





KIMBERLEY LEK

teacher knows best?



ON THE (DIS)ADVANTAGES OF
TEACHER JUDGMENTS AND TEST
RESULTS, AND HOW TO OPTIMALLY
COMBINE THEM



KIMBERLEY LEK

teacher knows best?

**ON THE (DIS)ADVANTAGES OF
TEACHER JUDGMENTS AND TEST
RESULTS, AND HOW TO OPTIMALLY
COMBINE THEM**

Printed by Ipskamp Drukkers, Enschede
ISBN/EAN: 978-94-028-1950-2
Cover designed by Kimberley Lek with Canva.com
Copyright © 2020, K. M. Lek. All Rights Reserved.

Teacher knows best?

On the (dis)advantages of teacher judgments and test results,
and how to optimally combine them

De leraar weet het het beste?

Over de voor- en nadelen van docentoordelen en toetsresultaten,
en hoe deze optimaal te combineren zijn
(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 22 april 2020 des ochtends te 10.30 uur

door

Kimberley Martine Lek

geboren op 25 juni 1990
te Aarlanderveen

Promotor: Prof. dr. A.G.J. van de Schoot
Copromotor: Dr. R.C.W. Feskens

Dit proefschrift werd (mede) mogelijk gemaakt met financiële steun van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO Talent Grant Nr. 406-15-062).

Contents

Introduction	1
Brief history of testing	1
Testing debate	3
The example of the Dutch ‘Eindtoets Basisonderwijs’ (EPST)	4
Topics covered in this dissertation	5
Some final comments	7
I The test	9
1 How to interpret confidence intervals in intelligence tests	12
Abstract	12
Introduction	12
The steps that precede the construction of a confidence interval	13
The construction of confidence intervals	18
Interpreting WISC-III ^{NL} confidence intervals	21
Assumptions	25
Some frequent misinterpretations of the confidence interval	25
Practical implications	26
Take home message	28
Acknowledgements	28
2 A comparison of the single, conditional and person-specific Standard Error of Measurement: what do they measure and when to use them?	30
Abstract	30
Introduction	31
Classical Test Theory	32
Single Standard Error of Measurement	33
Person-specific Standard Error of Measurement	35

Conditional Standard Error of Measurement	36
Bias-variance trade-off	39
Simulation	45
Results	50
Discussion	67
Practical recommendations	69
3 Approximate measurement invariance	74
Introduction	74
The multigroup confirmatory factor analysis	76
Illustration	78
Discussion and Conclusion	87
Acknowledgments	89
II The teacher	91
4 Does socio-economic status, ethnicity and gender matter? Investigation of teacher recommendations in the transition from primary to secondary education in the Netherlands	94
Abstract	94
Introduction	95
Literature review	96
The current study	98
Results	106
Discussion	112
5 Development and evaluation of a digital expert elicitation method aimed at fostering elementary school teachers' diagnostic competence	120
Abstract	120
Introduction	120
Background Expert Elicitation	122
Methods	130
Results	132
Discussion	138

III	Teacher versus test	143
6	The transition from primary to secondary education in the Netherlands: Who knows best, the teacher or the test?	146
	Abstract	146
	Introduction	146
	Description of the data	148
	Discrepancy EPST and Teacher's Advice	148
	Match of advised level and the level eventually taken	148
	Pupils' switching-behaviour	150
	Adjustment in light of a higher EPST advice	152
	Conclusion	154
	Discussion	155
	Practical Implications	156
7	How the Choice of Distance Measure Influences the Detection of Prior-Data Conflict	158
	Abstract	158
	Introduction	158
	Prior-Data Conflict Criteria	159
	Simulation Study	165
	Results	167
	Conclusion	169
	Discussion	174
	False dichotomization	174
	Why limit ourselves to a <i>single</i> test result?	175
	Why should the teacher recommendation be reached <i>holistically</i> ?	176
	Why should the teacher judgment be <i>unaided</i> ?	176
	Why not <i>combine</i> teacher and test?	177
	Feedbackcycle	178
	Smart Testing	179
	How to proceed?	180
	Footnotes	182

Nederlandse samenvatting	186
Deel I: De toets	187
Deel II: De leerkracht	187
Deel III: Leerkracht versus toets	188
Curriculum Vitae	189
Publications	189
Acknowledgements	191
References	193

Introduction

“ In the sixth grade, everything is much more serious than in the previous years. It is not about having fun, it is about grades. I want to score as highly as possible, which makes the sixth grade so stressful. ”

Lesly, 2015

These words - translated from Dutch to English - were spoken by twelve-year-old Lesly. Lesly and other sixth graders were interviewed for the documentary ‘Citostress’ (Jansen, 2015) during their last year of primary education (sixth grade) in the period before, during and after the ‘Cito Eindtoets Basisonderwijs’ [translated: Cito’s End-of-Primary-School-Test (EPST)]. In the Netherlands, the result of the EPST-test is used in the track allocation process to secondary education and is therewith one of the most important and well-known tests in the school career of pupils. The documentary illustrates the emotional ‘roller coaster’ that sixth grade often is from the perspective of Lesly and other sixth graders, their teachers and their parents. We - for instance - see Lesly’s teacher trying to explain the school’s perspective on Lesly’s learning potential in secondary education as well as a sad and angry mum who believes her son is capable of much more. The response of Lesly and his parents to his disappointing EPST-score is also captured. We see how his classmate Farah is working very hard on her preparation for the EPST, trying to explain her work to her Turkish mum. In the meantime, we see how two other children are trained for the EPST by an external training institute, paid for by their parents. Before opening the envelope containing the result of the EPST, we see how some children and their parents have a clear preference for a score(range) in mind, whereas others - such as the Turkish mum - are less aware of the scorerranges, the accompanying track levels and the implications of these track levels for the future career of their child(ren). At the end of the documentary, two things have become crystal clear. First, the transition from primary to secondary education, including the EPST-test, is very stressful to all persons involved. Second, how children are prepared for the EPST and how much pressure they experience to do well largely depends on their teacher and parents.

Brief history of testing

The documentary by Jansen (2015) shows that tests like the EPST play a vital role in a child’s school career. The roots of these tests can be traced back to both European and American pioneers, where especially the American development of test theory encouraged

and enabled the large-scale applications of (high- and low-stakes) tests we witness today (Mertens, 2016). In the mid-nineteenth century, the American Horace Mann was the first to introduce a nation-wide written exam called the ‘common test’. In a time in which American educators changed their focus from educating the elite to educating the masses (see Haladyna, Haas & Allison, 1998), Mann believed that such a test would provide objective information about the quality of learning and teaching for all children. Horace’s ideas resulted in the first ‘testing movement’, in which nearly all U.S. cities adopted a series of tests to sort students, schools and teachers (Gallagher, 2003). Nearly half a century later, the influential American professor Thorndike promoted the application of scientific principles to educational testing, with the goal to systematically identify and segregate students according to their intellectual abilities. Other important pioneers include the French psychologist Binet who, at the beginning of the twentieth century, developed the first individually administered test of intelligence and Terman (1916) who expanded the usage of intelligence tests to facilitate educational placement and career tracking. As Gallagher (2003) describes it: “In this era, academic tracking became entrenched in schools as scores from intelligence tests dramatically changed the ways in which students were classified. Standardized tests were used to stratify students of different abilities into different curricular paths, thereby restricting their academic and social choices.” (p. 88). The American Educational Testing Service is founded in 1947, devoted to educational research and assessment.

A pioneer of educational testing in the Netherlands worth mentioning here is professor Adriaan De Groot. After visiting the U.S. Educational Testing Service in 1958, De Groot concluded that the evaluation of children’s achievement in the Netherlands would benefit from a more objective and scientific approach, as it was primarily based on ‘habits, intuitions and prejudices’ (Mertens, 2016). He developed the ‘Amsterdamse Schooltoets’ (translated: ‘Schooltest of Amsterdam’) and founded the National Institute for Educational Test Development ‘Cito’; a European counterpart of the Educational Testing Service in the U.S. With his work and that of his PhD students (including Mellenberg, van Naerssen and Wesdorp) a psychometric tradition was initiated in the Netherlands (see Van Strien & Hofstee, 1995) and the ‘Amsterdamse Schooltoets’ evolved into a test well-known to every Dutch citizen: the ‘Cito Eindtoets Basisonderwijs’ (Cito’s version of the EPST) to be completed at the end of primary education. Due to De Groot’s efforts, the Netherlands became one of the early adopters of a ‘measurement approach’ to education, in which learning goals were operationalized (see De Groot, 1961) and systematically assessed, based on scientifically sound tests.

Fast forwarding to this decade, the ‘measurement approach’ is embraced by international organizations as the OECD¹ and is now the norm in many countries worldwide. The use of national large-scale tests in education has expanded exponentially, especially in OECD- and middle-income countries (Verger, Fontdevilla & Parcerista, 2019). Not only the number of tests has increased, educators are also more and more held accountable for the results of their pupils. The term ‘era of testing and pressure’, used by Deci and Ryan (2016), seems an appropriate summary of the global testing culture observed by education policy analysts (see, for instance, the work of C. Smith). The flourishing role of high-stakes tests in the educational career of pupils is, however, a recurrent subject of heated debate among educational specialists, psychometricians, policy makers, educational practitioners and test takers.

Testing debate

Originally, tests were often introduced to reach more objective, transparent and fair judgments for all pupils. Horace Mann's 'common test', for instance, provided an alternative to the then prevailing public oral examinations, which were mainly aimed at 'showing off brilliant pupils' and 'glorifying teachers' (Gallagher, 2003). And Thorndike aimed to replace 'subjective assessments' with more scientific and thus objective measurement tools. In the Netherlands, the aim of the 'Amsterdamse Schooltoets' and later on the 'Cito Eindtoets Basisonderwijs' was and is noble: to prevent pupils' abilities to be misjudged based on - for instance - socio-economic status and ethnicity (see De Groot's bestseller 'Vijven en zessen', 1966 and the review by Mertens, 2016). Despite these applaudable aims, not everyone was - however - convinced of the merits of standardized testing. As Linn (2001) describes, there has been a long history of a "combination of enthusiastic support of standardized testing and strong disapproval" (p. 29). Already in 1922, Walter Lippmann, for instance, wrote a series of articles in the *New Republic* fiercely criticizing the usage of intelligence tests in schools (see Haladyna, Haas & Allison, 1998). By then, the Alpha and Beta standardized test - an intelligence test applied when recruiting soldiers during World War I - was used to draw the dubious (and faulty) conclusion that a man's intelligence was related to its skin color, favoring white men. Within schools, the popular IQ-tests were increasingly used to channel pupils, making the IQ prophecies self-fulfilling (see Knoester & Au, 2017). During the 1960s, the 'testing debate' was fuelled again by the civil rights movement, as they claimed that standardized tests were biased in terms of social class and racial/cultural background (see Gallagher, 2003). In line with Lippmann, they argued that tests "are artifacts of culture, and culture may not diffuse equally into all households. [...] subcultures vary in ways that inevitably affect test scores" (Kincheloe et al., 1996, p. 32; Gallagher, 2003). More recently, Freedle (2003) published a highly controversial article in the *Harvard Educational Review*, stating that the SAT - the most widely taken college admissions test in the U.S. - is biased against African American Students. Although Freedle's work is strongly criticized (see, for instance, Dorans & Zeller, 2004), an improved² replication study in 2010 by Santelices and Wilson (again in the *Harvard Educational Review*) was unable to invalidate Freedle's claims of the unfair treatment of African American students, at least for some parts of the test. Freedle's, Santelices and Wilson's work shows that eventhough current tests are not comparable with the early criticized (IQ-)tests due to the many advancements in the educational testing field, their results should just as well be interpreted with care. The examples so far have focused on the debated possible influence of racial- and socio-economic background on test results. It is important to stress, however, that the debate surrounding standardized testing is broader than that. One other possible dangerous characteristic of test results is that they *appear* objective, simply because they "are bathed in the language of science and measurement" (Knoester & Au, 2017, p. 6) and are expressed in a - sometimes misleading - precise metric (Blauw, 2018). Test users (e.g., test takers, schools, government, media) are therefore often tempted to interpret the results of tests too strictly. Misuses of tests occur regularly³ whereas the measurement profession has little resources to prevent or remedy such misconduct (Brennan, 2006). An other persisting criticism of standardized tests is that they are unable to cover all learning objectives and intangible effects of schooling. Odell already discussed this issue in 1928 and also De Groot was aware of this in 1966, although he did not perceive it as a major problem (see Mertens, 2016).

Finally, the judgments and expertise of educational professionals (e.g., teachers and other educational practitioners) are taken less seriously because of the increased focus on standardized tests. In the Netherlands, some might even argue that the degradation of the teacher as undisputed professional has started with De Groot's bestseller 'Vijven en Zessen' in 1966 (see Mertens, 2016). I'm briefly turning to the example of De Groot's 'Cito Eindtoets Basisonderwijs' (EPST) below.

The example of the Dutch 'Eindtoets Basisonderwijs' (EPST)

I started this introduction with a passage on the Dutch 'Cito Eindtoets Basisonderwijs' (Cito's EPST). As this test is one of the most important tests in a Dutch pupil's school career, it unsurprisingly is the test that is most fiercely debated in the Netherlands. While psychometricians defend the psychometric qualities of this test, especially educational experts and practitioners have argued against its usage. According to these practitioners, the EPST might actually enlarge rather than reduce the gap between pupils of different racial and socio-economic backgrounds. Arguments for this statement closely reassemble those discussed in the previous section. On 'Wij-leren.nl', a popular knowledge platform for teachers, Karels (2019) for instance criticizes the limited scope of the test, the impossibility to measure less tangible abilities such as discipline, motivation, self-efficacy, curiosity, et cetera and the need for reasonably advanced language skills to understand the test items. As reflected in the opening passage, additional issues, underlined by Karels, are the availability of practice materials and trainings, especially for relatively wealthy parents and the practice of 'teaching to the test', stimulated by the fact that schools are held accountable for the results on the EPST.

Based on the criticism of the Cito EPST, the role of this and other End-of-primary-school tests (EPSTs)⁴ in the Netherlands changed in 2015. No longer was ability tracking in secondary education primarily based on the result of the EPST, but on the advice of the teacher instead. Basically, this meant a return to the situation before de Groot's 'Amsterdamse schooltoets'. A reset to the 1960s, although the EPST could now be administered as a second-opinion after the teacher advice was finalized. In the ongoing debate that followed this policy change, questions emerged as 'is one of the two (the teacher or the EPST-test) better in guiding the transition from primary to secondary education?' 'To what extent are test results and teacher judgments biased by factors as the pupil's ethnicity, socio-economic background and gender?' 'How can we prevent or limit these biases?' In this PhD thesis, I have had the opportunity to investigate these and related questions, in the specific context of the EPST and beyond. My thesis is divided in three parts. Part I and Part II examine some dangers, biases and merits of test results and teacher judgment, respectively. Part III focuses on the comparison and potential combination of teacher judgment and test results.

Topics covered in this dissertation

Part I: the test

My work in this part touches upon three fundamental test theoretic concepts, which all influence the quality of a test result: validity, reliability and measurement invariance (see Brennan, 2006). According to the early work of Kelley (1927), a test is valid “if it measures what it purports to measure” (see Borsboom, Mellenbergh & Van Heerden, 2004). Over the last century, this broad definition of validity has changed and expanded repeatedly (Brennan, 2006; see for instance Messick, 1989; Kane, 1992; 2001), with validity now often being used as an ‘umbrella term’ covering many kinds of test-related issues (Borsboom, Mellenbergh & Van Heerden, 2004). One such test-related issue that threatens validity is that test results are regularly misinterpreted or even misused in the daily educational and psychological testing practice (Borsboom, 2006b; Brennan, 2006), as not all test users (e.g. teachers and educational psychologists) are ‘(test) data literate’ (Schenke & Meijer, 2018; Schildkamp & Poortman, 2015; Schildkamp et al., 2017). Test results are, for instance, often interpreted too strictly by these practitioners (Gardner, 2013). In other words, the uncertainty in the test result (expressed by, for instance, the standard error of measurement in tests based on the Classical Test Theory) is not taken into account or misjudged. As mentioned previously, this is a common phenomenon when people are presented with numbers that *seem* precise (Blauw, 2018). The phenomenon is termed ‘false precision’ and it is a well-known problem in the context of IQ-testing (Huff, 1993). Working closely with educational psychologists, I discovered that the many misconceptions about the precision of IQ-scores are aggravated by the often erroneous and vague description of confidence intervals (i.e., error bars based on the standard error of measurement) in the test manuals of common intelligence tests. Working closely with these psychologists, I wrote an easy-to-read article on the different types of confidence intervals in common intelligence tests, together with an explanation of the often ignored or unknown assumptions that underlie these confidence intervals. The resulting article can be read in **chapter 1** of this dissertation.

Opposed to the definition of validity, the definition of reliability has barely changed over the last decades. According to Brennan (2006, p. 3), reliability refers to “consistency of (test) scores across replications of a measurement procedure”. How exactly this ‘consistency’ is expressed depends on whether the test is based on Classical Test Theory (Spearman, 1904; Guilford, 1936; Gulliksen, 1950), Generalizability Theory (Cronbach et al., 1972; Cronbach, 1991) or Item Response Theory (Lord, 1952; Lord & Novick, 1968; Rasch, 1960). Within the Classical Test Theory, tests often report the earlier mentioned standard error of measurement (SEm; and/or the confidence intervals that are based on this SEm). One strict assumption underlying this SEm is that all pupils are measured with the same amount of measurement error (i.e., the same degree of consistency). Since this assumption is hardly tenable (Sijtsma, 2009; Molenaar, 2004), **chapter 2**, discusses two alternatives in which this assumption is loosened. By means of a simulation study, I show when to opt for which alternative based on characteristics of the test and target population. Note that just as the first chapter, this second chapter is written based on input from educational psychologists. Specifically, while talking with the educational psychologists, I came to realize that the ‘equal measurement error’-assumption underlying the single SEm does not match the ‘testing reality’ as

experienced by educational psychologists.

The final chapter of part I, **chapter 3**, is written on the topic of measurement invariance⁵. Loosely formulated, when a test is ‘measurement invariant’, this means that the test does not disadvantage or favor a certain group (i.e., group with certain socio-economic background, gender or ethnicity) over another. As such, Measurement non-invariance is another important test-related issue influencing the validity of a test (Borsboom, Mellenbergh & Van Heerden, 2002). The first advancements in measurement invariance testing stem from the 1960s and were developed as a response to critics who were worried that tests were unfair to minority examinees (see section ‘Testing debate’; Angoff, 1993). Since then, the measurement invariance field has faced many methodological advances. Recent advantages deal for instance with the problem that existing tests for measurement invariance are often cumbersome and overly strict (Van De Vijver, 2019). One such advancement is called Bayesian approximate measurement invariance (Muthén & Asparouhov, 2013; Van De Schoot et al., 2013) and is discussed in detail in chapter 3. Note that the context of this third chapter differs from the earlier two chapters in two major ways. First, this third chapter is written for models which relate observed (test) scores to attributes by using latent variables (examples of attributes being ‘math ability’ or ‘intelligence’; see Borsboom & Van Heerden, 2003). Measurement invariance testing for these models involves testing whether the relationship between observed (test) scores and the unknown attribute of interest is the same over groups (assuring that any two persons with the same level of the attribute obtain comparable observed (test) scores, regardless of group membership). In contrast, Classical Test Theory - as used in the previous two chapters - fixes the relationship between the observed test scores and the attribute (the so-called ‘true score’) axiomatically (Borsboom, 2006b), which makes testing for measurement invariance impossible (at least without introducing additional assumptions, see Millsap, 2011). Second, the example in the third chapter is drawn from survey research rather than from educational practice, as I had the opportunity to work with some prominent survey methodologists on this chapter. Note that the technique of Bayesian approximate measurement invariance testing is, however, equally applicable to educational and psychological attributes.

Part II: the teacher

Many daily decisions influence the learning opportunities of pupils, including ability grouping and instructional decisions, but also selection, grade retention and track allocation. All of these decisions are at least partly based on the judgment of a teacher. Therefore, it is of paramount importance that these teacher judgments⁶ are of a high quality. Other than for tests (see Part I), there are however few guidelines to assess the quality of teacher judgments⁷. In the context of the ‘Eindtoets’ (see section ‘The example of the Dutch Eindtoets (EPST)’), it is especially important to ensure that the teacher judgments are not unintentionally influenced by unrelated pupil characteristics (e.g., ethnicity, socio-economic status), akin to the measurement invariance requirement of tests (see Part I). Likewise, it is important to guarantee that pupils are not limited or privileged in their teacher’s judgments due to school characteristics (such as school region and school composition). The first chapter of Part II, **chapter 4**, therefore explores whether and how teacher judgments for almost all Dutch pupils in their transition from primary to secondary education were influenced by pupil- and school characteristics.

This chapter focuses on the period in which the role of the EPST changed (see again the section ‘The example of the Dutch Eindtoets (EPST)’), and is based on a large dataset collected by Statistics Netherlands (In Dutch: Centraal Bureau voor de Statistiek, abbreviated ‘CBS’).

One reason why the quality of teacher judgments is hard to assess, is that these judgments are often intangible and implicit. In this part’s second chapter, **chapter 5**, I therefore investigate how teacher judgments can be made explicit through a technique called ‘expert elicitation’ (O’Hagan, 2006). The advantage of this endeavour is that both the teacher judgments and the confidence teachers have in these judgments (see, for instance, Gabriele & Park, 2016) can now be formally assessed and evaluated, just as is possible with test scores⁸. Note that doing so also opens up other possibilities, such as the possibility to combine teacher judgment(s) with test results in a statistical way (see this dissertation’s discussion and chapter 7).

Part III: teacher versus test

In the context of the Dutch Eindtoets (EPST), the million dollar question is whether the EPST-result *or* the teacher judgment is more suitable in the track allocation process from primary to secondary education. In **chapter 6**, I attempt to answer this question, again using a large dataset collected by CBS. Among other things, teacher judgment and EPST-result are compared with the pupil’s placement in secondary education three years later, to investigate how predictive teacher judgment and EPST-result generally are.

The final chapter, **chapter 7**, resulted from the idea that the discussion surrounding the EPST-test might be too black and white. Instead of focusing on the question “who is ‘right’, the teacher *or* the test?” we could ask ourselves the question “how can we combine teacher judgment and test result such that we maximize the advantages and minimize the biases and dangers of both?”. When teacher judgments are elicited as explained in chapter 5, it is possible to combine teacher judgment and test results using Bayesian statistics (for a gentle introduction see Van De Schoot et al., 2014). The expert elicitation produces a so-called ‘prior’, which together with the (test) data results in a Bayesian posterior. This posterior constitutes a compromise of the teacher judgment and the test result, weighted by the judgment uncertainty of the teacher and the measurement uncertainty of the test (i.e., the standard error of measurement). Such a ‘compromise’ between teacher and test is, however, only meaningful when teacher and test do not disagree ‘too much’. ‘Too much’ disagreement can result in something called a prior-data conflict (Moshonov, 2006). Chapter 7 compares two different checks for prior-data conflict, specifically focusing on the question how robust these checks are when one of the ingredients of the checks (i.e., the choice of distance measure) is altered.

