa, bedankt dat ik altijd over je schouder mocht meekijken. Dankzij jou beheers ik nu een perfecte balans van plannen en probleemoplossend vermogen, wat niet alleen handig is tijdens het klussen maar ook in mijn onderzoekswerk. Bedankt dat ik alvast even kennis mocht maken met het gepensioneerde leven, ik kijk er zeker niet tegen op.