Some final comments

I would like to end this introduction with a few comments that are important, I believe, to place my dissertation into perspective. First, I made the deliberate choice to focus my first two chapters on Classical Test Theory rather than the in many respects superior Item Response Theory. Although the field of psychometrics has made impressive progressions

in Item Response Theory, just as Borsboom (2006b) concluded more than a decade ago, psychological practice and applied psychological research lack behind. The Dutch intelligence tests I was working with in chapter 1, for instance, was built on Classical Test Theory and the educational psychologists only completed a basic training in this theory. As I believe the psychometric field has a responsibility to bridge the ever increasing gap between psychometric advances and psychological and educational practice, I decided to focus on the current understanding of the psychologists and how I could help them one step further. Although I believe that we should work towards a future in which Item Response Theory is the prevailing testing theory, for now I agree with Heisser (2006) that “[...] throwing classical test theory out of the window would only impair the credibility of psychometrics, and increase the gap with psychology” (p. 458).

Second, for two chapters (chapter 1 and chapter 6) I decided that a Dutch, non-technical and easy-to-read article would be more suitable than a publication in an international scientific journal. While such a Dutch article is no doubtless less prestigious than a scientific one, I believe that with such an article I was able to make a larger societal impact. For the first chapter, I for instance had the possibility to present at the NIP (The Dutch Association of Psychologists) and the ‘day about intelligence’ (in Dutch: dag van de intelligentie). The results of the sixth chapter were extensively discussed in the media (see De Bruin, 2 April 2019, De Bruin, 3 April 2019 and Remie, 3 April 2019), by the ‘PO raad’ (the Primary Education Council) and in official policy documents (see the report by the ministry of Education, Culture and Science of 21 juni 2019). For this dissertation, both chapter 1 and chapter 6 were translated into English.

A final, practical comment: to save space, all additional materials and appendices are collected on my Open Science Framework (OSF) page (<https://osf.io/qrg4e/>).

Enjoy reading my dissertation!

PART I

Test



Chapter 1



HOW TO INTERPRET CONFIDENCE INTERVALS IN INTELLIGENCE TESTS?

Kimberley Lek, Wenneke Van De Schoot-Hubeek, Evelyn Kroesbergen & Rens Van De Schoot

A Dutch version of this article was published in *De Psycholoog*, November 2017

Chapter 1

How to interpret confidence intervals in intelligence tests

Abstract

Intelligence test reports usually mention a 95% confidence interval. Roughly speaking, such a confidence interval acts as a ‘margin of error’ with respect to the IQ score acquired by the test subject. But the exact meaning of this interval is often misunderstood. In this chapter, we brush up the use and correct interpretation of the confidence interval in intelligence tests. The examples in our exposition are based on the WISC-III^{NL}.

Keywords: confidence interval, intelligence test, WISC-III^{NL}

Introduction

A number of factors influence the IQ score attained by a child⁹ at a given moment in time. Such factors include work attitude and attention problems (Pameijer, 2014), observer-expectancy effects, search mistakes and calculation errors in processing the results, disturbances or malfunctions during the test (Tellegen, 2004), fatigue, mood, performance anxiety (Schouws, 2015), and so on. For these reasons, the attained IQ score cannot be interpreted as an absolute value (Tellegen, 2004); hence the use of a confidence interval which acts as a margin of error. The correct interpretation of this confidence interval, however, is another matter. Imagine the following: Julia has taken a WISC-III^{NL} test and her (T)IQ score is 97, with the corresponding 95% confidence interval ranging from 90 to 104. How would you interpret this confidence interval? In late 2014, we put this question to 293 NIP psychologists (the NIP is the Dutch Association of Psychologists). An analysis of the results yielded a variety of interpretations of the 95% interval. As such this is not surprising, since the WISC-III^{NL} manual does not explain anything about how these confidence intervals come about, nor does it offer any concrete interpretation (Tellegen, 2002). Apart from that, these intervals are known to be misinterpreted often (for example, see Hoekstra, Morey, Rouder et al., 2014). Our presentation on confidence intervals in intelligence tests at a NIP meeting for psychologists in education (22 January 2016) led to the present chapter aimed at brushing up knowledge about confidence intervals: how they come about and how they should be interpreted. We begin with the steps

preceding the construction of the confidence interval, which are comparable for most intelligence tests, but which we explain by referring to the WISC-III^{NL}. Next, we explain the required steps for constructing a 95% confidence interval, which again are similar for most intelligence tests.¹⁰ We then proceed to discuss how the confidence interval should be interpreted. We will address the assumptions behind the confidence interval and will also involve the frequent (but mistaken) interpretations we saw in our late 2014 survey. Our chapter will close with a practice related discussion on how psychologists should use the 95% confidence intervals in tests and questionnaires. Our explanations here are supplemented with an interactive app that we designed for the purpose of demonstrating the interpretation of confidence intervals. It can be used (for free) for educational purposes (see <https://osf.io/qrg4e/>). It is not limited to the WISC-III^{NL} but can be used with other intelligence tests as well.

The steps that precede the construction of a confidence interval

From item score to raw subtest score

First, the scores on items belonging to a specific subtest are added up to arrive at a raw subtest score (see table 3.1., p. 45, WISC-III^{NL} manual)

From raw score to standard score

Next, the raw subtest scores are converted to standard scores in order to (1) compare the (raw) subtest scores and (2) to make sure that a specific subtest score has the same relative meaning for the different age groups. The conversion is based on data collected from the ‘norm population’ (i.e., a more or less representative sample of 1.239 children aged 6-16). Conversion of raw scores to standard scores (mean 10, standard deviation 3) occurs for each age group (see the ‘slides’ in Figure 1.1) as well as for each subtest (the 10 subsets in each slide) separately. This begins with determining the distribution of raw test scores attained by the relevant norm sample for each of the age groups and subtests. This is shown in the upper part of Figure 1.1. For example, for the age group of 6.5, the mean raw test subtest score on incomplete drawings (OT) is 12.36, with a standard deviation of 2.98. The upper part of Figure 1.1 clearly illustrates that the raw score distributions can take up any form: either or not symmetrical, with or without a clear peak, and so on. The assumption of the WISC-III^{NL}, however, is that the raw scores would take up a normal distribution in the hypothetical case that all possible children within a specific age group had been tested (Glutting & Oakland, 1993). This means that, whatever the shape of the raw score distribution, the raw score distributions¹¹ are converted to normal distributions. This is shown in the lower part of Figure 1.1.

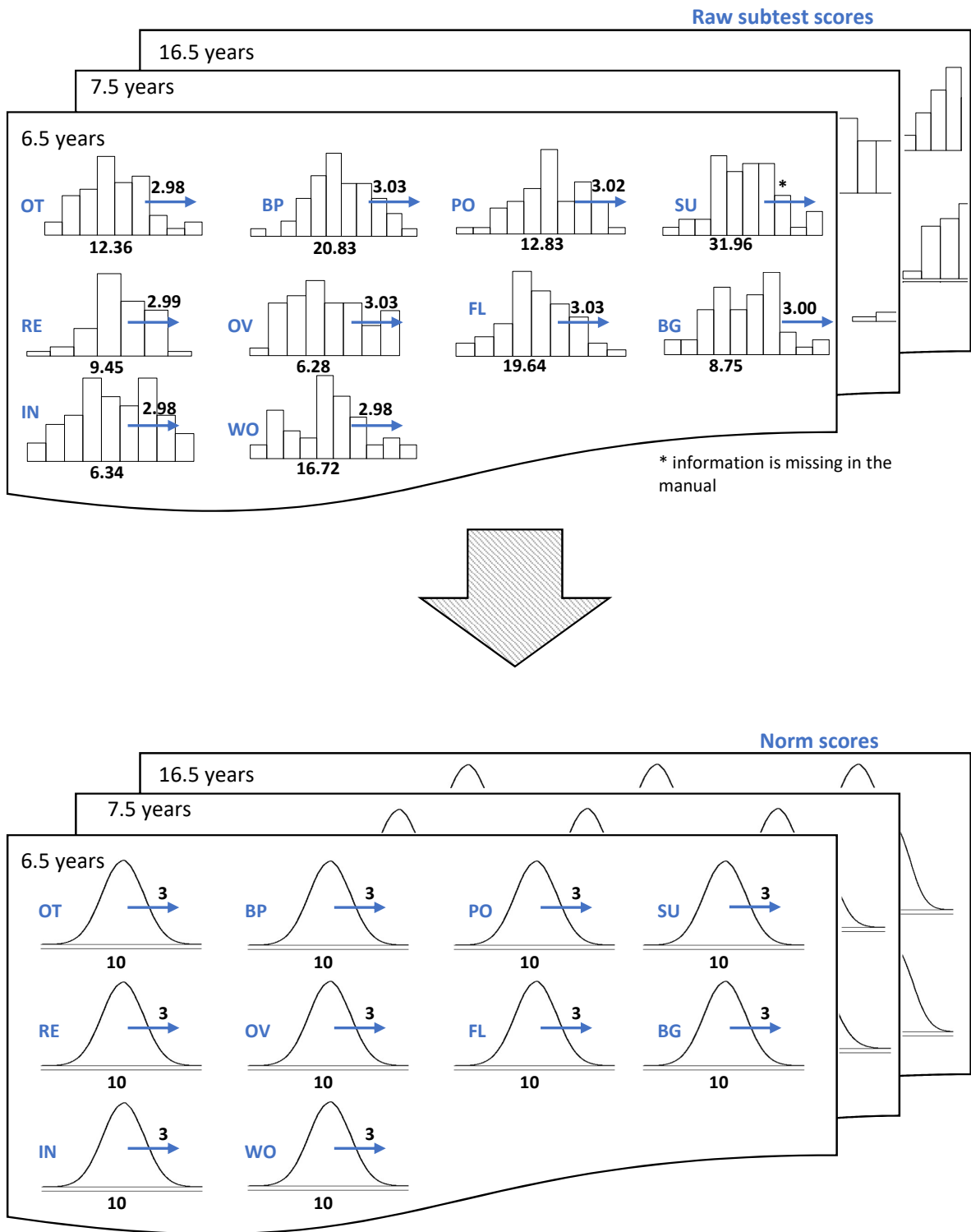


Figure 1.1: The WISC-III procedure for converting raw subtest scores (above) to standard scores (below).

Technical intermezzo 1: How are raw subtest scores converted to standard scores?

Figure 1.2. illustrates how raw subtest scores are converted to standard scores. After determining the raw score distribution for a particular age group and subtest (part A in Figure 1.2.), the first step is to determine the percentile scores (Crawford, 2004). Figure 1.2B. shows such percentile scores for two fictive children with raw subtest scores of 15.5 and 17. Percentile scores express the percentage of children in a particular age group that have acquired a particular raw subtest score or a score lower than that. Next, these percentile scores are matched with the z -scores within the standard distribution for the standard scores (part C in Figure 1.2.). For example, a z -score indicating the position of 90% of scores in the normal distribution of standard scores is matched with the raw score corresponding to the percentile score of 90. Multiplying the z -scores with the desired standard deviation 3 and adding this up to the desired standard score mean of 10 yields a standard score for a particular z -score, and thus, indirectly, for a particular raw score as well (part D, Figure 1.2).

From standard score to total score

The standard scores for the 10 subtests are added up to arrive at a single total score per child; see the first arrow in Figure 1.3.

From total score to IQ score

In order to arrive at an intelligence measurement comparable to outcomes of other intelligence tests, the distribution of total scores is converted to an IQ distribution; see the second arrow in Figure 1.3.¹²

Technical intermezzo 2: How are total scores converted to IQ scores?

Figure 1.4 visualises the conversion of total scores (x-axis) to the corresponding IQ scores (y-axis)¹³. Total scores have a large range (10 to 190), which by referring to the norm population, is reduced to the allowed range of WISC-III^{NL} IQ scores (45 - 145). The jittery line is caused by rounding the numbers. Note that in comparison to the top end of the WISC-III^{NL} scale, equally wide differences in total scores at the bottom end of the WISC-III^{NL} scale lead to a smaller bandwidth of differences in the IQ scores. The red dots show what the conversion of total scores to IQ scores looks like for extremely low (< 3 SDs) and extremely high (> 3 SDs) total scores. It is especially worth pointing out that children with a total score higher than 152 all get the same IQ score: 145. In other words, the WISC-III^{NL} IQ, does not differentiate for children with a total score range of 152 - 190.

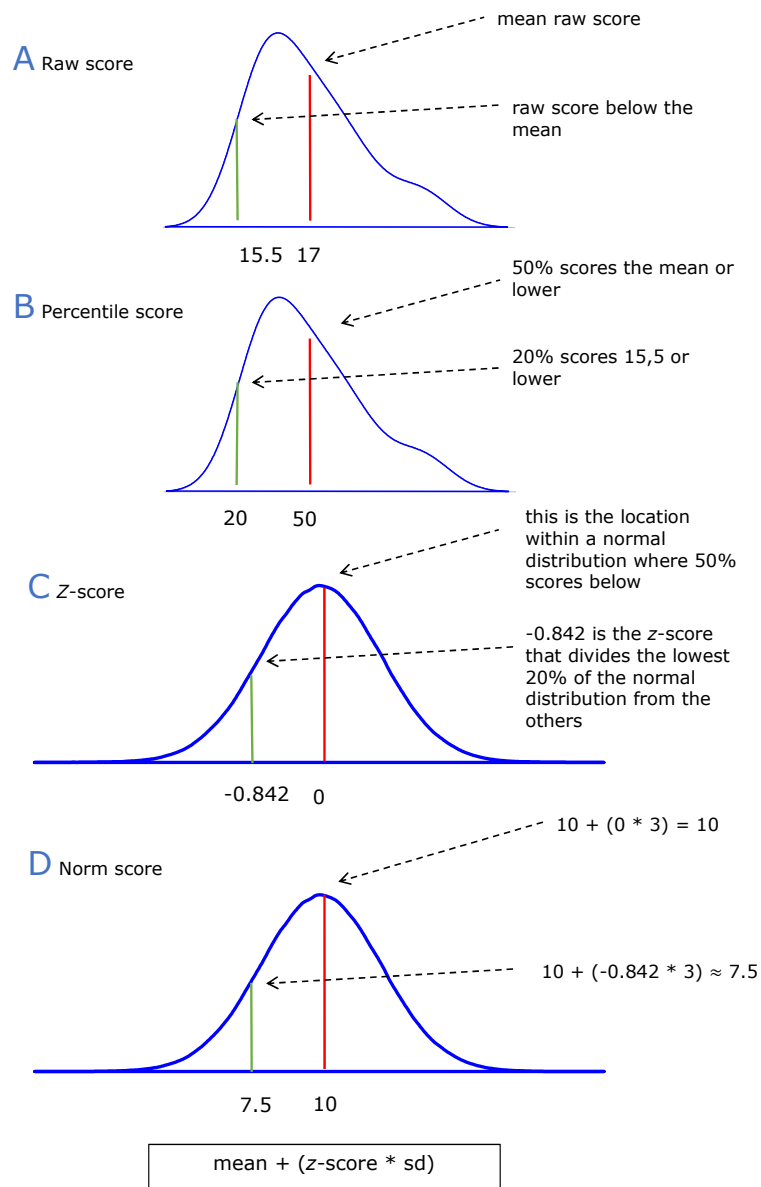


Figure 1.2: Example of converting of raw scores (A) to percentile scores (B), z-scores (C) and finally standard scores (D). Note that A and B observe the distribution according to the data, while C and D observe the normal distribution.

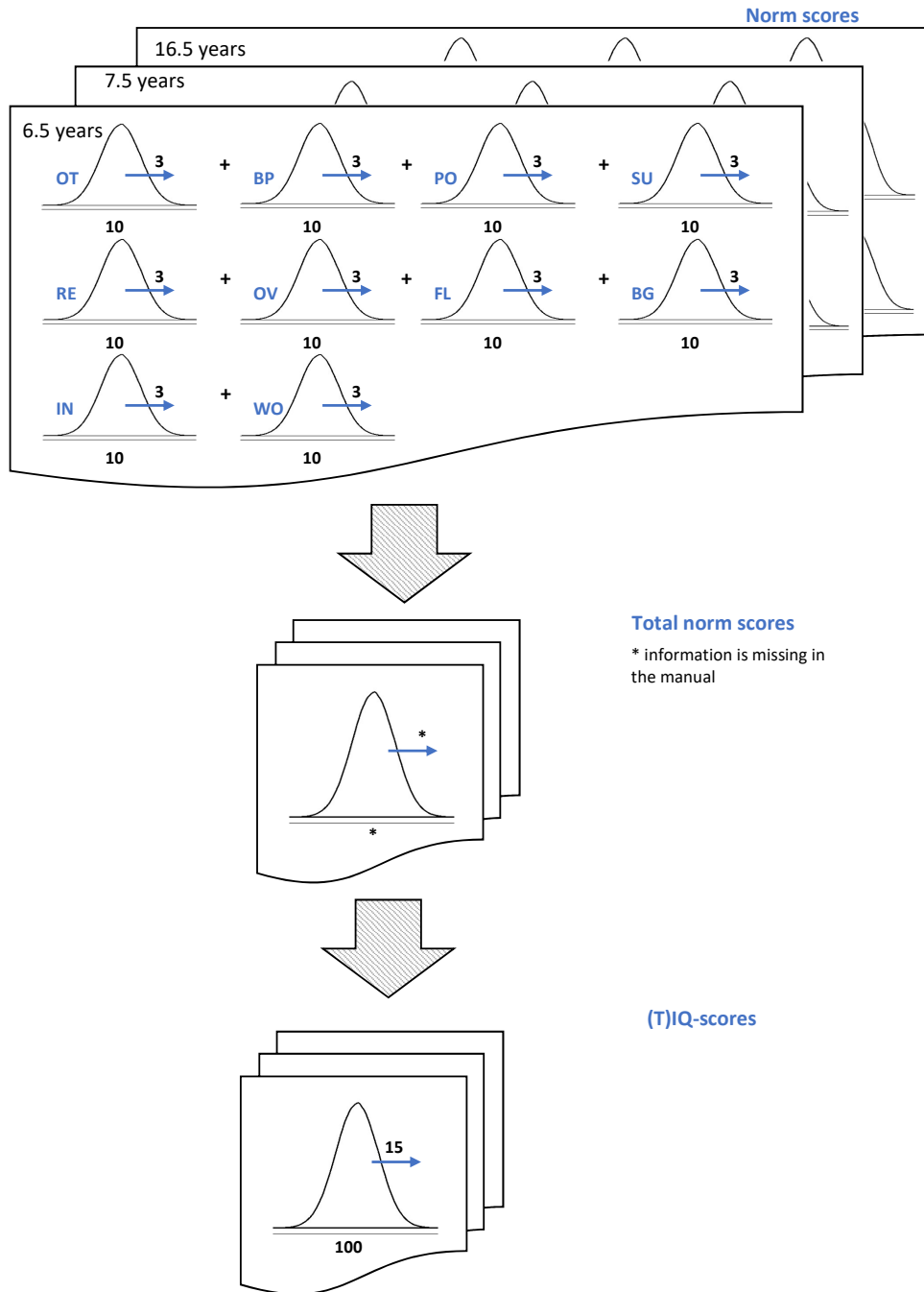


Figure 1.3: WISC-III procedure for converting standard scores first to total scores and then to IQ scores.

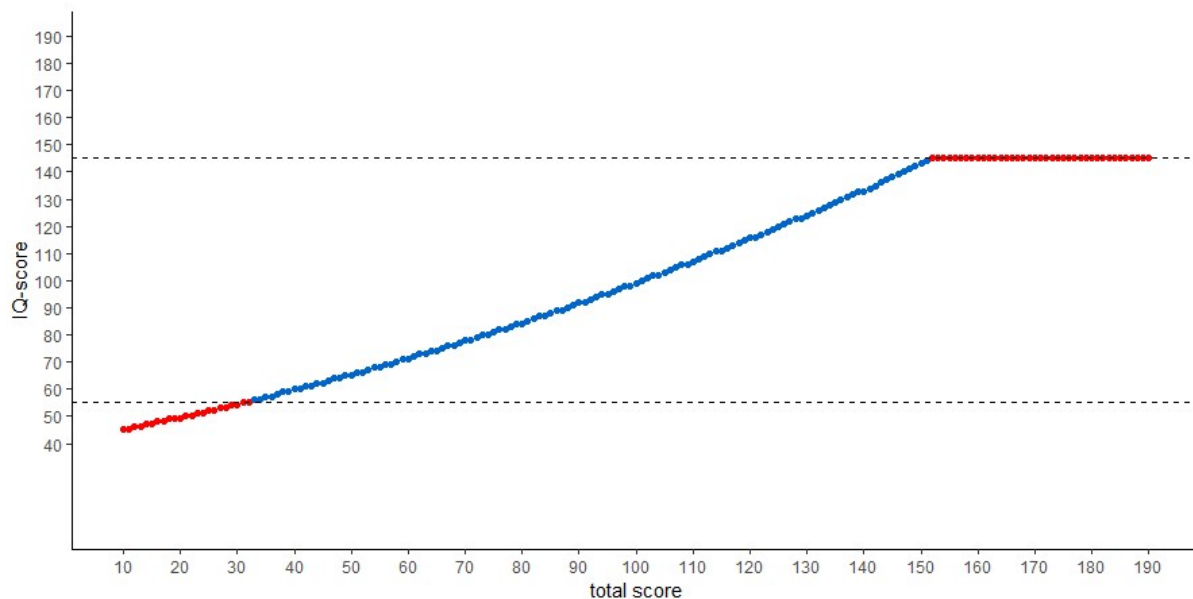


Figure 1.4: Example of converting total scores (x axis) to IQ scores (y axis).

The construction of confidence intervals

A confidence interval is computed by subtracting and adding up an estimate of error of measurement to the estimate of the IQ score for a ‘ z -number of times’:

$$\text{Estimate IQ} \pm z * \text{estimate error of measurement} \quad (1.1)$$

The z in formula (1.1) can take up different values; for example, with a value of 1.96 for z , a 95% confidence interval can be obtained. The error of measurement that the confidence interval is based on can be estimated with two different formulae (see COTAN, p. 27). The result of the one is called ‘standard error of measurement’; of the other ‘standard error of estimation’ — see technical intermezzo 3. The choice of formula has an effect on how the resulting confidence interval should be interpreted (see the next sections).

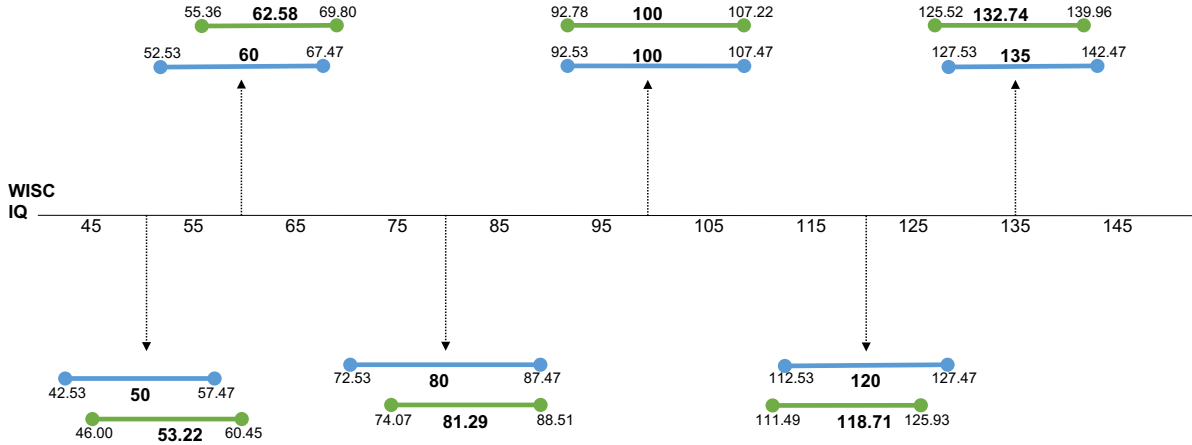


Figure 1.5: Comparison of confidence intervals based on options 1 (blue) and 2 (green) for different IQ values on the WISC-III scale.

Technical intermezzo 3: The two formulae for estimating error of measurement

Both formulae for estimating error of measurement contain two ingredients:

1. the reliability coefficient (reliability)
2. the standard deviation of the IQ score distribution (SD)

The standard deviation (ingredient 2) shows the variation in IQ scores for the children in the norm population. The reliability coefficient (ingredient 1) is used to estimate which part of the differences in the norm population children's IQ scores (ingredient 2) is caused by 'true' differences in IQ¹⁴. As regards the WISC-III^{NL}, for example, the reliability coefficient (ingredient 1) is estimated at about 0.94. The expectation therefore is that 94% of the IQ scores variance can be attributed to 'true' differences in IQ. The other 6% indicates the part played by error of measurement in the norm population. The WISC-III^{NL} assumes that if 6% of the differences in IQ in the norm population are caused by errors of measurement, the same percentage also holds for any specific child that has been tested. In short, there are two formulae for estimating error of measurement (COTAN, 2010; McManus, 2012):

$$\text{Standard error of measurement} = \text{SD} * \sqrt{(1 - \text{reliability})} \quad (1.2)$$

The standard error of measurement WISC-III^{NL} is 3.81.

$$\text{Standard error of estimation} = \text{SD} * \sqrt{(1 - \text{reliability})} * \sqrt{\text{reliability}} \quad (1.3)$$

The standard error of estimation in WISC-III^{NL} \approx 3.68.

Apart from estimating error of measurement, formula (1.1) also contains an estimate of the child's 'true' IQ score. The IQ score resulting from the test can be used for this estimate, but another option is to use Kelly's formula to correct for 'regression to the mean' (Kelley, 1947). This can be explained as follows. If a child gets a relatively high or relatively low IQ score the first time it is tested, then statistically the chance is high that with (hypothetical) subsequent intelligence tests a less extreme IQ score is attained. In other words, that the scores will be closer to the norm sample. This means that there is a risk of overestimating the 'true' IQ score for children with relatively high IQ scores, and similarly of underestimating the 'true' IQ score with children attaining a relatively low IQ score (Charter & Feldt, 2001). In these cases, then, the 'true' IQ score will on average approximate the norm sample mean of 100 to a greater degree than the observed IQ score implies (Barnett, Van Der Pols & Dobson, 2005). Kelley's formula counters this effect by correcting the observed IQ score with respect to the norm sample mean.

Generally speaking, intelligence tests either use the standard error of measurement and the IQ score resulting from the test to compute the confidence interval (option 1), or they use the standard error of estimation in combination with Kelly's formula (option 2). The WISC-III^{NL} uses option 2. The two options can be illustrated, using WISC-III^{NL} data, with the fictitious cases of Joshua, who attained an IQ score of 100, and Thomas, who attained an IQ score of 70.

In option 1 (standard error of measurement + the attained IQ score), Joshua's confidence interval is:

$$100 - (3.81 * 1.96) = 92.53$$

$$100 + (3.81 * 1.96) = 107.47$$

Rounding these values results in the following confidence interval for Joshua: 93–107.

Thomas's confidence interval in option 1 is:

$$70 - (3.81 * 1.96) = 62.53$$

$$70 + (3.81 * 1.96) = 77.47$$

Thus, the confidence interval for Thomas is: 63–77.

In option 2 (standard error of estimation + Kelly's formula to correct for regression to the mean), the attained IQ score is first multiplied with the reliability coefficient (see technical intermezzo 3) according to Kelly's formula.

In the case of Joshua, this is:

$$100(\text{attained IQ score}) * 0.935484(\text{reliability coefficient}) = 93.5484$$

And with Thomas:

$$70(\text{attained IQ-score}) * 0.935484(\text{reliability coefficient}) = 65.48388$$

To these values, the value of '1 minus the reliability coefficient multiplied by 100' is added up.

In the case of Joshua, this results in:

$$(1 - 0.935484) * 100 = 6.4516$$

$$93.5484 + 6.4516 = 100$$

The norm population mean (100) is the same as Joshua's attained IQ score (100), which means that with Kelly's formula Joshua's estimated IQ score is also 100.

With Thomas, however, correcting with Kelly's formula does lead to a difference:

$$(1 - 0.935484) * 100 = 6.4516$$

$$65.48388 + 6.4516 = 71.93548$$

The estimate of Thomas's IQ score (71.94) based on Kelly's formula thus leads to a score almost 2 points higher than his attained IQ score (70). With the estimated IQ scores of Joshua and Thomas and the standard error of measurement (3.685048), we can now calculate the confidence intervals according to option 2 (corresponding to the confidence intervals in table D.4 of the WISC-III^{NL} manual):

Joshua

$$100 - (3.685048 * 1.96) = 92.77731 \approx 93$$

$$100 + (3.685048 * 1.96) = 107.2227 \approx 107$$

Thomas

$$71.93548 - (3.685048 * 1.96) = 64.71279 \approx 65$$

$$71.93548 + (3.685048 * 1.96) = 79.15817 \approx 79$$

Note that using option 1 or 2 for calculating the confidence interval makes no difference for Joshua, but that it does for Thomas. Generally speaking, the further the value of the attained IQ score is removed from 100, the greater the differences are between options 1 and 2 (Figure 1.5).¹⁵

Interpreting WISC-III^{NL} confidence intervals

The confidence intervals of option 1 (standard error of measurement + the attained IQ score) and option 2 (standard error of estimation + Kelly's formula to correct for regression to the mean) are both based on Classical Test Theory (CTT; Guilford, 1936; Lord & Novick, 1986). Here, the CTT based starting point is the assumption that every child has one 'true' IQ score. The attained IQ score however can deviate to greater or lesser extent from this 'true' IQ score. CTT considers this deviation as measurement error. In real-life test situations we can never know a child's 'true' IQ score and thus we cannot establish the extent of measurement error for a single IQ test outcome. The solution in CTT is in the assumption that if we were able to test a child an infinite number of times, the resulting IQ scores would follow a normal distribution (Wang & Osterlind, 2013). The peak in this normal distribution then equals the 'true' IQ score of the child, meaning that on average the attained IQ score is a correct indication of the 'true' IQ score. Since the normal distribution is symmetrical, it is assumed that underestimation and overestimation of the IQ score occur with equal frequency. The confidence intervals of both options 1 and 2 should be interpreted in light of these CTT assumptions.

Option 1

Let's assume that a child takes a large number of intelligence tests. We expect then that the child attains IQ scores that approximates its 'true' IQ score — its 'peak'- relatively often. This implies that if we determine a confidence interval for these attained IQ scores, they will relatively often contain the child's 'true' IQ score. How often 'relatively often' is, depends on the bandwidth of the confidence interval. With option 1, the idea is to determine the width of the confidence interval in such a way that, if a child took an intelligence test for an infinite number of times (hypothetically), and we calculated the corresponding 95% confidence intervals, these would contain the child's 'true' IQ score in 95% of the cases. Specifically, this width is determined with the standard error of measurement (as discussed above). Figure 1.6 illustrates the 95% confidence intervals of 100 (hypothetical) intelligence tests taken by Julia. As you can see, about 95% of these confidence intervals contain Julia's 'true' IQ score of 105 (the blue intervals).

Summing up, the definition of option 1 is: "If an intelligence test is taken a large number of times, we expect that 95% of the corresponding confidence scores contain the child's 'true' IQ score. Prior to an intelligence test, there is therefore a 95% chance that the resulting 95% confidence interval actually contains the child's 'true' IQ score."

Option 2

Let's assume now that an intelligence test is not taken an infinite number of times by a single child, but just once by an infinite number of children. In that case the children can be clustered according to their attained IQ score. Figure 1.7 illustrates this for children with scores of X, Y, and Z. The idea with option 2 is that the confidence intervals are determined by asking, for all children with the same attained IQ score, what their 'true' IQ scores could be. Because the attained IQ score can deviate to greater or lesser extent from the 'true' IQ scores of every child within a cluster, we do not expect one single 'true' IQ score, but a distribution such as the three normal distributions in Figure 1.7. Based on CTT we can estimate the mean and the standard deviation of this distribution for a specific attained IQ score (see Charter & Feldt, 2001). In section 2 above, this mean is explained as 'Kelly's formula' and the standard deviation as 'standard error of estimation'. Based on the estimated distribution of 'true' IQ scores for one particular IQ score, an area can be determined that would contain 95% of these children's 'true' IQ scores. Figure 1.7 shows this as the three blue arced areas in each of the normal distributions for the scores X, Y, and Z.

Summing up, the definition of option 2 is: "For children with an IQ score equal to that of the child who was tested, we expect that 95% of these children actually has an IQ score that falls within the bandwidth of the confidence interval" (Charter & Feldt, 2001).

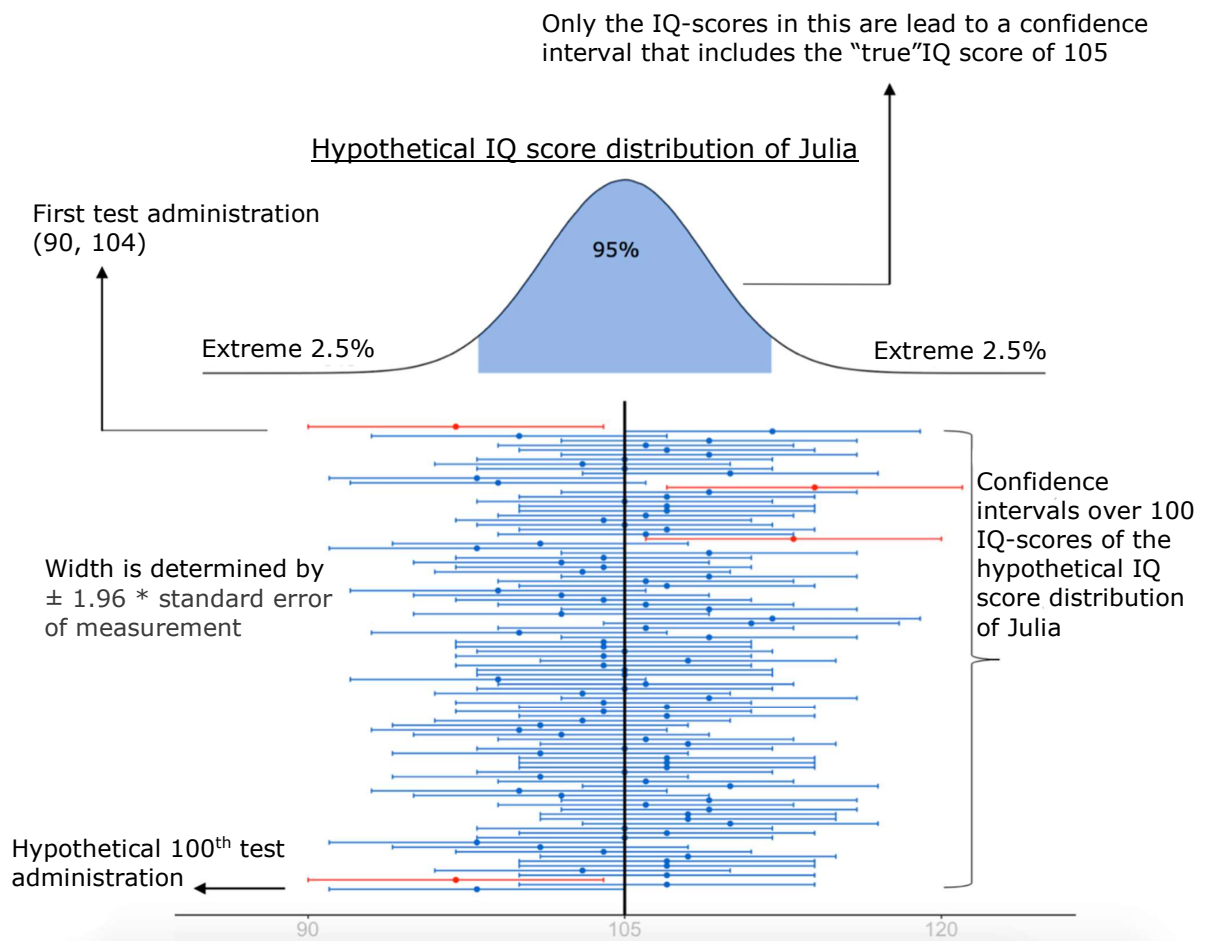


Figure 1.6: 100 hypothetical option 1 confidence intervals for Julia. The black vertical line indicates Julia's 'true' IQ score (105). The blue confidence intervals contain this 'true' IQ score, the red don't.

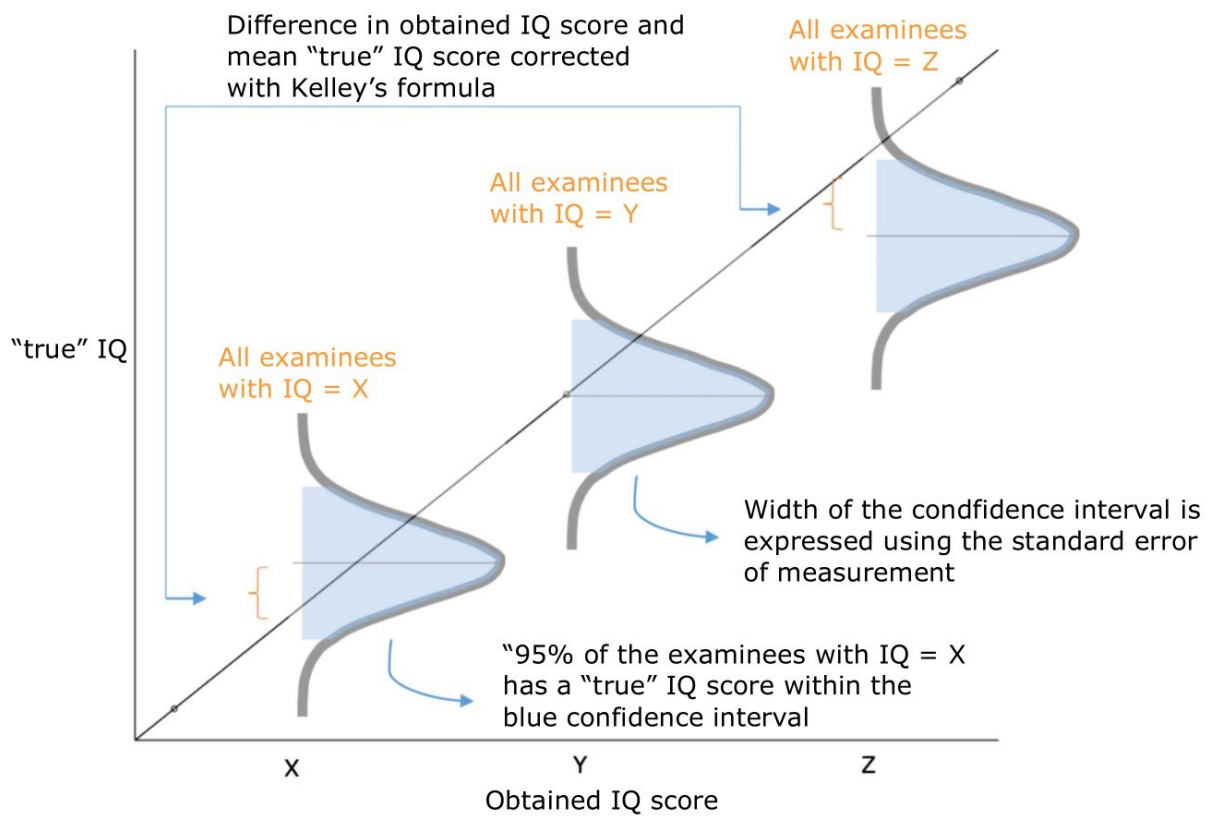


Figure 1.7: Example of the interpretation of option 2 confidence intervals (as in the WISC-III)

Assumptions

It is important to bear in mind – with both option 1 and 2 – that the interpretation of the confidence intervals as described here can only be true if all the assumptions are correct. Options 1 and 2 for example both assume that each child is measured with an equal measure of certainty or uncertainty. In fact, however, there are numerous reasons for calling this a rather unrealistic assumption (Sijtsma, 2009; Molenaar, 2004). When the test items, for example, are (much) too hard or (much) too easy for a child, this complicates estimating its intelligence correctly by using this test. The test will in this case cause a relatively larger number of errors in measurement, compared to using the same test with children whose level of intelligence is a better match for the test items' degree of difficulty. There can also be other reasons why there will be more measurement (un)certainty from one child to the next. These can include factors such as concentration, performance anxiety, multilingualism, tester (experienced or not) or observer-expectancy effects. In addition, with option 2 it is assumed that there are no fundamental differences in the average (sub)test scores between girls and boys, or between children with a western and non-western background. If there are such differences, they are not considered, which implies that Kelly's formula actually gives a wrong impression of the expected 'true' IQ score. This is called Kelly's paradox (see Wainer, 2000 and the explanation in Borsboom, Romeijn & Wicherts, 2008). In other words, it is good idea to bear in mind that for a specific child or group of children the errors of measurement could actually be larger and that the confidence intervals represent an underestimation of the uncertainty of measurement. Or, as one of our reviewers put it, it is important to be aware of the test data's 'softness'. So, the fact that it is possible to calculate the confidence interval with a great degree of precision (see section 2, above) does not mean that IQ scores can be wholly excluded with the same exactitude (cf. 'false precision'—Huff, 2010).

Some frequent misinterpretations of the confidence interval

In November 2014, 293 NIP psychologists participated in our survey. Our questions included issues such as how the participant interprets 95% confidence intervals and whether they mentioned confidence intervals in their intelligence test reports. Definitions of the WISC-III^{NL} confidence interval that resemble the definition of option 1 (see above) prevailed with 65% of the NIP psychologists. However, this is not a valid definition of the WISC-III^{NL} confidence interval, which, as we mentioned above, corresponds to option 2. Apart from that, interpreting option 1 is tricky, as some of the survey definitions indicate. For example, 14 participants gave this definition: "If Julia took the IQ test 100 times, this would yield an IQ between 90 and 104 in 95% of the cases." This definition deviates in two significant ways from the option 1 definition of confidence intervals given above. Firstly, this definition does not distinguish between attained and 'true' IQ. Secondly, by referring to the specific 90-104 interval, it does not address the general concept of confidence intervals. Compare Figure 1.6, which illustrates the definition of the option 1 confidence interval, with Figure 1.8, which illustrates a misinterpretation of the confidence interval. In both figures the upper confidence interval (90-104) is the one that belongs to the first time the test was taken; the subsequent 99 confidence intervals belong to

the hypothetical subsequent tests, with a ‘true’ IQ score of 105. All confidence intervals in Figure 1.6 that do not contain the ‘true’ IQ score of 105, including the present test, are inked in red. This is the case for about 5% of the confidence intervals. We say ‘about’ because the 100 confidence intervals result from the random IQ scores in Julia’s hypothetical IQ distribution. Should we randomly select another set of 100 IQ scores, then by chance 3, or 4, 5, 6 (etc.) confidence intervals may not contain the ‘true’ IQ score. Figure 1.8 is copy of figure 1.6, with one difference: now those confidence intervals are inked in red whose attained IQ scores are not between 90 and 104, in agreement with the survey definition “if Julia took the IQ test 100 times, this would yield an IQ between 90 and 104 in 95% of the cases.” Since 90-104 by chance belongs to the 5% that does not contain Julia’s ‘true’ IQ score (see figure 1.6), the percentage of intervals that are inked in red is much higher than 5% in Figure 1.8: 61%, to be exact. So the correct definition of option 1 and the survey’s definition only agree if the attained IQ score matches the ‘true’ IQ score exactly. Another definition frequently given (41% of the answers) was: “There is a 95% chance that Julia’s ‘true’ IQ score ranges between 90 and 104”. This definition however relies on an inaccurate use of the concept ‘chance’. Once an intelligence test is completed, the chance that the ‘true’ IQ score falls within the range of any confidence interval is either 1 or 0. Strictly speaking then, such a 95% chance is true only before the test is taken, not after its completion.

Practical implications

The most important practical difference between both types of confidence intervals is that in the case of ‘option 2’ the confidence limits – e.g. 90 and 104 – can be interpreted directly. With option 2 we can therefore say that there is a 95% chance that Julia’s true IQ falls somewhere within the limits of the confidence interval. This is because 95% of all children with similar attained IQ scores have a ‘true’ IQ score within the range of the confidence interval’s limits (if all assumptions are met). In the case of option 1, in contrast, the confidence limits cannot be interpreted directly. Here, the interval’s bandwidth is determined by the long term chance (95%) that the confidence interval actually contains the child’s ‘true’ IQ score: the confidence limits are the artefact of this long-term chance. In other words, we expect that the child’s ‘true’ IQ falls within the range of the present confidence interval limits (since we have a 95% chance prior to the test that this is the case), but we cannot be sure. So, how much emphasis can be put on the confidence interval’s limits depends on the ‘type’ of confidence interval (option 1 or 2). There are several ways of finding out whether a test report uses the option 1 or option 2 type of confidence intervals. Sometimes it is mentioned in the test manual (look for the terms ‘standard error of measurement’, ‘standard error of estimation’, or ‘Kelly’s formula’ and check whether the confidence interval is given as a formula – see technical intermezzo 3). If this is not the case, have a good look at the confidence interval of the highest and lowest possible IQ scores. If the (rounded) middle number of this confidence interval is higher (lowest possible score) or lower (highest possible score) than the attained IQ score, this (almost certainly) indicates the use of Kelly’s formula. A third option is to calculate the confidence intervals for options 1 and 2 yourself, and then to compare them to the intervals reported in the manual (accounting for small rounding differences). With our app this can be easily done for the option 1 and 2 confidence intervals; all you have to do is enter the values for the ingredients (see my OSF page <https://osf.io/qrg4e/>).

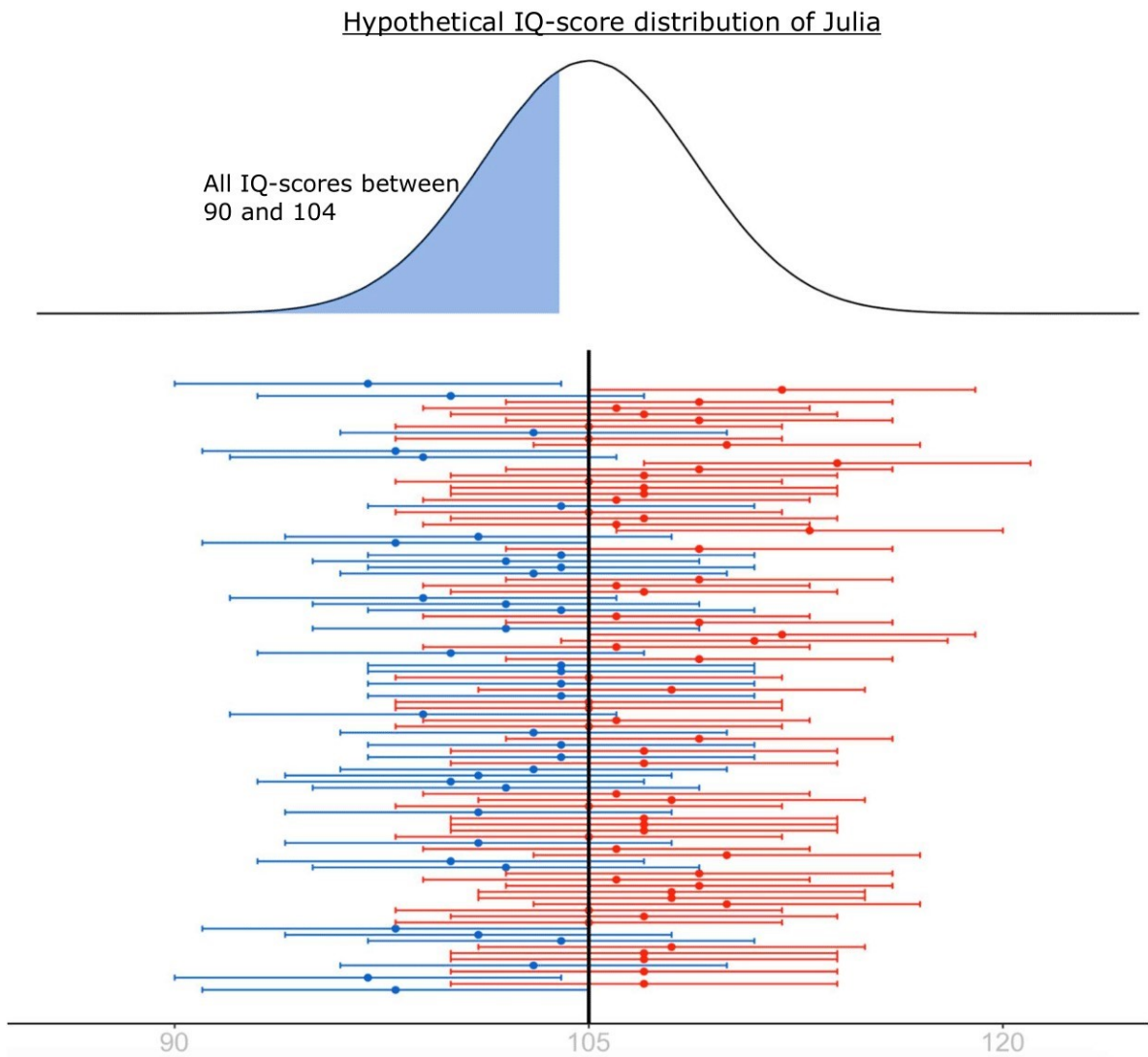


Figure 1.8: Example of the (incorrect) interpretation of "if Julia took the IQ test 100 times, this would yield an IQ between 90 and 104 in 95 out of the 100 cases"

Take home message

The stability of IQ scores is quite relative and often they are more unreliable and unstable than is frequently assumed. It is important therefore to consider more than the IQ score alone. This is to some extent also true for the classification of IQ scores (e.g. see Resing & Blok, 2002), because here a term like ‘below average’ is assigned on the basis of acquired IQ scores (Tellegen, 2004). If interpreted correctly and with due consideration of assumptions and limitations, the option 1 and 2 confidence intervals such as described in this article can help to put the IQ score in perspective.

Acknowledgements

We would like to thank dr. Helen Bakker, dr. Esmee Verhulp, Noëlle Pameijer, dr. Jelte Wicherts and prof. dr. Rob Meijer for their valuable comments on an earlier version of this article. We would also like to thank the NIP section board psychologists in education as well as the participants in our 2014 survey. The first author was supported by an NWO talent scholarship (406-15-062); the last by an NWO-VIDI grant (452-14-006).

Chapter 2



**A COMPARISON OF THE SINGLE,
CONDITIONAL AND PERSON-
SPECIFIC STANDARD ERROR OF
MEASUREMENT: WHAT DO THEY
MEASURE AND WHEN TO USE
THEM?**

Kimberley Lek & Rens Van De Schoot

Chapter 2

A comparison of the single, conditional and person-specific Standard Error of Measurement: what do they measure and when to use them?

Abstract

Tests based on the Classical Test Theory often use the Standard Error of Measurement (SEm) as an expression of (un)certainty in test results. Although by convention a single SEm is calculated for all examinees, it is also possible to (1) estimate a person-specific SEm for every examinee separately or (2) a conditional SEm for groups of comparable examinees. The choice for either of these SEms depends on their underlying assumptions and the trade-off between their unbiasedness and estimation variance. These underlying assumptions are discussed in the present article, together with a mathematical expression of the bias and estimation variance of each of the SEms. Using a simulation study, we furthermore show how characteristics of the test situation (i.e., test length, number of items, number of parallel test parts, overall reliability, relationship between ‘true’ score and true (un)certainty in test results and rounding/truncation) influence the SEm-estimates and impact our choice for one of the SEms. Following the results of the simulation study, especially rounding appears to hugely affect the person-specific and – to a lesser extent – the conditional SEm. Therefore, when a test is small and an examinee is only tested once or a few times, it is safer to opt for a single SEm. Overall, a conditional SEm based on coarse grouping appears to be a suitable compromise between a stable, but strict estimate (like the single SEm) and a lenient, but highly variable estimate (like the person-specific SEm). More practical recommendations can be found at the end of the article.

Keywords: Standard Error of Measurement, Classical Test Theory, intra-individual variation, conditional Standard Error of Measurement

Introduction

Within the Classical Test Theory (CTT), the Standard Error of Measurement (SEm; Lord & Novick, 1968) is often used as an expression of (un)certainty of test results in education, psychological assessment and health related research. Tests based on CTT usually report the traditional ‘single’ SEm, which briefly works as follows. The test is administered once to a group of persons from the desired population (i.e., the ‘norm population’) and one CTT-based SEm is estimated for all future examinees, hence the term ‘single’ (Spearman, 1904; Guilford, 1936, Gulliksen, 1950; Lord & Novick, 1968). Another possibility is the estimation of a person-specific SEm, which is calculated – as the name implies – for every examinee separately. This person-specific SEm is usually seen as an intermediate step in the estimation of the conditional SEm (see Feldt & Qualls, 1996), but as we show it can also be used on its own. When the expected value is taken of the person-specific SEm over all examinees with the same (expected) ‘true’ test score, it yields a SEm for every group of examinees separately called the conditional sem (see, for instance, Feldt & Qualls, 1996). The choice for either of the three SEMs depends on the assumptions that each of these SEMs make about measurement (un)certainty. The single SEm, for instance, can only be used when two strict assumptions are met (Molenaar, 2004). First, all persons must be measured with the same accuracy; i.e., the measurement error variance must be constant. Second, the person’s intra-individual variation in measurements must equal the inter-individual variation in measurement of the norm-population examinees; i.e., intra-individual variation and inter-individual variation must be interchangeable. Multiple authors have questioned the reasonableness of these assumptions. Sijtsma (2009), for instance, writes: “... there is no reason whatsoever to assume that the propensity distributions of different persons must be identical to one another and to the between-person distribution based on single-administration data” (p. 118). With ‘propensity distribution’ Sijtsma (2009) refers to the hypothetical distribution of repeated measurements of single examinees. When picturing a class of – in many aspects – different examinees, it is hard to imagine that the test result of each of these examinees is governed by the same underlying propensity distribution. Furthermore, it seems unlikely that these propensity distributions equal an aggregated, between-person distribution calculated on the basis of a single test take. Molenaar (2004) concludes: “[CTT] gives rise to serious questions regarding its applicability to individual assessment” (p. 203). The two assumptions are relaxed when we move from the single to the conditional SEm. Instead of constant measurement error variance, the conditional SEm is based on the more lenient assumption of homogeneous measurement error variance (see Novick, 1966). This assumption states that measurement error variance is constant conditional on ‘true’ test score. Thus, measurement error variance is only assumed constant for examinees with the same, underlying ‘true’ test score. The same holds for the second assumption; only within a group of examinees with the same ‘true’ test score, intra-individual variation and inter-individual variation in measurement are assumed interchangeable. Finally, a move from the conditional to the person-specific SEm leads to a drop of both assumptions altogether. Based on the assumptions above, one might be inclined to opt for a conditional SEm or even for a person-specific one. However, there is no free lunch. The increasingly lenient assumptions with respect to measurement (un)certainty are accompanied by increasingly stricter assumptions with respect to the structure of a test and its items (we illustrate this in the next sections). Furthermore, there is a

trade-off between the unbiasedness of SEM-estimates and their estimation variability. For instance, if examinees have their own unique measurement (un)certainty, the person-specific SEM will provide us with an unbiased estimate at the cost of being the most variable estimate of SEM.

Conventionally, CTT test makers and test users have ‘blindly’ adopted the single SEM for individual inferences. We propose to not base this decision on convention, but on a deliberate consideration of characteristics of the test, its norm population and assumptions one is willing to make and their influence on the trade-off between unbiasedness and estimation variance. To aid this deliberate consideration, this chapter first discusses the Classical Test Theory and how the single SEM, the person-specific SEM and the conditional SEM fit in this framework. In this discussion, differences in the estimation of the three SEMs are shown and underlying assumptions are highlighted. Then, we zoom in on the trade-off between unbiasedness and estimation variance of the three SEMs, which is followed by a simulation study. In this simulation study, we investigate how characteristics of the test and its norm population (i.e., number of repeated test takes, number of items, overall reliability, divisibility in parallel test parts, anticipated relationship between true test score and SEM and whether the scores are continuous or rounded and truncated) influence the trade-off between unbiasedness and estimation variance. Based on this simulation study, we end with a set of practical recommendations when to use each of the SEMs.

Classical Test Theory

Throughout this chapter, we built on the ideas and notation of Classical Test Theory¹⁶ (CTT; Spearman, 1904; Guilford, 1936; Gulliksen, 1950; Lord & Novick, 1968). CTT comprises a set of assumptions and conceptualizations that make it possible to model measurement error (Gulliksen, 1950; DeVellis, 2006). In this section, we discuss the fundamentals of CTT. In the next section, we discuss how respectively the single Standard Error of Measurement (SEM), the person-specific SEM and the conditional SEM fit within this CTT framework. In essence, CTT assumes that every person has a ‘true’ overall test score. This ‘true’ overall test score is assumed to be time invariant. The difference between the fixed ‘true’ score and the score obtained on a measurement is considered random measurement error:

$$e_{gi} = \tau_i - x_{gi} \tag{2.1}$$

In Equation (2.1), e_{gi} denotes person’s i measurement error on measurement occasion g . τ_i refers to the person’s ‘true’, overall test score and x_{gi} to person’s i obtained score on measurement occasion g . For this equation to hold, CTT assumes that person i is part of a well-defined population of persons \wp ($i \in \wp$) and that the specific test, let’s call this test a , is part of a well-defined set of possible tests \mathfrak{R} ($a \in \mathfrak{R}$). Assume we can repeat the measurement of person i with test a many times independently. In that case, the random variable E_{*i} would contain all measurement errors e_{gi} of person i over G repeated measurements and the random variable X_{*i} would contain all observed scores x_{gi} of person i over G repeated measurements. Note the replacement of g with an asterisk to denote random measurements. τ_i in Equation (2.1) can then be conceptualized as the expected value of X_{*i} :

$$\tau_i = E_i[X_{*i}] \quad (2.2)$$

Since the expected value of x_{gi} equals τ_i , the expected value of E_{*i} is zero:

$$E_i[E_{*i}] = 0, \quad (2.3)$$

where E_{*i} is expected to be normally distributed with finite variance $\sigma^2(E_{*i})$:¹⁷

$$E_{*i} \sim N\left(0, \sigma^2(E_{*i})\right), \quad (2.4)$$

where $\sigma^2(E_{*i})$ here refers to intra-individual variation. Although equation (2.1)-(2.4) are useful for the conceptualization of the conditional and the person-specific Standard Error of Measurement later, they are not generally the focus of CTT (Novick, 1966). Rather, CTT shifts the focus from the repeated measurement of a single, specific examinee to the single measurement of repeatedly, randomly selected examinees (Lord & Novick, 1968). Given that Equation (2.1) holds for a specific examinee i , it also holds for any randomly selected examinee. Therefore, if we denote the error random variable over randomly selected examinees by E_{g*} , the true score random variable by T_* and the observed score random variable by X_{g*} ,

$$E_{g*} = T_* - X_{g*} \quad (2.5)$$

in \wp or any subpopulation of \wp (note the substitution of i by $*$ to denote randomly selected examinees). Again, E_{g*} is expected to be normally distributed with finite variance $\sigma^2(E_{g*})$:

$$E_{g*} \sim N\left(0, \sigma^2(E_{g*})\right), \quad (2.6)$$

where $\sigma^2(E_{g*})$ refers to inter-individual variation. As we explain in the coming sections, $\sigma^2(E_{g*})$ in Equation (2.6) forms the basis for the single SEM while $\sigma^2(E_{*i})$ from Equation (2.4) is crucial to the conditional and person-specific SEM.

Single Standard Error of Measurement

Definition

The single Standard Error of Measurement, abbreviated here as single SEM, is defined as $\sqrt{\left(\sigma^2(E_{g*})\right)}$; the square root of the variance in measurement errors of randomly selected examinees (see Equation (2.6)). Because of the linear relationship in Equation (2.5) and zero correlation between E_{g*} and T_* , the variance in measurement errors of randomly selected examinees is simply the difference between the total variation in observed scores and true-score variance (for proof see Novick, 1966):

$$\sigma^2(E_{g*}) = \sigma^2(X_{g*}) - \sigma^2(T_*) \quad (2.7)$$

Estimation

As mentioned in the introduction, the single SEM is estimated based on a norm population. Since this norm population is only tested once, the true-score variance $\sigma^2(T_*)$ is unknown and the single SEM cannot be estimated directly using Equation (2.7). In order to estimate $\sigma^2(E_{g*})$, and eventually $\sqrt{\sigma^2(E_{g*})}$, the unobservable quantity $\sigma^2(T_*)$ must be rewritten in a potentially observable quantity. CTT does this by splitting the single measurement of the norm population into two parallel parts. ‘Parallel’ means that the items and/or scores in the parts have the same underlying true score variable T_* and experimentally independent errors with equal variance $\sigma^2(E_{g*})$ in \wp and every subpopulation of \wp (see definition 2.13.4, Lord & Novick, 1968). ‘Experimentally independent’ (definition 2.10.1; Lord & Novick, 1968) implies that the correlation between errors of the parts equals zero. From a practical point of view, parallel test parts can be constructed by matching item content and item characteristics over parts. Examples of such item characteristics are item difficulty – the proportion of persons of the norm population answering the item correctly - and the reliability index, which is the point-biserial correlation between item and total score multiplied by the item standard deviation. Thus, even though items in a test are likely to differ in terms of difficulty and reliability, the idea is to group them in such a way that the groups of items (the parallel test parts) are comparable. An example of an algorithm capable of grouping items in an optimal way can be found in Sanders and Verschoor (1996). In the remainder of this article, we use K to denote the number of parallel test parts and k to refer to a specific parallel part. Multiple coefficients exist based on the correlation between K parallel measurements, such as the Spearman-Brown split half coefficient (Spearman, 1910; Brown, 1910). These coefficients, here denoted by $\rho_{XX'}$, are meant to estimate the reliability of the test, i.e. the ratio $\sigma^2(T_*)/\sigma^2(X_{g*})$ (see Lord & Novick, 1968) adjusting for the change in text length and hence the single SEM can be estimated using:

$$\sigma^2(E_{g*}) = \sigma^2(X_{g*}) - \sigma^2(T_*) = \sigma^2(X_{g*}) - \sigma^2(X_{g*}) \frac{\sigma^2(T_*)}{\sigma^2(X_{g*})} = \sigma^2(X_{g*})(1 - \rho_{XX'}) \quad (2.8)$$

In the remainder, we will use the Spearman-Brown split half coefficient (Equation (2.8)). Note, however, that alternatives exist for the estimation of reliability that are based on more lenient assumptions of parallelism. Test halves can, for instance, be tau-equivalent, meaning that the true scores in both test parts are allowed to differ by a constant (equal for all examinees) and that their error variances are allowed to differ as well (Brennan, 2010; Traub, 1997; also see Haertel, 2006 for an overview of reliability coefficients and their assumptions of parallelism). From a practical point of view, items in tau-equivalent test parts do not have to be matched as strictly on item characteristics such as item difficulty as is the case for parallel test parts.

Person-specific Standard Error of Measurement

Definition

Whereas the single SEM is based on the variance in measurement errors of randomly selected examinees $\sigma^2(E_{g*})$, see Equation (2.6), the person-specific SEM can be expressed as $\sqrt{\sigma^2(E_{*i})}$; the square root of the variance in person i 's measurement errors over repeated measurements (see Equation (2.4)). As shown by Lord and Novick (1968), the single SEM and the person-specific SEM are related to one another:

$$\sigma^2(E_{g*}) = E\left[\sigma^2(E_{*i})\right], \quad (2.9)$$

indicating that the between person variance on which the single SEM is based, is equal to the expected value of the within person error variance on which the person-specific error variance is based (Lord & Novick, 1968, p. 35; for a graphical illustration see our Figure 2.1). In other words, the single SEM $\sqrt{\sigma^2(E_{g*})}$ is the mean Standard Error of Measurement in \wp (Lord & Novick, 1968). Estimating $\sqrt{\sigma^2(E_{*i})}$ for every person separately thus leads to a correct estimate of $\sqrt{\sigma^2(E_{g*})}$ on average while allowing the possibility that some persons' measurements are inherently more consistent than others. Equation (2.9) shows that it is possible to aggregate $\sigma^2(E_{*i})$ on the level of specific examinees to reach an estimate of $\sigma^2(E_{g*})$ on a population level, but not the other way around. Thus, if we want to use $\sigma^2(E_{g*})$, on which the single SEM is based, to make inferences on an individual level, we need to assume that (1) every examinee has the same error variance $\sigma^2(E_{*i})$ and that this error variance equals $\sigma^2(E_{g*})$; the assumptions of constant measurement error and interchangeability of intra-individual and inter-individual variation discussed previously.

Estimation

Just as with the single SEM, the estimation of the person-specific SEM is based on splitting the test in parallel test parts. When a test is administered multiple times, the person-specific SEM can be estimated based on the variances within parallel test parts, $\sigma^2(X_{*ki})$, operationalized as $\frac{\sum_{g=1}^G (x_{gki} - \bar{x}_{*ki})^2}{G-1}$ where \bar{x}_{*ki} denotes the mean of the k th parallel test part. Since the overall test score x_{gi} is the sum over (let's for simplicity assume unit-weighted) scores for each of the parallel parts, the error variance of x_{gi} can be expressed as (Lord & Novick, 1968):

$$\sigma^2(E_{*i}) = \sigma^2(X_{*i}) = \sum_{k=1}^K \sigma^2(X_{*ki}) + 2 \sum_{k < p} \sigma^2(X_{*ki}, X_{*pi}) \quad (2.10a)$$

$$\sigma^2(E_{*i}) = \sigma^2(X_{*i}) = \sum_{k=1}^K \sigma^2(X_{*ki}). \quad (2.10b)$$

In other words, the variance of the composite test score x_{gi} , and thus $\sigma^2(E_{*i})$, is a function of the (unit-weighted) within variances of the parallel parts and their covariance. Since the parallel parts are experimentally independent, these covariances equal zero (Equation (2.10b), see Hu et al., 2016 for the estimation of the person-specific SEM in the context of many repeated measurements). As stressed in Sijtsma (2009), there generally is a “. . . practical impossibility to administer the same test to the same individuals repeatedly – even twice is nearly impossible” (p. 117). Therefore, Equation (2.10) must be modified to accommodate a situation with only one test take. This is done by replacing the *within* parallel parts variances, $\sigma^2(X_{*ki})$, with the variance *between* parallel parts, $\sigma^2(X_{g*i})$:

$$\sigma^2(E_{*i}) = \sigma^2(X_{*i}) = \sum_{k=1}^K \sigma^2(X_{g*i}) = K\sigma^2(X_{g*i}). \quad (2.11)$$

This replacement is justified since, if we operationalize the variance between parallel parts as $\frac{\sum_{k=1}^K (x_{gki} - \bar{x}_{g*i})^2}{K-1}$ where \bar{x}_{g*i} denotes the mean of the g th measurement, the expected value of this between variance equals the expected value of the parallel parts’ within variance, calculated across repeated measures:

$$E_i \left[\sigma^2(X_{*ki}) \right] = E_i \left[\sigma^2(X_{g*i}) \right]. \quad (2.12)$$

The proof for this is straightforward:

$$E_i \left[K\sigma^2(X_{g*i}) \right] = K E_i \left[\frac{\sum_{k=1}^K (x_{gki} - \bar{x}_{g*i})^2}{K-1} \right] = K \frac{\sum_{g=1}^G \frac{\sum_{k=1}^K (x_{gki} - \bar{x}_{g*i})^2}{K-1}}{G-1}$$

and

$$E_i \left[K\sigma^2(X_{*ki}) \right] = K E_i \left[\frac{\sum_{g=1}^G (x_{gki} - \bar{x}_{*ki})^2}{G-1} \right] = K \frac{\sum_{k=1}^K \frac{\sum_{g=1}^G (x_{gki} - \bar{x}_{*ki})^2}{G-1}}{K-1}.$$

In other words, the expected variance between and within parallel parts is the same.

For normally distributed parallel measurements, the between variance $\sigma^2(X_{g*i})$ is known to follow a scaled chi-squared distribution (see Knight, 2000)¹⁸. Therefore, the between variance $\sigma^2(X_{1*i})$ at a specific test take can also be perceived as a random sample of the scaled chi-squared distribution of a fixed, unknown $\sigma^2(E_{*i})$ of person i .

Conditional Standard Error of Measurement

Definition

Just as the person-specific SEM, the conditional SEM is based on Equation (2.4). This conditional SEM can simply be expressed as the expected value of the person-specific SEMs of examinees belonging to the same group j :



Figure 2.1: Illustration of the relationship between the single, conditional and person-specific Standard Error of Measurement. The rows symbolize examinees N and the columns parallel parts K . The different colors indicate different test scores for the K parallel parts. $E[\]$ refers to the expectation over the variances between the brackets.

$$\sigma_{i \in j}(E_{*i}) = \sqrt{E_{i \in j}[\sigma^2(E_{*i})]} \quad (2.13)$$

Ideally, every examinee in group j has the same ‘true’ test score. Since examinee’s ‘true’ scores are unknown, they are typically grouped according to their total test scores¹⁹. Building on Equation (2.9), the relationship between the conditional and the single SEM can be described as follows (for a graphical illustration see Figure 2.1):

$$\sigma^2(E_{g*}) = E\left[\sigma_{i \in j}^2(E_{*i})\right], \quad (2.14)$$

that is, when we take the expected value over all the error variances in groups j , we find the between person variance on which the single SEM is based. Estimating $\sigma_{i \in j}^2(E_{*i})$ for every group separately thus leads to a correct mean estimate of $\sigma^2(E_{g*})$ while allowing the possibility that low-scoring, average-scoring and/or high-scoring examinees have measurements that are inherently more or less consistent. Equation (2.13) shows that whereas the conditional SEM can be estimated by aggregating the person-specific SEMs, it does not work the other way around. Thus, if we want to use $\sigma_{i \in j}^2(E_{*i})$, on which the conditional SEM is based, to make inferences on an individual level, we need to assume that every examinee within group j has the same error variance; this is the homogeneous measurement error variance assumption discussed previously.

Estimation

Many procedures have been proposed in the literature to estimate the conditional SEM, including but not restricted to the binomial procedure (Brennan & Lee, 1999), the compound binomial procedure [Brennan & Lee, 1999] the Feldt-Qualls procedure (Feldt & Qualls, 1998; for a comparison of this procedure and the procedures previously mentioned see Lee, Brennan & Kolen, 2000), the item response theory procedure by Wang, Kolen, and Harris (1996) and the bootstrap procedure advocated by Colton, Gao and Kolen (1996). In the present article, we focus on the Feldt and Qualls (1996) procedure, since this procedure is very similar to the procedure for estimating the person-specific SEM discussed previously. The Feldt and Qualls (1996) procedure extends the original method proposed by Thorndike (1951), where a test is split in two parallel halves. Thorndike (1951) showed that the difference between two half-test scores equals the difference between their two errors (up to a constant, when the half tests are essentially tau-equivalent instead of classically parallel). With the assumption that the two errors are independent, the variance of the difference between the two half-test scores equals the error variance we are interested in. Feldt and Qualls (1996) generalize this finding of Thorndike (1951) to a situation with more than two parallel (or essentially tau-equivalent) parts:

$$\sigma^2(E_{*i}) = K \left[\frac{\sum_{k=1}^K (x_{gki} - \bar{x}_{g*i})^2}{K - 1} \right], \quad (2.15)$$

where K in Equation (2.15) corrects for test length. That is, since the division of tests in parallel parts shortens the test by a factor K , we multiply by K to obtain an estimate for

tests with the original test length (Feldt & Qualls, 1996). Note that Equation (2.15) is equal to Equation (2.11) discussed earlier. To obtain the conditional SEM, the expected value needs to be taken over the result in Equation (2.15) for all examinees in a certain group j . The person-specific SEM can thus also be perceived as a special case of the conditional SEM when $n_j = 1$:

$$\sigma_{i \in j}^2(E_{*i}) = E_{i \in j} \left\langle K \left[\frac{\sum_{k=1}^K (x_{gki} - \bar{x}_{g*i})^2}{K-1} \right] \right\rangle \quad (2.16)$$

When the parts are (essentially) tau-equivalent instead of classically parallel, Equation (2.15) can be replaced with:

$$\sigma^2(E_{*i}) = K \left[\frac{\sum_{k=1}^K ([x_{gki} - \bar{x}_{g*i}] - [\bar{x}_{g*i} - \bar{x}_{g**}])^2}{K-1} \right] \quad (2.17)$$

where the part $[\bar{x}_{g*i} - \bar{x}_{g**}]$ accounts for differences in means of the parallel parts (\bar{x}_{g**} denotes the ‘grand’ mean over all parallel test parts) for the person-specific SEM with $n_j = 1$. Equation (2.16) can be replaced with:

$$\sigma_{i \in j}^2(E_{*i}) = E_{i \in j} \left\langle K \left[\frac{\sum_{k=1}^K ([x_{gki} - \bar{x}_{g*i}] - [\bar{x}_{g*i} - \bar{x}_{g**}])^2}{K-1} \right] \right\rangle \quad (2.18)$$

for the conditional SEM with $n_j > 1$.

Bias-variance trade-off

Ultimately, we want the estimate of SEM to be as reflective of the examinee’s error variation as possible. Keeping in mind that the examinee is tested once, there are two factors that influence the adequacy of the estimated SEM for the examinee: (1) the (un)biasedness of the estimate and (2) its estimation variance.

Variance

Looking at the variance first, the following (in)equalities are at play:

$$\text{Rule 1.} \quad \text{variance}_{\text{single}} \leq \text{variance}_{\text{conditional}} \leq \text{variance}_{\text{person-specific}} \quad (2.19)$$

The justification for this rule is as follows. As discussed before, the variance between parallel parts of a person follows a scaled chi-squared distribution. The variance of this distribution equals:

$$\sigma^2\left(\sigma^2(X_{g*i})\right) = \frac{2(E_i[\sigma^2(X_{g*i})])^2}{K-1} \quad (2.20)$$

Since the estimate of the person-specific error variance $\sigma^2(E_{*i})$ simply follows by multiplying $\sigma^2(E_{g^*i})$ by K (see Equation (2.11)), the estimation variance of $\sigma^2(E_{*i})$ is:

$$\sigma^2\left(\sigma^2(E_{*i})\right) = K^2\sigma^2\left(\sigma^2(X_{g^*i})\right) \quad (2.21)$$

(Feldt & Qualls, 1996). By rules of error propagation, we know that the variance of n_j independent examinees equals $\frac{\sum_i \sigma^2\left(\sigma^2(E_{*i})\right)}{n_j^2}$ and thus the variance of the conditional SEM can be expressed as:

$$\sigma^2\left(\sigma_{i \in j}^2(E_{*i})\right) = \frac{\sum_i K^2\sigma^2\left(\sigma^2(X_{g^*i})\right)}{n_j^2} \quad (2.22)$$

The variance of the conditional SEM thus also depends on the size of groups J . Finally, since the single error variance $\sigma^2(E_{g^*}) = E[\sigma^2(E_{*i})]$ is usually computed for all future examinees, it is a constant and therefore its variance equals zero:

$$\sigma^2\left(\sigma^2(E_{g^*})\right) = 0. \quad (2.23)$$

Comparing Equation (2.22) and (2.23), we see that Equation (2.22) can *only* equal Equation (2.23) when the numerator of Equation (2.22) results in zero. Since K^2 cannot equal zero (K has to be at least 2), the expression $\sigma^2\left(\sigma^2(X_{g^*i})\right)$ has to become zero, which means that there is no variation in the between variances $\sigma^2(X_{g^*i})$. When this is true, also Equation (2.21) results in zero, explaining why the conditional and person-specific estimation variances are larger or equal to the estimation variance of the single SEM. On top of the situation in which $\sigma^2\left(\sigma^2(X_{g^*i})\right) = 0$, Equation (2.21) and (2.22) result in the same value when $n_j = 1$ (in this case, Equation (2.22) simplifies into Equation (2.21)). In other instances, the person-specific SEM always exceeds the estimation variance of the conditional SEM. Figure 2.2 illustrates Rule 1 graphically (see the R file ‘Figure 2.R’ on the OSF page ‘<https://osf.io/qrg4e/>’ for detailed information about this plot). The flat, mint colored surface shows the estimation variance of the single SEM and the tilted vertical surface illustrates the estimation variance of the person-specific SEM. The darkest colored surface belongs to the estimation variance of the conditional SEM. All three surfaces touch when there is no variation in test score variance over test takes (left side x-axis). The surface of the person-specific SEM and the conditional SEM also touch when n_j equals one (right side of the y-axis). Other than at these two touching points, the dark-colored surface lies above the mint-colored surface, implying that the estimation variance of the conditional SEM exceeds the estimation variance of the single SEM, whereas the surface of the person-specific SEM consistently lies above the surface of the conditional SEM.

Bias

Turning to the (un)biasedness, the following rule holds:

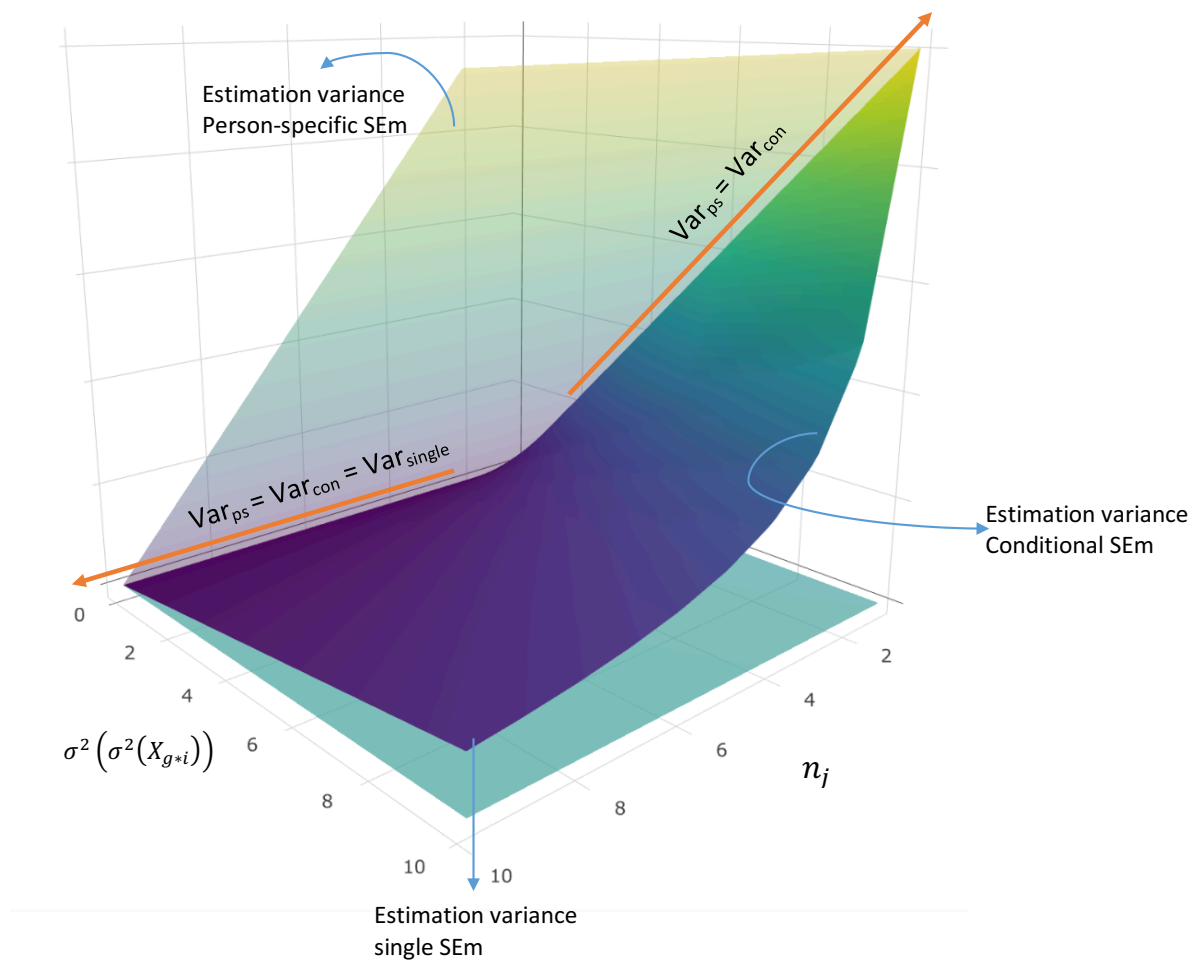


Figure 2.2: Relationship between the estimation variance of the person-specific, conditional- and single error variance, for a fictional examinee.

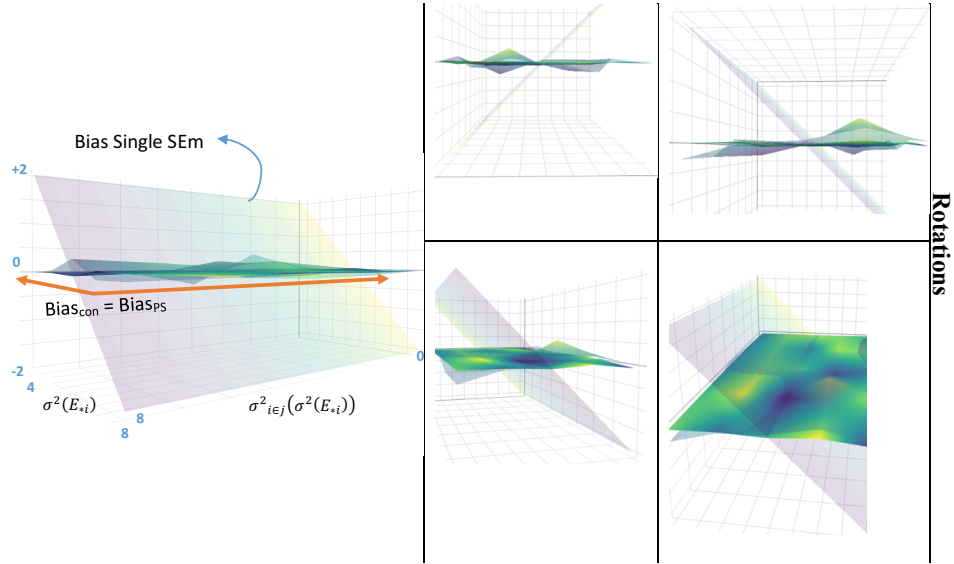
$$\text{Rule 2.} \quad \text{bias}_{\text{single}}, \text{bias}_{\text{conditional}} \geq \text{bias}_{\text{person-specific}} \quad (2.24)$$

The rationale for this rule is as follows. Using Equation (2.11) and (2.12), it is easy to show that the expected value of the estimator $\widehat{\sigma^2(E_{*i})}$ equals $\sigma^2(E_{*i})$:

$$E_i \left[\widehat{\sigma^2(E_{*i})} \right] = KE_i \left[\sigma^2(X_{g*i}) \right] = KE_i \left[\sigma^2(X_{*ki}) \right] = \sigma^2(E_{*i}) \quad (2.25)$$

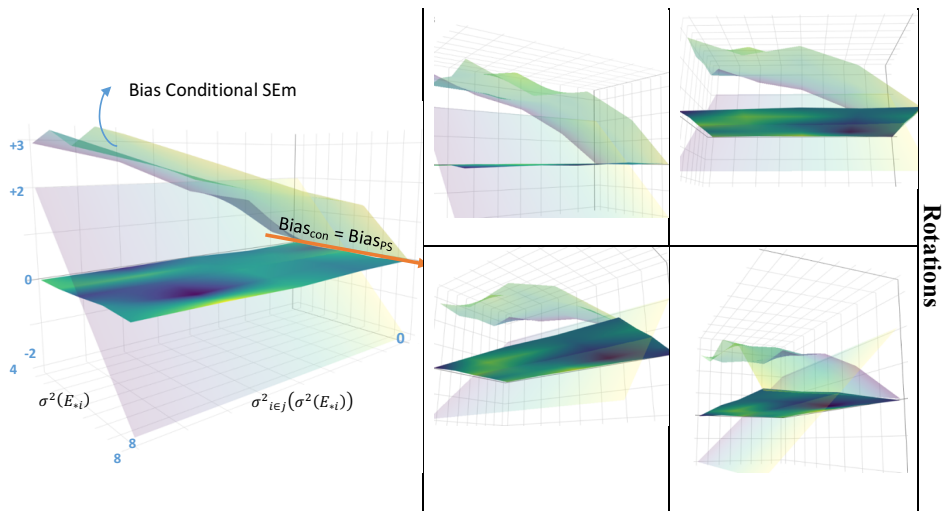
Thus, in the long run the error variance of examinee i is correctly estimated. Other estimators are either as unbiased as $\widehat{\sigma^2(E_{*i})}$ or more biased. Using Equation (2.8), we see that the expected value of the estimator $\widehat{\sigma^2(E_{g*})}$ equals $\sigma^2(E_{g*})$. As Equation (2.9) shows, $\widehat{\sigma^2(E_{g*})}$ is only an unbiased estimator for $\sigma^2(E_{*i})$ when the $\sigma^2(E_{*i})$ of examinee i equals $E \left[\sigma^2(E_{*i}) \right]$ over all examinees in φ . This is either the case when all examinees have the same error variance or when examinee i happens to have an error variance that equals the mean error variance of the population. Finally, based on Equation (2.16), the expected value of the estimator $\widehat{\sigma_{i \in j}^2(E_{*i})}$ equals $\sigma_{i \in j}^2(E_{*i})$ when the grouping is based on true scores²⁰. $\widehat{\sigma_{i \in j}^2(E_{*i})}$ is thus only an unbiased estimator for $\sigma^2(E_{*i})$ when the $\sigma^2(E_{*i})$ of examinee i equals $E_{i \in j} \left[\sigma^2(E_{*i}) \right]$ over all examinees in group j . Again, this occurs when either all examinees in group j have the same error variance or when the error variance of examinee i equals the mean error variance of the group. It is impossible to know whether $\sigma^2(E_{*i})$ of examinee i happens to be closer to the population mean $\widehat{\sigma^2(E_{g*})}$ or the group mean $\widehat{\sigma_{i \in j}^2(E_{*i})}$ and therefore the bias for the single- and conditional SEM are separated by a comma in Rule 2. Figure 2.3 illustrates Rule 2 graphically (for more detailed information on how Figure 3 was constructed, see the R file ‘Figure 3A.R’, ‘Figure 3B.R’ and ‘Figure 3C.R’ on the OSF page <https://osf.io/qrg4e/>). The x-axes vary the true error variance of fictional examinee i and the y-axes the variation in true error variances within group j ($n_j = 100$) where examinee i belongs to. Figure 2.3a illustrates the bias (estimate of error variance – true error variance) of the single-, conditional- and person-specific SEM over 10,000 fictional test takes when $\sigma^2(E_{*i})$ of examinee i equals the group mean $E_{i \in j} \left[\sigma^2(E_{*i}) \right]$, Figure 2.3b the bias when $\sigma^2(E_{*i})$ is one standard deviation removed from $E_{i \in j} \left[\sigma^2(E_{*i}) \right]$ and Figure 2.3c the bias when $\sigma^2(E_{*i})$ is two standard deviations removed from $E_{i \in j} \left[\sigma^2(E_{*i}) \right]$. The (almost) flat surface for the person-specific SEM in all three plots illustrates that the person-specific SEM is unbiased regardless of the true error variance and variation of true error variances in group j . The tilted vertical surface for the single SEM shows that this SEM is only unbiased when $\sigma^2(E_{*i})$ collides with the $\widehat{\sigma^2(E_{g*})}$. As the first plot shows, when $\sigma^2(E_{*i})$ equals $E_{i \in j} \left[\sigma^2(E_{*i}) \right]$, both the conditional and person-specific SEM are unbiased (although the surface of the conditional SEM is ‘bumpier’ due to the introduction of variation over the 100 group members). When that is not the case (second and third plot), bias quickly rises when the variance in personal error variances within group j increases.

Although Rule 1 and 2 and Figure 2.2 and 2.3 provide insight into the estimation variance and (un)biasedness of the three estimators, it is hard to choose one of the three looking at



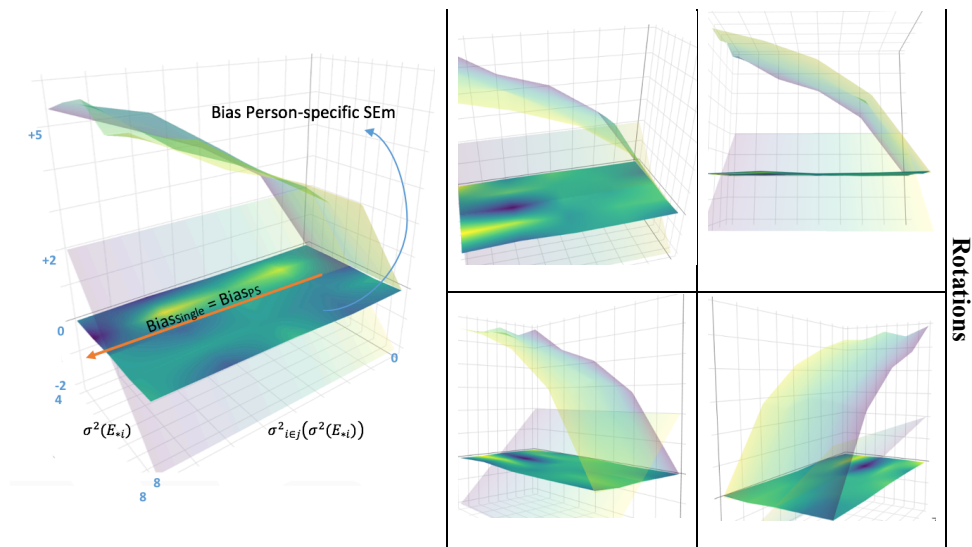
$$\sigma^2(E_{*i}) = E_{i \in j}[\sigma^2(E_{*i})]$$

(a)



$$\sigma^2(E_{*i}) = E_{i \in j}[\sigma^2(E_{*i})] - \sqrt{\sigma^2_{i \in j}(\sigma^2(E_{*i}))}$$

(b)



$$\sigma^2(E_{*i}) = E_{i \in j}[\sigma^2(E_{*i})] - 2\sqrt{\sigma^2_{i \in j}(\sigma^2(E_{*i}))}$$

(c)

Figure 2.3: Relationship between the bias of the person-specific -, conditional-, and single SEM variance, for a fictional examinee i . When the error variance of examinee i equals the average error variance in group j (a), both the person-specific SEM and the conditional SEM provide an unbiased estimate. When this is not the case (b,c), the bias of the conditional SEM increases rapidly, with increasing variance over the error variances within group j (y-axis). The single SEM (tilted vertical surface) is only unbiased when the error variance of the examinee (x-axis) coincides with the average population error variance (visible in a-c). To aid interpretation, different rotations of the a-c figure are included at the right side.

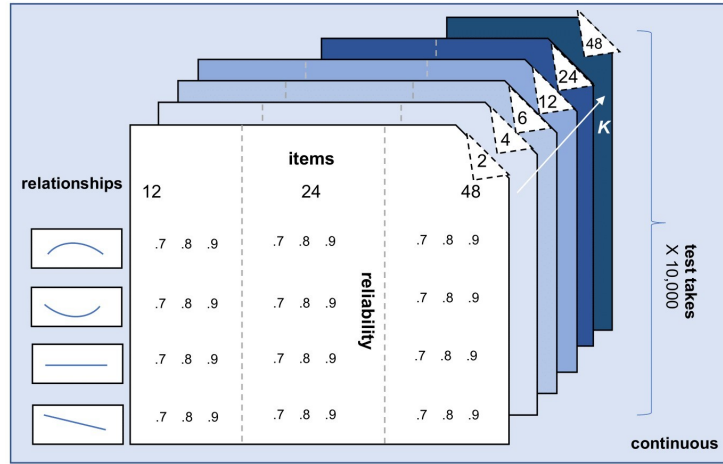
these two figures separately. Ideally, we would want to choose between the person-specific -, conditional – and single SEM based on the (un)biasedness and variance simultaneously. A measure which naturally fits the need to balance bias and estimation variance is the mean squared error (MSE), which can be expressed as the sum of the bias squared and the estimation variance. In the next sections, we show how to use the MSE to choose between the single, conditional and person-specific SEM in different testing situations.

Simulation

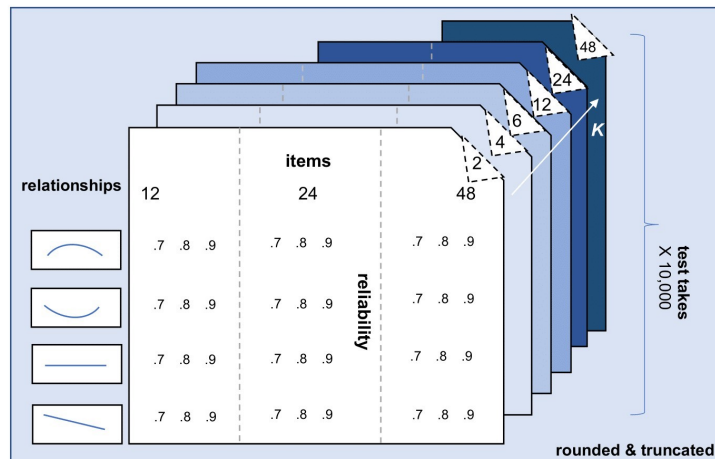
Simulation design

To compare the single, conditional and person-specific SEM, we simulated item and test scores for 1,000 examinees and 10,000 repeated test takes. Six characteristics of the test and the norm population were varied: (1) whether the parallel test scores are continuous and unrestricted or rounded to integers and truncated, (2) the number of repeated test takes, (3) the number of items of the test, (4) the number of parallel test parts K , (5) the relationship between the 1,000 ‘true’ examinee scores and their error variances and (6) the overall reliability of the test. Figure 2.4 summarizes these six characteristics. Each of these characteristics is also briefly discussed below.

- (1) **Continuous versus rounded and truncated scores.** As can be seen in Figure 2.4, all of the other five characteristics are varied for a scenario with continuous item-, parallel- and total scores and a scenario with rounded and truncated scores. In the ‘rounded and truncated’ scenario (see lower part Figure 2.4), each item can be answered correctly (item score = 1) or incorrectly (item score = 0), leading to a total score with a possible range between 0 (every item answered incorrectly) and the maximum number of items considered (every item answered correctly). In the ‘continuous’ scenario (see upper part Figure 2.4), the item- and total scores are on a similar scale as the ‘rounded and truncated’ ones, but these scores are not restricted or rounded. Note that the Classical Test Theory and the formulas for the person-specific SEM, the conditional SEM and the single SEM discussed previously are all based on this continuous scenario. The ‘rounded and truncated’ scenario, however, provides a more realistic reflection of test practice.
- (2) **Number of repeated test takes.** In this simulation study, we simulate 10,000 repeated test takes for each of the examinees. This large number makes it possible to estimate the long run bias, estimation variance and MSE. In practice, however, we do not have 10,000 test takes per examinees and therefore we will also look at fewer test takes (e.g., 1, 2, 3, 4, 5, 10, 25, 50 and 500) to see how accurate SEMs estimates are in the short run.
- (3) **Number of items.** Respectively 12, 24 and 48 items are used in this simulation study. Based on previous studies (see, for instance, Feldt & Qualls, 1996) 12 items can be perceived as a small test, whereas between 24 and 48 items can be seen as a realistic test size. To keep the simulation results comparable over the varying number of items, for the ‘rounded and truncated’ scenario the relative proportion of items correct of the 1,000 examinees for a certain test take were kept equal (i.e., if



(a)



(b)

Figure 2.4: Schematic overview of the simulation design. The characteristics number of repeated test takes (2), number of items (3), number of parallel test parts K (4), relationship between 'true' test score and error variance (5) and overall test reliability (6) are varied for both the scenario with continuous test scores (1; upper part of the figure) and rounded and truncated test scores (lower part of the figure). Each slide within 'continuous' and 'rounded and truncated' belongs to a specific number of parallel test parts K . Within each slide $\leq K = 12$, 3 (number of items) \times 4 (relationships) \times 3 (reliabilities) = 36 conditions are being simulated. The slides $K = 24$ and $K = 48$ add another 24 ($K = 24$) and 12 ($K=48$) conditions.

person i has 6 items correct on a 12-item test at test take j , he has 12 items correct on a 24-item test at the same test take, et cetera). For the ‘continuous’ scenario, the average item score was kept the same for the varying number of items.

- (4) **Number of parallel test parts.** To compare the effectiveness of various partitions, the 12, 24 and 48 items were respectively divided in parallel parts of $K = 2$, $K = 4$, $K = 6$ and $K = 12$. For the tests with items > 12 , we additionally added $K = 24$ (24- and 48-item test) and $K = 48$ (48-item test). By keeping K fixed over the different number of items, the number of items per parallel test part varies. For $K = 2$, for instance, there are respectively 6, 12 and 24 items in each of the 2 parallel test parts, depending on the number of items in the test. The simulation of the parallel test parts is based on the premise that division of the test in parallel test parts is possible (something that should be checked by the test maker or user in a realistic test situation). On the item level, we thus implicitly assume that the mean item difficulty of the items in every parallel test part is the same. Note that this assumption is harder to fulfill when there are few items in a parallel test part.
- (5) **Relationship true score and error variance.** The choice for either the single-, conditional- or person-specific SEM also depends on the anticipated relationship between examinees’ true scores and their error variance. Users of the conditional SEM, for instance, assume that there is a relationship between true score and error variance whereas such an assumption is not (necessarily) considered when opting for the single SEM or the person-specific SEM. In this simulation study we investigate the influence of four different relationships between true score and error variance, which are all depicted in Figure 2.5.

The relationships in Figure 2.5 are all based on the ‘continuous’ scenario. Thus, these relationships show what error variances we might expect for different true scores if there were no test restrictions (i.e., no minimum or maximum scores and no rounding). It is important to stress that the error variances (y-axis) in the ‘continuous’ scenario might differ somewhat from the error variances when test restrictions such as truncation and rounding are taken into account. Due to rounding and truncation, examinees at the extremes, for instance, might have a lower error variance over test takes than in the scenario with continuous scores. On the scale of the test, the results of these examinees are thus relatively consistent and even though the test might make an error theoretically (i.e., the ‘true’ score of an examinee is -15 while he/she keeps getting the minimum score 0 at the test), the test makes this error consistently over test takes, lowering the error variance $\sigma^2(E_{*i})$ over test takes. In the results section, the continuous ‘true’ error variances (Figure 2.5) are used to assess bias and the MSE in the ‘continuous’ scenario whereas the bias and MSE in the ‘rounded and truncated’ scenario are assessed with error variances that are adjusted to take this rounding and truncation into account.

- (6) **Overall reliability of the test.** To test the influence of (overall) reliability, we created data for an overall reliability of .7, .8 and .9. According to often used rules-of-thumb, .7 is the minimum acceptable level of reliability, .8 is preferred and .9 is desirable for high stakes assessment (see Nunnally & Bernstein, 1994). As discussed previously, the reliability of a test for a certain norm population can be

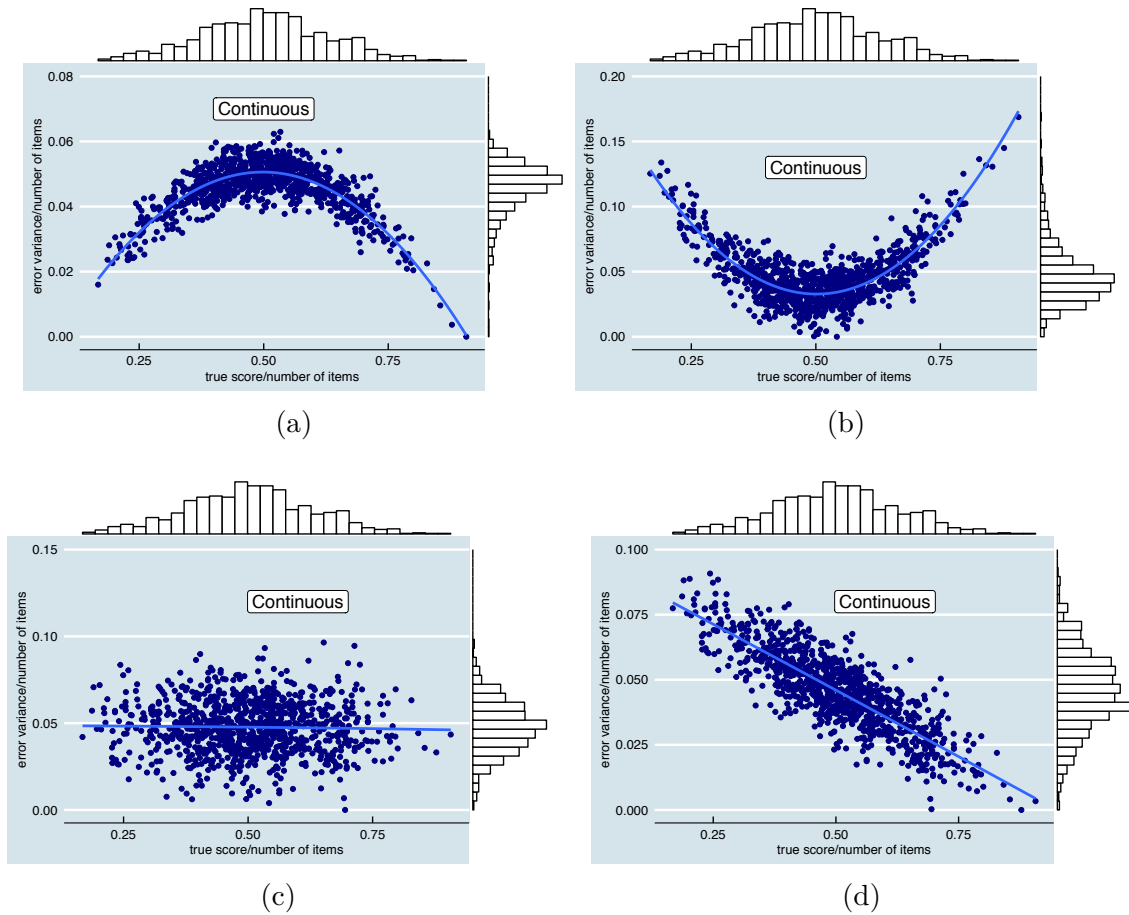


Figure 2.5: Visualization of the four simulated relationships between ‘true’ total test score (x-axis) and error variance (y-axis). Each dot represents one of the 1,000 fictional examinees. To ease comparison, the error variance and the true score are divided by the number of items in the plots. Note that the y-axis scale depends on the chosen level of reliability. (a) Quadratic relationship between true score and error variance, with lower error variances for examinees with extreme true scores. (b) Quadratic relationship between true score and error variance, with higher error variances for examinees with extreme true scores. (c) Absence of a relationship between true score and error variance. (d) Linearly decreasing relationship between true score and error variance.

expressed as: $\frac{\sigma^2(T_*)}{\sigma^2(X_{g*})}$; the ratio of the ‘true’ score variance to the total variance in X_{g*} where $\sigma^2(X_{g*}) = \sigma^2(T_*) + \sigma^2(E_{g*})$ (see Equation (2.7)). In our simulation study, we simulate ‘true’ scores for the 1,000 examinees once for all scenario’s (see “Simulated dataset(s) and the x-axis in Figure 2.5”). Therefore, the variance term $\sigma^2(T_*)$ in the divisor of the ratio is a constant. To increase the reliability, we manipulated the overall variance (denominator) by lowering the error variance $\sigma^2(E_{g*})$. Since $\sigma^2(E_{g*}) = E[\sigma^2(E_{*i})]$ (see Equation (2.9)), we divided every of the 1,000 simulated error variances $\sigma^2(E_{*i})$ by a certain number d to reach the required reliability level. Note that this division results in a higher overall reliability but does not alter the relationship between true scores and error variances as depicted in Figure 2.5 (see section above).

Simulated dataset(s)

True scores τ_i

Following the simulation design described above, data was simulated in a top-down way, starting with the simulation of the distribution of true test scores τ_i over all 1,000 examinees in φ . To accommodate different number of items (see Figure 2.4) we simulated true scores using a standardized metric²¹ and consequently transformed these scores to the scales of 12, 24 and 48 items. Note that the distribution of true scores is fixed over all other scenarios depicted in Figure 2.4.

Error variances $\sigma^2(E_{*i})$

Next, the error variances $\sigma^2(E_{*i})$ were simulated. Just as with the true scores τ_i , the error variances were rescaled for the different number of items. Other than in the case of the true scores, however, the error variances were not fixed over all other scenarios in Figure 2.4. Rather, the error variance of all examinees varied depending on the anticipated relationship between true score and error variance (see Figure 2.5) and the overall reliability of the test (which simply shifts the values in Figure 2.5 down by a factor d). In total, four (relationships true score-error variance) x three (reliabilities) = 12 error variances $\sigma^2(E_{*i})$ were thus simulated for each of the 1,000 examinees.

Total-, item- and parallel test scores

For each of the 1,000 examinees and their 12 error variances $\sigma^2(E_{*i})$ (see above), a vector of 10,000 error scores E_{*i} was simulated. Since the three types of SEM are all estimated on the level of parallel test scores, we simulated these error scores on the level of the parallel test parts, E_{*ik} , first. For this, we made use of the fact that $\sigma^2(E_{*i})$ is a sum over the parallel test error variances: $E_{*ik} \sim N\left(0, \sigma^2(E_{*i})/K\right)$. E_{*i} simply resulted from adding the parallel errors for a certain test take g . To create parallel test scores for each of the 10,000 test takes, the parallel error score was added to the examinee’s ‘true’ parallel test score τ_i/K (scaled to the number of items in the test). For the ‘rounded and

truncated' scenario, these parallel test scores were rounded to the nearest integer. When the combination of true score and error resulted in a rounded parallel test score lower than 0 or higher than the number of items/ K , the parallel test score was truncated to be zero or equal to the number of items/ K , respectively.

Estimation of the SEms, their bias and estimation variance

After creating the data (see above), the single SEM, conditional SEM and the person-specific SEM were estimated for each of the 1,000 examinees and each of the 10,000 test takes. This process was repeated for all conditions and the 'continuous' and 'rounded and truncated' scenario separately (see Figure 2.4). The person-specific SEM was estimated using Equation (2.11). We subsequently calculated the conditional SEM with Equation (2.16), using the total score as grouping factor. Since literature on the conditional SEM (see for instance Feldt & Qualls, 1996) mentions the additional possibility of grouping examinees in narrow intervals of total scores, we also used Equation (2.16) for groupings including examinees with different – but adjacent – total scores. Specifically, we used groupings in which the total score ± 1 was used as grouping factor (in the remainder denoted by “con. +1”), total score ± 2 (“con. +2”) and total score ± 3 (“con. +3”). When it was impossible to add or subtract a number from the examinee's total score without crossing the threshold 0 or the maximum score, the largest number was taken which could still be added and subtracted. Thus, “con. +2” means that grouping of examinees is based on total score ± 2 for all total scores ≥ 2 and \leq the maximum number of items - 2, total score ± 1 for total scores 1 and the maximum number of items - 1 and just the total score for scores 0 and the maximum number of items, for instance. The single SEM was estimated last based on the Spearman-Brown prophecy formula (see Spearman, 1910) between the parallel test parts over all examinees within every repeated measurement; $\rho_{XX'}$ (see Equation (2.8)). Note that this estimation of the single SEM was based on $K=2$, whereas the estimation of the person-specific and the conditional SEMs was based on varying K . Finally, the bias squared $\left(\left[\widehat{\sigma^2(E_{*i})} - \sigma^2(E_{*i}) \right]^2 \right)$ was estimated for all 1,000 examinees and all conditions ($\widehat{\sigma^2(E_{*i})}$ is the average estimate of $\sigma^2(E_{*i})$ over the 10,000 test takes). This bias squared was added to the estimation variance $\sigma^2\left(\widehat{\sigma^2(E_{*i})}\right)$ over the 10,000 test takes to calculate the MSE.

Results

In the simulation design, six characteristics of the test and the norm population were varied (see again Figure 2.4): (1) whether the parallel test scores are continuous and unrestricted or rounded to integers and truncated, (2) the number of repeated test takes, (3) the number of items of the test, (4) the number of parallel test parts K , (5) the relationship between the 1,000 'true' examinee scores and their error variances and (6) the overall reliability of the test. Below, the influences of each of these characteristics on the bias, estimation variance and MSE of the three kinds of SEM are discussed.

Continuous versus rounded and truncated

When parallel test scores were rounded and truncated, we saw an alteration of the simulation results. This alteration had two general causes: (1) the error variances as depicted in Figure 2.5 were altered and (2) rounding and truncation limited the between variance over parallel test parts. Each of these two causes is discussed below. Because of the influence of rounding and truncation, in the other sections results are discussed separately for the ‘continuous’ and the ‘rounded and truncated’ case.

Alteration of the error variances

The error variances $\sigma^2(E_{*i})$ as displayed in Figure 2.5 are based on continuous parallel test scores. Rounding and truncating these parallel test scores alters the total test scores – the sum over the parallel test scores – and thus the error variances. In a sense, rounding and truncation thus ‘push’ the error variances in a certain direction that takes into account the limitations of the scale of the test. In Figure 2.6, an example of the possible extreme influence of rounding and truncation on the error variances $\sigma^2(E_{*i})$ is showed. This example is based on the relationship in Figure 2.5a and 12 items; comparable plots for the other relationships of Figure 2.5 and 24 and 48 items can be found in Appendix 2.A (see <https://osf.io/qrg4e/>). Based on Figure 2.6 and Appendix 2.A, a few observations can be made. First, since rounding and truncation take place on the level of parallel test scores, the choice for K influences the error variances. Figure 2.7 and 2.8 illustrate why this is the case. These two figures are based on two example examinees for a test with 12 items and $K = 4$. The first examinee has a ‘true’ test score of 4; the second a ‘true’ test score of 6. Before rounding and truncation, these two examinees have the exact same error variance of 0.20; after that an error variance of respectively 0.10 and 1.00. Looking at the distribution of parallel test scores (Figure 2.7) for the examinee with true score 4 (Figure 2.7a) and the examinee with true score 6 (Figure 2.7b) we see that there is indeed very little variation in the parallel test scores for the first examinee while there is a relatively large variation in the second examinee’s case, due to an (almost) equal amount of parallel scores 1 and 2. Looking at the underlying continuous error distributions for one of the parallel test parts (Figure (2.8)), we see that there is (almost) no difference between the two examinees. What is different is the effect of the error on the rounded parallel test score. Since the score 4 is divisible by $K=4$, the continuous ‘true’ parallel test score is 1. Only when the parallel test error is larger or equal to .5 or smaller or equal to -.5, the parallel test score will be rounded one score up and down, respectively. Since the score 6 is not divisible by $K=4$, the continuous ‘true’ parallel test score is 1.5. Any negative error larger than $-.000\dots 1$ leads to a rounding down to 1 whereas any positive error smaller than $.999\dots 9$ has no influence on the rounded score. Thus, although the original errors of examinee 1 and 2 are similar due to their similar (continuous) error variance, the effect of these errors on the rounded parallel test scores is different. Would we change K to another number, say 3, then the effect of the error on the rounded parallel test score would change accordingly.

A second observation is that the truncated and rounded error variances reassemble the continuous error variances more closely when the number of items is larger (see Appendix 2.A). Generally, when there are more than 3 items within each parallel test part, the distribution of truncated and rounded error variances is very similar to that

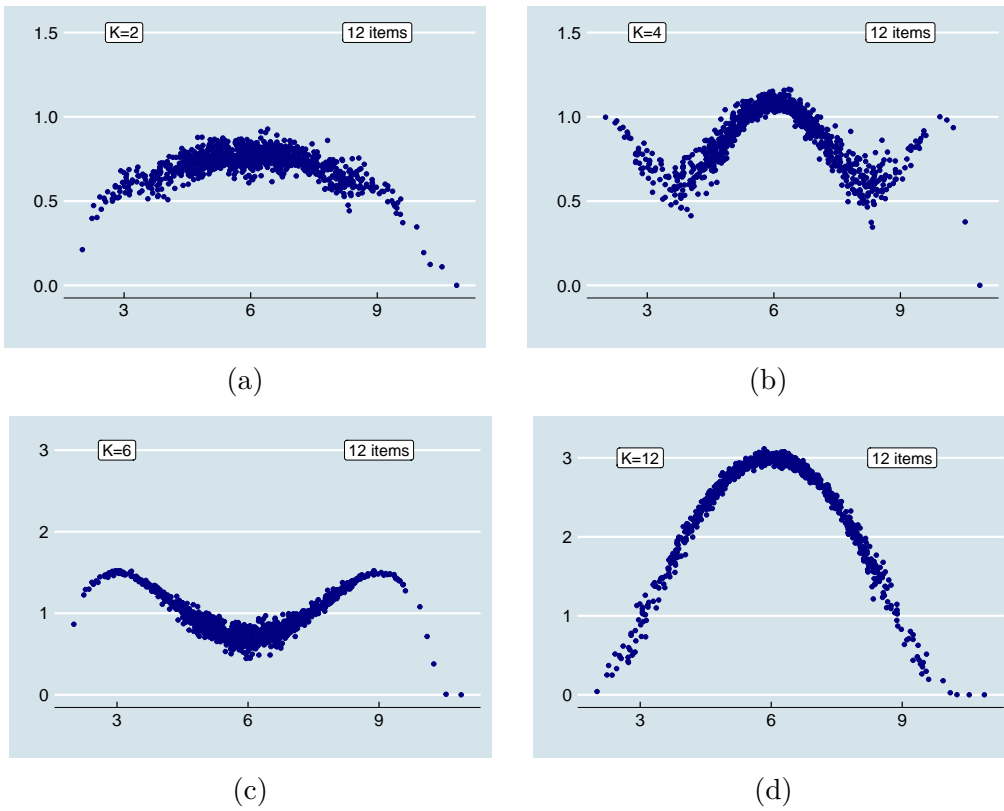


Figure 2.6: Influence of rounding and truncation on error variances for a test with 12 items, a .8 reliability and $K=2,4,6$ and 12 respectively. Note the difference in scale of (a) and (b) versus (c) and (d).

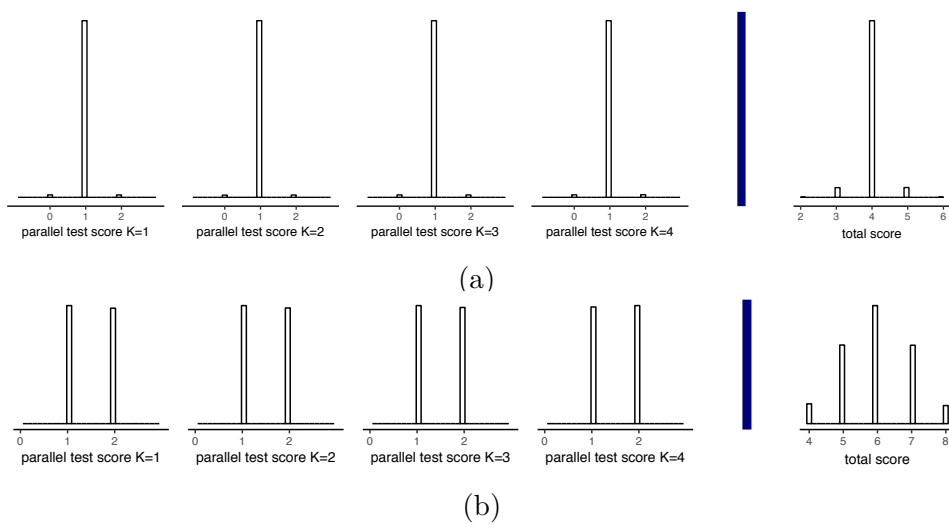


Figure 2.7: Parallel and total test scores for two example examinees with true score 4 (a) and true score 6 (b), both having a true error variance of 0.2 on a 12 item test with $K = 4$.

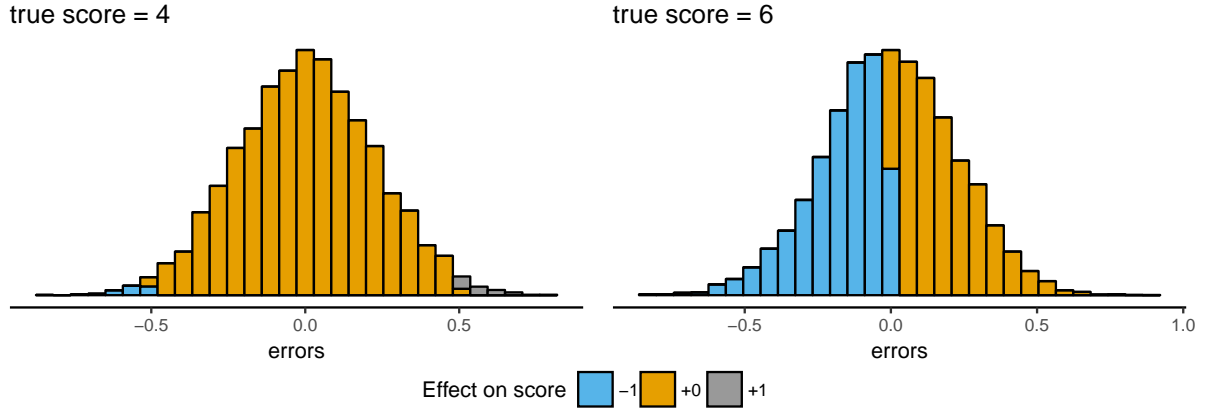


Figure 2.8: Error score distribution for the first parallel test and the effect on the parallel test score for the two examinees of Figure 2.7.

of the continuous one. Last, in the extreme case of having only one item within each parallel test part, the relationship between ‘true’ score and error variance becomes more or less parabolic regardless of the ‘original’ relationship as depicted in Figure 2.5 (see Appendix 2.A). In that case, the error variances approximate those of a binomial distribution (Brennan & Lee, 1999); they become a compromise between the simulated relationship and a binomial one. In Figure 2.9, the relationship between ‘true’ scores and error variances is shown for 12 items and $K=12$ together with the binomial error for comparison.

In the section ‘bias-variance trade-off’, it was explained that the person-specific SEM provides an unbiased estimate of $\sigma^2(E_{*i})$. As visualized in Figure 2.10, this holds when parallel test scores are continuous (Figure 2.10a) and when parallel test scores are rounded and truncated (Figure 2.10b). Equation (2.11) thus holds when we use it based on continuous parallel test scores to estimate the continuous $\sigma^2(E_{*i})$ as depicted in Figure 2.5. Likewise, when we use the person-specific SEM based on rounded and truncated parallel test scores we obtain an unbiased estimate of the rounded and truncated $\sigma^2(E_{*i})$ counterpart (see Appendix 2.A). The person-specific SEM based on rounded and truncated parallel scores only gives an unbiased estimate of the continuous $\sigma^2(E_{*i})$ in so far as the continuous and rounded and truncated $\sigma^2(E_{*i})$ are similar.

Limits on between variance

When the number of items in a parallel test part is small and/or the size of K is small, truncation and rounding lead to another problem: the number of possible between variances becomes limited. Since the person-specific and the conditional SEM are based on the between variance over parallel test parts, this also means that the possible estimates for $\sigma^2(E_{*i})$ are limited. With 12 items and $K=12$, for instance, only 13 unique combinations can be made of 0 and 1 scores (note that Equation (2.26) is an adjustment of the formula for combinations with replacement):

$$\frac{(n_s + K - 1)!}{K!(n_s - 1)!} = \frac{(2 + 12 - 1)!}{12!(2 - 1)!} = 13 \quad (2.26)$$

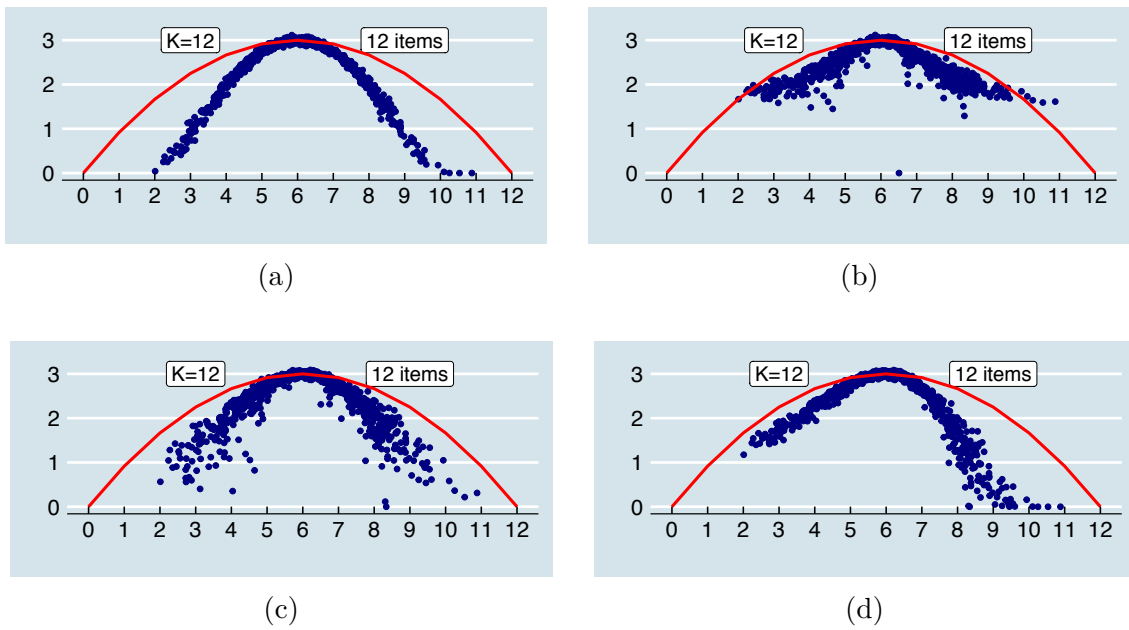


Figure 2.9: When the parallel test scores are truncated and rounded, the relationships between “true” scores and error variances as depicted in Figure 5 change. In the extreme case, when there are only 12 items and $K = 12$, the relationship between “true” scores and error variances of an originally quadratic increase (a), quadratic decrease (b), flat relationship (c) and linear decrease (d) all more or less reassemble that of the red parabola, which corresponds to the binomial error of a 12-item, dichotomously scored test.

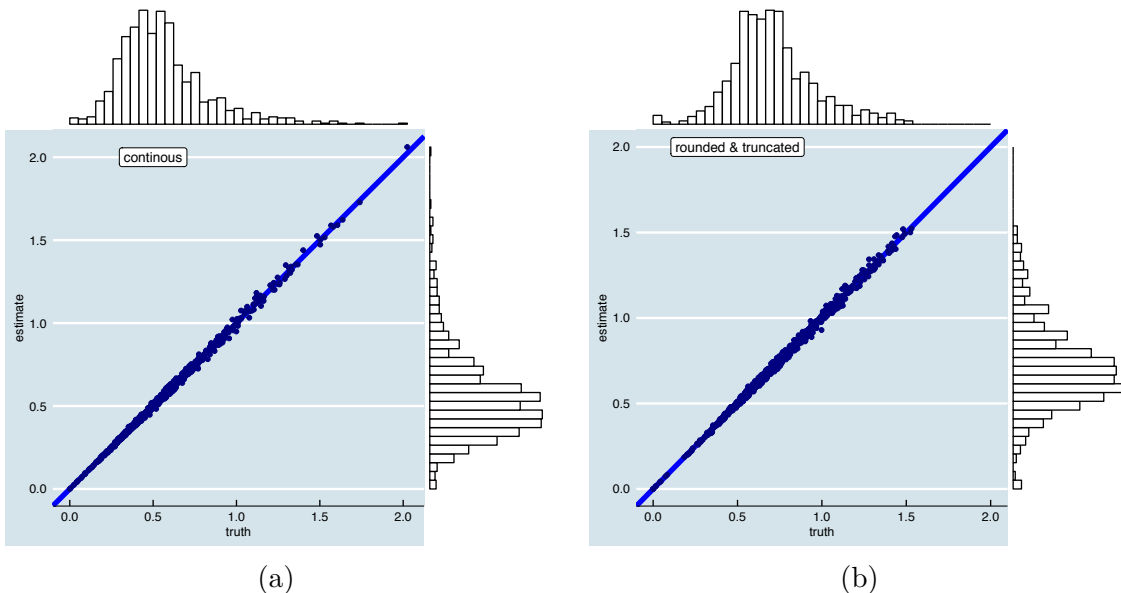


Figure 2.10: Visualization of unbiasedness of the person-specific SEM over 10,000 repeated test takes. The left figure (a) is based on continuous parallel test scores; the right figure (b) on rounded and truncated test scores. The two figures are examples of a test with 12-items, $K=2$, a reliability of .8 and a relationship between ‘true’ score and error variance as shown in Figure 2.5b.

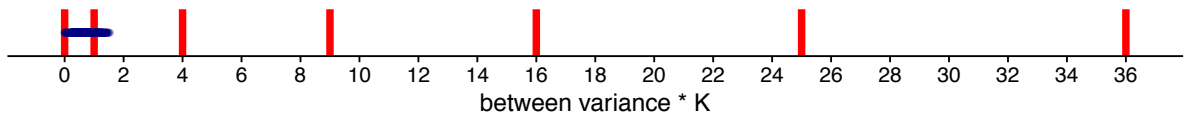
where n_s denotes the number of possible scores (in this case, 2; 0 and 1). With these 13 combinations, only 7 unique between variances can be obtained. Thus, using Equation (2.11), there are only 7 possible estimates of the person-specific SEM at every test take. When there are such few options, there is a high chance that none of the possible SEM estimates are close to the true $\sigma^2(E_{*i})$ of the examinee. In the long run this is not problematic, since we average over the $\sigma^2(E_{*i})$ -estimates and thus are still able to obtain an unbiased overall estimate. But when we only have one test take (as we usually do) this is troublesome. Figure 2.11 shows the options for the between variance when there are respectively 12 items with $K = 2, 4, 6,$ and 12 . Appendix 2.B (see <https://osf.io/qrg4e/>) contains the same figure but then for tests with respectively 24 and 48 items. Looking at Figure 2.11, there are the fewest options for the $\sigma^2(E_{*i})$ -estimate when either the size K is small (i.e., $K = 2$) or when there are few items within each K (i.e., $K = 12$). An important distinction between having a small K versus having few items within each K is, however, that most of the options are within the range of the actual error variances with $K = 12$ whereas most of these options are outside this range when $K = 2$. A reasonable number of items (see appendix 2.B) and a balance between not having too few parallel test parts but also not too few items within each K seems essential. The conditional SEM also suffers from having few estimation options when K and the number of items within K are small. However, since the conditional SEM consists of taking an average over similar examinees, the number of estimation options accumulates the more examinees are taken together. With only one extra examinee, for instance, there are 72 possible combinations of between variances for the example of 12 items and $K = 12$ of which 27 lead to a unique between variance estimate when averaged. This more than doubling notwithstanding, in practice the problem of having few options can still occur. Especially when there is a notable relationship between ‘true’ score and error variance, it is likely that we average over just one or a few between variance options.

Number of repeated test takes

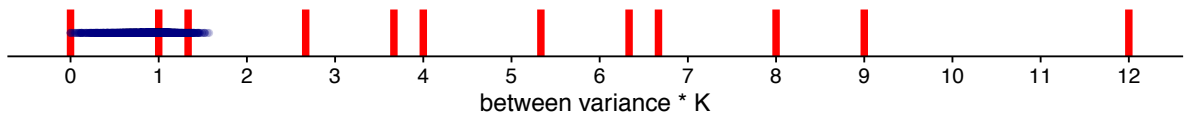
In the section ‘Bias-variance trade-off’ we discussed the long run properties of the three different SEM estimators. However, favorable long run properties do not guarantee that the SEM estimator is appropriate to use in the short run, when only having one or just a few test takes. Below, we discuss the short run properties of the person-specific and the conditional SEM (not the single SEM, since this SEM does not change with more or less test takes) and reasons why the person-specific and conditional SEM might not be as appropriate to use in the short run as in the long run.

Short run performance of SEM

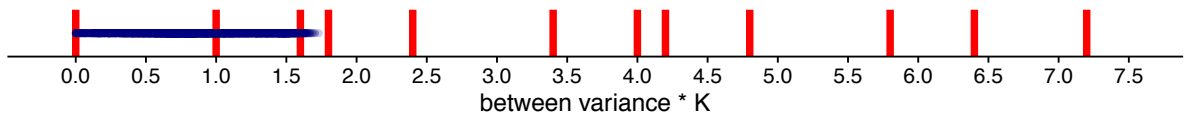
To see how well the conditional and person-specific SEM perform in the short run, we plotted these SEM estimates after only 1, 2, 3, 4, 5, 10, 25 and 50 test takes, together with the true error variances we are trying to estimate²², for both the scenario with continuous parallel test scores and rounded and truncated parallel test scores. The result can be seen in Appendix 2.E (see <https://osf.io/qrg4e/>). The Pearson correlation between the true error variance and the estimates are shown at the right side of the plots. Additionally, a table is presented with the within variance at true error variance .6 and the between variance over all examinees. When there is only one test take, the correlation between



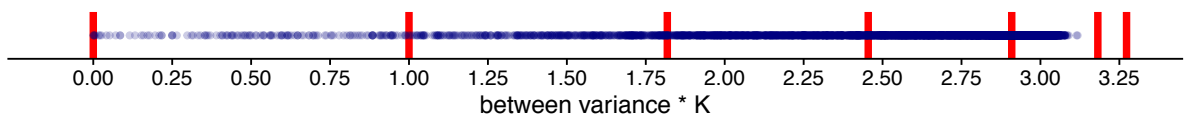
(a)



(b)



(c)



(d)

Figure 2.11: All options of the between variance multiplied by K (red vertical lines; see Equation 11) for $K = 2$ (a), $K = 4$ (b), $K = 6$ (c) and $K = 12$ (d) compared to the “true” truncated and rounded error variances of the 1,000 examinees (blue dots). Note the difference in x-axis scale.

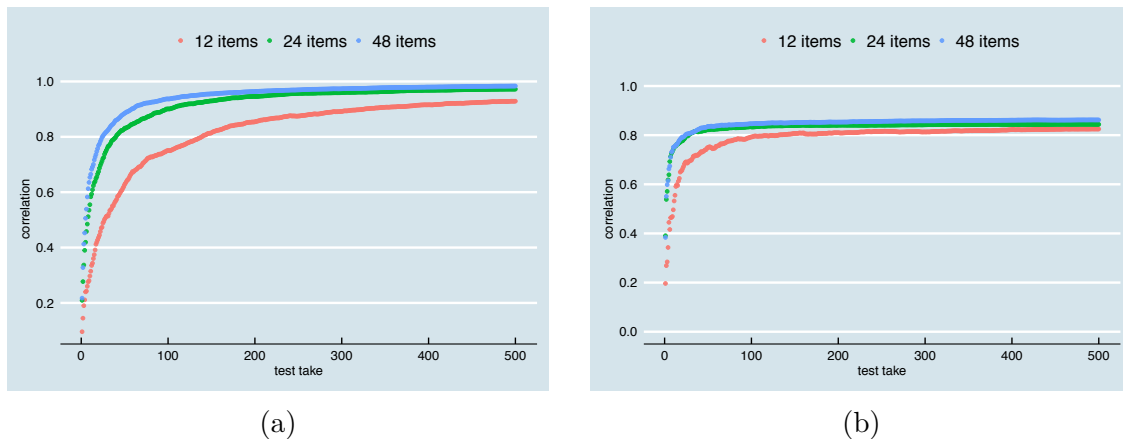


Figure 2.12: correlation between the true error variance and the SEM-estimate for different number of test takes (x-axis) and a test with respectively 12, 24 and 48 items. (a) shows the correlations for the person-specific SEM; (b) a comparable figure for the conditional SEM. Both are based on continuous parallel test scores, an overall reliability of .8 and a relationship between ‘true’ score and error variance as depicted in Figure 2.5a.

the truth and the SEM estimates is very low for the person-specific SEM (only .09 in both the continuous and rounded and truncated case). The conditional SEM is only slightly better (.20 in the continuous case and .12 in the rounded and truncated case). The SEM estimates for different true error variances appear to overlap to a large extent, although variation in the SEM estimate increases along the x-axis and higher estimates are observed occasionally for higher true error variances. After 50 test takes, the correlation has increased notably but the variance in estimated SEM for examinees with .6 as their ‘true’ error variance is still about the same as the overall between variance. As Figure 2.12 shows, about 100 (conditional SEM, Figure 2.12b) to 200 (person-specific SEM, Figure 2.12a) test takes are needed to reach high levels of correlation. When parallel test scores are rounded and truncated (second part Appendix 2.E), we generally need more test takes (see Appendix 2.F, which contains the equivalent to Figure 2.12 for rounded and truncated parallel test scores). The correlations are therefore generally lower than in the continuous case.

Reasons for bad short run performance

We found two general reasons for the suboptimal short run performance as observed in Appendix 2.E and Figure 2.12. One of these reasons was already discussed in the previous section and visualized in Figure 2.11: because of few possible between variances, the SEM for a single test can only take on few different values. This influence is clearly visible in the first plot of the ‘rounded and truncated’ scenario in Appendix 2.E. There is also another reason why the person-specific (and to a certain degree the conditional SEM) might not be suitable estimators of the error variance after one or just a few test takes, even when parallel test scores are continuous. As visualized in Figure 2.13, the distribution of possible between variances for different underlying error variances overlap to a high extent, especially in the small between variances range. Therefore, if we would randomly observe one between variance it can be the result of very different underlying error variances. As Figure 2.13) shows, a problematic feature of the distributions of

between variances is that the variance of this distribution directly depends on the size of the underlying error variance (see Equation (2.20) and (2.21) and the sd in Figure 2.13). Therefore, the accuracy of any estimate based on the between variances directly depends on that what we are trying to estimate; the error variance. Since the error variance is unknown, we do not know exactly how accurate we can expect one estimate to be. Other problematic features of the distribution of between variances are that zero is always the most probable between variance, no matter how large the underlying error variance (see again Figure 2.13) and the overabundance of zeroes in Appendix 2.E) and that the distribution has a large tail. Regarding the former, we might be tempted to conclude that there is little error variation when we observe a between variance of zero but the underlying error variance might actually be relatively large. With regard to the latter, for comparison purposes Figure 2.12 only shows the histogram values until .75, but as the blue arrow and the “max = .” shows, the distributions are stretched out over a far wider range. The large between variances are rare, but they are necessary to reach an accurate estimate on average. Due to the relationship between error variance and variation in the possible between variances, observing a low between variance does not contain as much information about the underlying error variance as does observing a high between variance. Whereas a low between variance can be the result of small, medium or even large error variances, a large between variance is very unlikely to be the result of a small or even medium error variance. Finally, the distribution of between variances does not become less spread out when we include more items or increase K .

Number of items

By incrementing the number of items, the scale of the parallel test scores and hence the size of the true error variances (see Figure 2.5 and Appendix 2.A) increased. Consequently, comparing the absolute bias, estimation variance and MSE is not very insightful, as it simply reflects this change in scale. To compare the bias and estimation variance for the different number of items, we therefore decided to use a relative version. Specifically, we express bias as a percentage of the true error variances:

$$\text{bias} = \frac{\sum_i \left(\widehat{\sigma^2(E_{*i})} - \sigma^2(E_{*i}) \right)}{\sum_i \sigma^2(E_{*i})} \quad (2.27)$$

and we use the coefficient of variation (see Walther & Moore, 2005) as a relative measure of estimation variance, which is simply the estimation standard deviation expressed as a percentage of the mean:

$$\text{CV} = \frac{\sum_i \sqrt{\sigma^2(\sigma^2(E_{*i}))}}{\sum_i \sigma^2(E_{*i})} \quad (2.28)$$

Below, we discuss the percentage bias and coefficient of variation when compared over tests with different lengths (i.e., with 12, 24 and 48 items). We also discuss whether and how having a larger test is beneficial for the SEM-estimates.

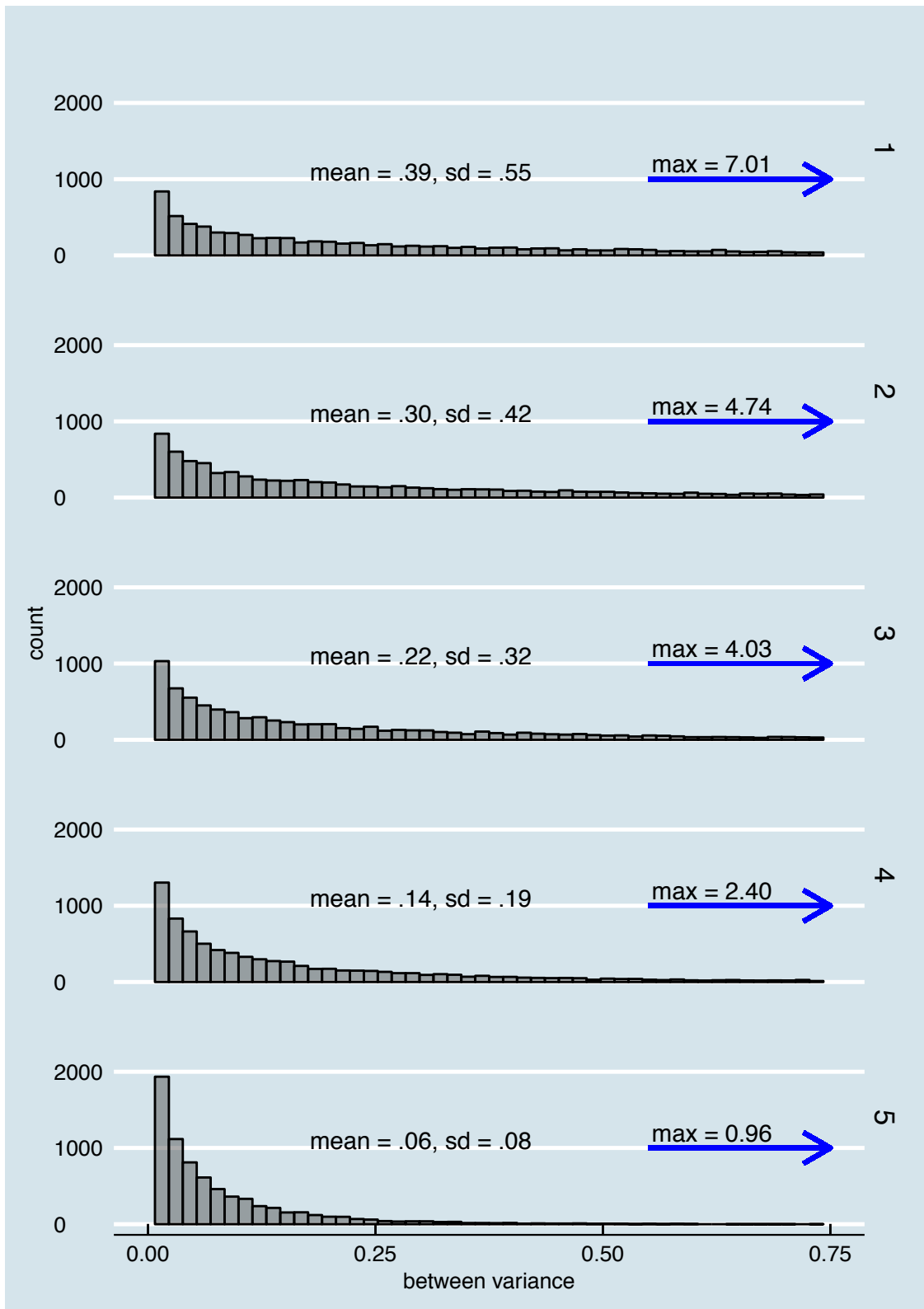


Figure 2.13: Overlap in the distributions of possible between variances for 5 examinees with underlying error variances .75 (examinee 1), .60 (examinee 2), .44 (examinee 3), .28 (examinee 4) and .12 (examinee 5).

Percentage bias

In Figure 2.14, the percentage bias is visualized for different number of items, for the relationship depicted in Figure 2.5a and with continuous parallel test scores (for the other relationships, see Appendix 2.D). According to this figure, bias is not dependent on the number of items in the test (all bars are of equal height for the person-specific, conditional and single SEM). Note that the percentage bias for the person-specific and conditional SEM is very close to zero, but not exactly null. The small deviation from zero stems from the fact that we have a limited norm population size (1,000) and a limited number of repeated test takes (10,000) over which we determine the bias. The covariance between the parallel test parts (see Equation (2.10a)) over the 10,000 test takes are, for instance, very close to zero, but not exactly zero (as Equation (2.10b) assumes). The “ $2 \sum_{k < p} \sigma^2(X_{*ki}, X_{*pi})$ ” part in Equation (2.10a) for the first five examinees for a test with 12 items, 2 parallel test parts, an overall reliability of .8 and a relationship as in Figure 2.5a, are for instance -0.00056, -0.001686, -0.004748, -0.000355 and -0.014109. Therefore, the between variance (see Equation (2.12)) on which the person-specific and conditional SEM are based is not exactly equal to the within variance of Equation (2.10a), leading to a small bias over the limited set of test takes. As Equation (2.12) postulates, this bias is expected to disappear with an infinite number of tests. When parallel test scores are rounded and truncated, the percentage bias as observed in the continuous case changes (see Appendix 2.D). Due to truncation and rounding, the percentage bias is not exactly the same anymore for different test lengths. The percentage bias now depends on the changed relationship between ‘true’ test score and error variance (see Appendix 2.A) and the between variances that can be estimated with a certain number of items and K (see Appendix 2.B). Additionally, the percentage bias decreases in the case of the “con.+1”, “con.+2” and “con.+3” when moving from a 12-item to respectively a 24-item and 48-item test. This has to do with the fact that “+1” is a larger step in a 12-item test (1/12th) than in a 24-item test (1/24th) or a 48-item test (1/48th). Hence, we introduce relatively more bias when we include examinees with +1 scores when the scale of the test is smaller (and similarly for examinees with +2 and +3 scores).

Coefficient of variation

Figure 2.15 shows the coefficient of variation for different number of items for the relationship depicted in Figure 2.5a with continuous parallel test scores (for the other relationships, again see Appendix 2.D). Not surprisingly, the coefficient of variation is exactly the same over different numbers of items for the person-specific SEM and the single SEM. As expressed earlier in Equation (2.20)-(2.21), the estimation variance of the person-specific SEM does not depend on the number of items. The single SEM has a small random variation from test take to test take (since the scores of the norm population change randomly over test takes), but this random variation also does not depend on the number of items. For the conditional SEM, there is a small increase in the CV when moving from a 12- to a 24- and 48-item test. Inspection of Equation (2.21) shows that the estimation variance for the conditional SEM not only depends on K (as the variance for the personal SEM) but also on n_j ; the number of examinees within one group j . Since we kept the overall number of examinees (1,000) equal over the different test lengths, the number of examinees with the same total score decreased when doubling the number of items from 12 to 24 and from 24 to 48; leading to the

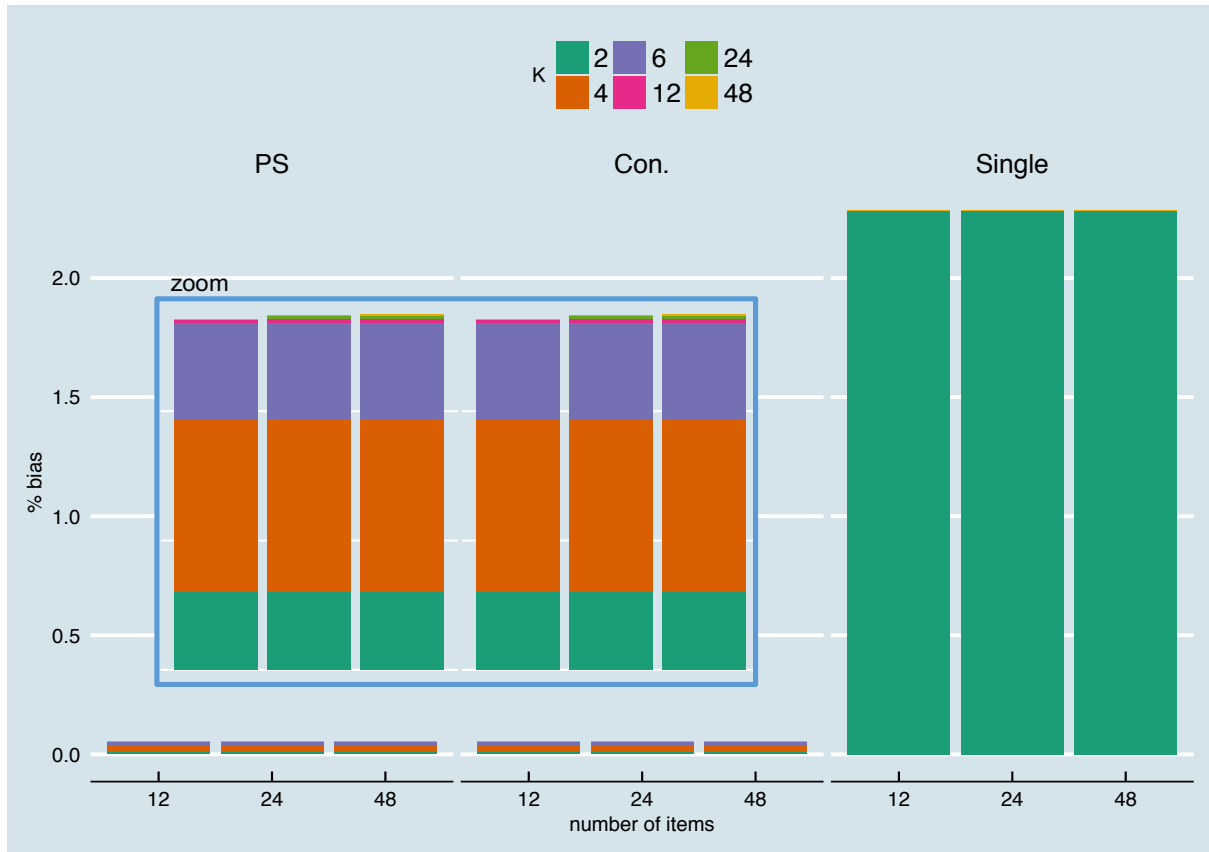


Figure 2.14: Percentage bias (see Equation 25) for different number of items, for the relationship depicted in Figure 2.5a and with continuous parallel test scores.

slight increase in CV. When parallel test scores are rounded and truncated, the CV changes (see Appendix 2.D). For the person-specific SEM, the CV is not exactly the same anymore over the different test sizes. Additionally, the CV becomes smaller when moving from a conditional to a conditional “+1”, “+2” and “+3” for a fixed test size. This again has to do with the n_j in Equation (2.21). Because of the +1, +2 and +3, more examinees are included in a group j leading to a smaller estimation variance. This decrease in CV is largest for the 12-item test since – as discussed previously – the +1, +2 and +3 are relatively larger intervals in a smaller test. Including examinees with +1 therefore increases n_j faster for a smaller test.

Benefits of having a larger test

Looking at Figure 2.14, Figure 2.15 and Appendix 2.D, in the long run there is no advantage in having a larger test when estimating the SEM. The relative bias and estimation variance do not change when increasing the number of items, as bias and estimation variance do not depend on test length (see section “Bias-variance trade-off”). When using rounded and truncated parallel test parts, however, there is an advantage in having a larger test as discussed in the previous sections. With a larger test size, the estimable error variance depends less on K and the between variance can take on more different values. Additionally, in the short run the SEM estimates are more accurate when based on a larger test. As Figure 2.12 showed, with a larger test, initial correlations

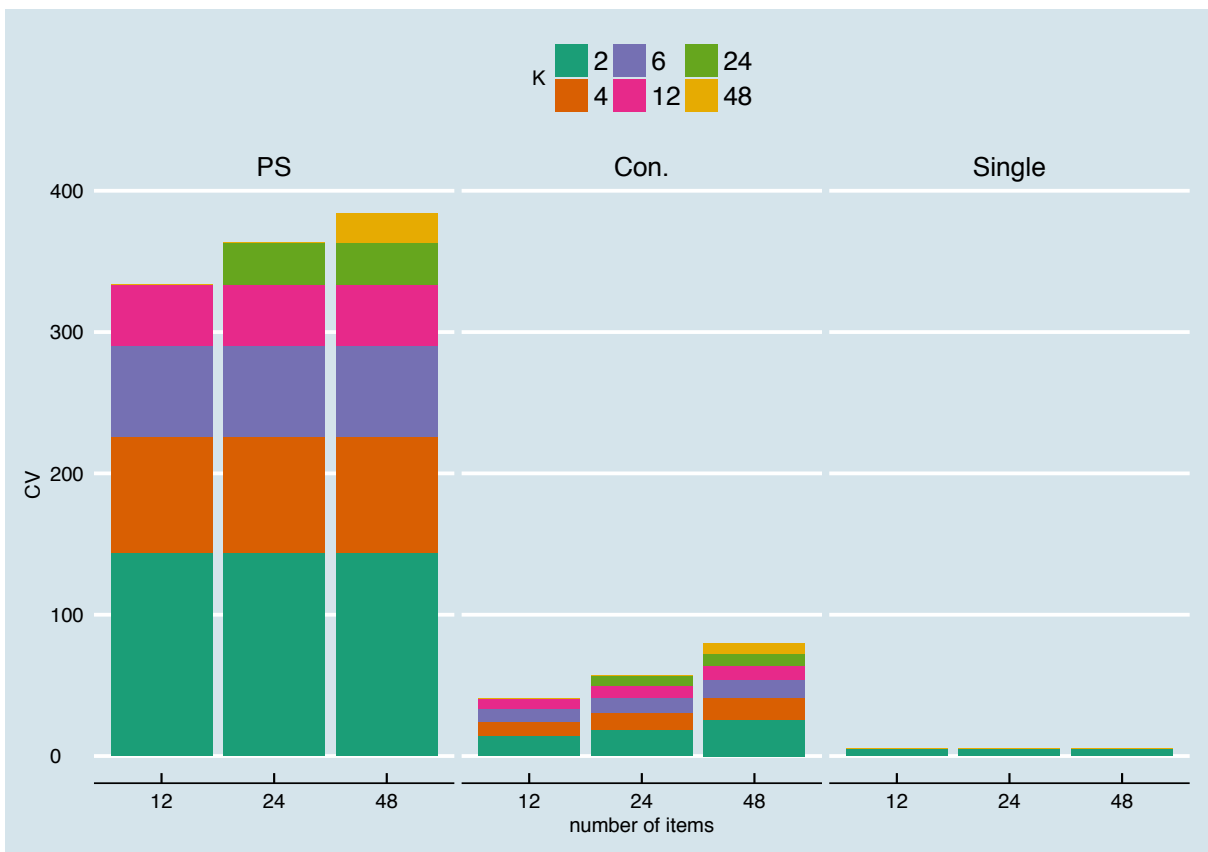


Figure 2.15: Coefficient of variation (see Equation 26) for different number of items, for the relationship depicted in Figure 2.5a and with continuous parallel test scores.

between true error variance and the SEM estimates are higher and there are fewer repeated test takes needed to reach a high correlation.

Number of parallel test parts

Figure 2.15 (previous section) shows why it is beneficial for the person-specific and conditional SEM to choose K as high as possible: the estimation variance goes down with increasing K . This relationship between the estimation variance and K was also discussed in the section “Bias-variance trade-off”. A more puzzling result of the previous section (see Figure 2.14) is that the relative bias observed after 10,000 test takes also seems to depend on K to some extent. In Figure 2.14), the relative bias of $K = 12$ is, for instance, structurally lower than the relative bias of $K = 4$ (also see Appendix 2.D). It is unclear why this difference in bias for different values K occurs, especially since it is not linearly related to K . Since we are using a limited number of repeated measures (10,000) and a limited sized norm-population (1,000 examinees) and R has a limited precision in its calculations, the differences observed may be a reflection of these limitations instead of real differences over K . Inspection of the average SEM-estimates with different sizes for K at least shows that the absolute differences are neglectable in any practical situation. All in all, the comment that it is “advantageous to use as many [parallel] parts as possible” (Feldt & Qualls, 1996, p. 154) seems to hold in the case of continuous parallel test scores. When parallel test scores are rounded and truncated, the choice for K becomes more delicate. As was already discussed in the section “Continuous versus rounded and truncated”, the observable error variance depends on characteristics of the test, including the size of K . When every parallel part contains few items, the resulting error variance estimate highly depends on K and therefore the estimate of the error variance cannot be generalized to a similar test with larger or smaller K . Furthermore, the choice for K influences the between variances that can be observed (see again Figure 2.11). Instead of opting for the largest K possible, it therefore seems advisable in case of rounded and truncated scores to balance K and the number of items within each parallel test part.

Relationship true score and error variance

In this simulation, we varied the relationship between ‘true’ score and error variance according to Figure 2.5. The question is whether the preference for one of the different SEMs depends on these underlying relationships. Figure 2.16 shows a grid with all possible combinations of K , the number of items and the overall reliability for the relationships depicted in Figure 2.5a (Figure 2.16a), Figure 2.5b (Figure 2.16b), Figure 2.5 (Figure 2.16c) and Figure 2.5d (Figure 2.16d). The colors within this grid show which of the three SEMs had the lowest absolute MSE and thus would be favored in which situation. Figure 2.17 shows similar figures but for the case with rounded and truncated parallel test parts. In these figures, also the conditional “+1”, “+2” and “+3” are included. Turning first to Figure 2.16, we see a comparable pattern for all relationships except for the ‘flat’ or ‘no relationship’ case (Figure 2.5c and Figure 2.16c). When no relationship between ‘true’ score and error variance is anticipated, the single SEM is generally the best choice unless the number of items and K are large. When we do anticipate a relationship between ‘true’ score and error variance, the conditional SEM is generally favored, unless K is small relative to the number of items (favoring the single

SEm) or when both the number of items and K are large (favoring the person-specific SEm).

Looking at Figure 2.17, we see again some comparability over all relationships, except for Figure 2.17c. For the ‘flat’ or ‘no relationship’ case (Figure 2.17c), the single SEm is again favored most, together with the conditional SEm with a large interval width. Opposed to the single SEm, the “con.+3” can adjust its estimate such that the error variances for ‘extreme’ true scores are correctly estimated. Since variation at these extremes of the scale (see Figure 2.5c) is less than in the middle, having a SEm-estimate close to the real error variances at these extremes lowers the overall discrepancy between the estimated SEm and the true error variances. In the other three figures, the “con.+3” option is also a popular choice, especially when the test has 48 items. To a lesser extent, also the “con.+1” and “con.+2” options are favored over the person-specific and the ‘regular’ conditional SEm, which are no longer leading to the lowest MSE in any of the conditions. The popularity of the conditional SEms with a larger interval makes sense, because they are stable (like the single SEm) but also flexible enough to capture the anticipated relationship between ‘true’ score and error variance. They also lead to the largest possible group size n_j , which directly lowers the estimation variance (see section “Bias-variance trade-off”). Other than in the case of continuous parallel test scores, the single SEm is barely chosen when K is small relative to the number of items in Figure 2.17a, b and d. For rounded and truncated scores, the single SEm is for instance favored when the number of items is small (i.e., 12) and K relatively large. In the section on rounding and truncation, we saw that the relationship between true score and error variance was highly altered when the number of items was small relative to K (see Figure 2.6). For a 12-item test, this influence was already present with a K larger than 2. When such an alteration occurs, it is better to stick with the single SEm (which is always based on $K = 2$, see section “Simulated dataset(s)”) than to opt for a conditional SEm with a larger K .

Overall reliability

As explained under “simulation design”, the overall reliability of the test was manipulated by dividing every examinee’s error variance by d . Since this lowers the size and therefore the scale of the error variances, we expect the bias², estimation variance and MSE to go down when moving from a .7 to a .8 and .9 overall reliability. Figure 2.18 shows that this is indeed the case for a test with 24 items, with K running from 2 to 6 and error variances as in Figure 2.5a. In this figure, the reliability is varied on the x-axis and every type of SEm is shown separately (note the difference in y-axis for the person-specific SEm compared to the other SEms). Despite the decrease in bias² in absolute sense, there is no clear downward trend in the percentage bias (see Table 2.1 and Equation (2.27)). Additionally, the relative differences in MSE, bias² and estimation variance between the single, person-specific and conditional SEms are comparable regardless of whether there is an overall reliability of .7, .8 or .9 (see Figure 2.18). Figure 2.18 furthermore shows that the proportion of the MSE ‘caused’ by bias is not notably different for the different overall reliability except for the conditional SEm (see percentages on top of the bars in Figure 2.18). This bias inflation for relatively low reliabilities is caused by using the total score instead of the true score of examinees as grouping factor in the calculation of the conditional SEm (see Woodruff, 1990). The higher the overall reliability, the smaller

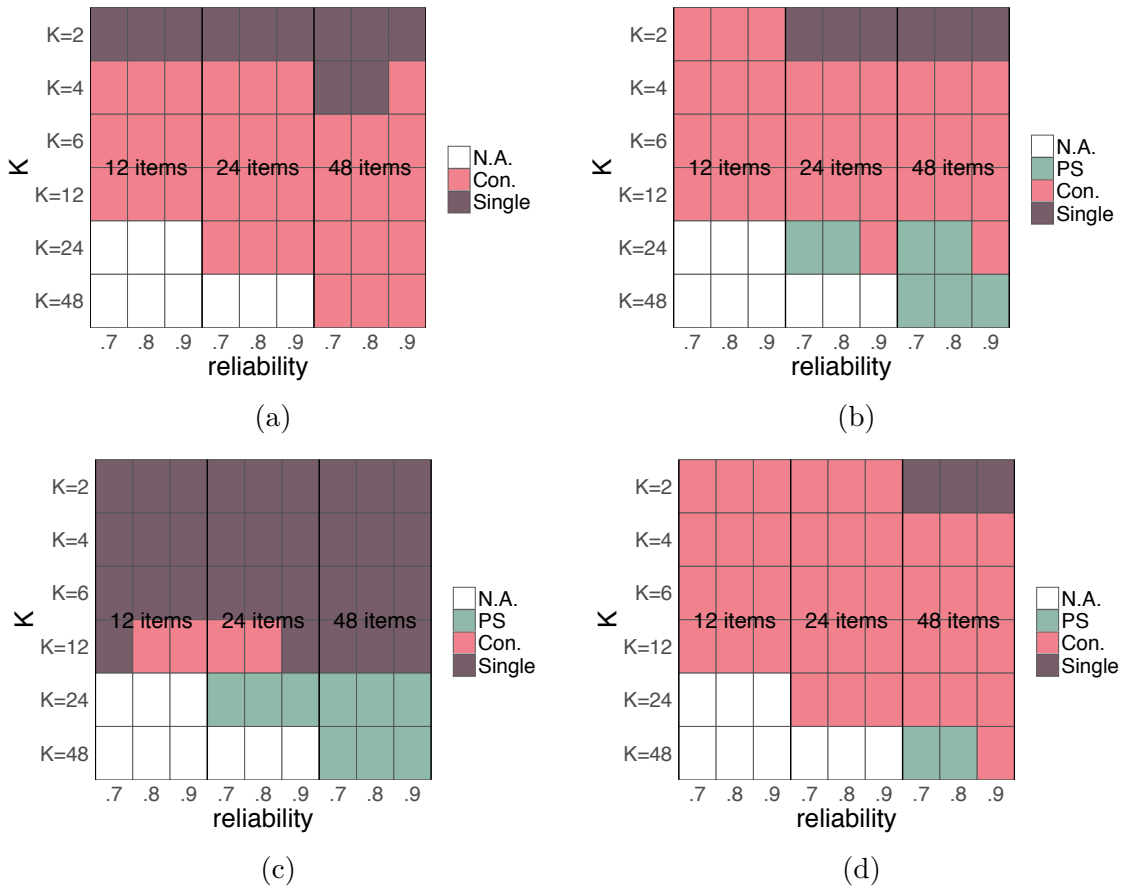


Figure 2.16: Preference for the person-specific SEM (green), the conditional SEM (pink) or the single SEM (purple) based on the MSE for different K 's, number of items and overall reliability. Plot (a) corresponds to the relationship between 'true' scores and error variances as depicted in Figure 2.5a, (b) to the relationship of Figure 2.5b, (c) to the relationship in Figure 2.5c and (d) to the relationship in Figure 2.5d. All four plots are based on continuous parallel test scores.

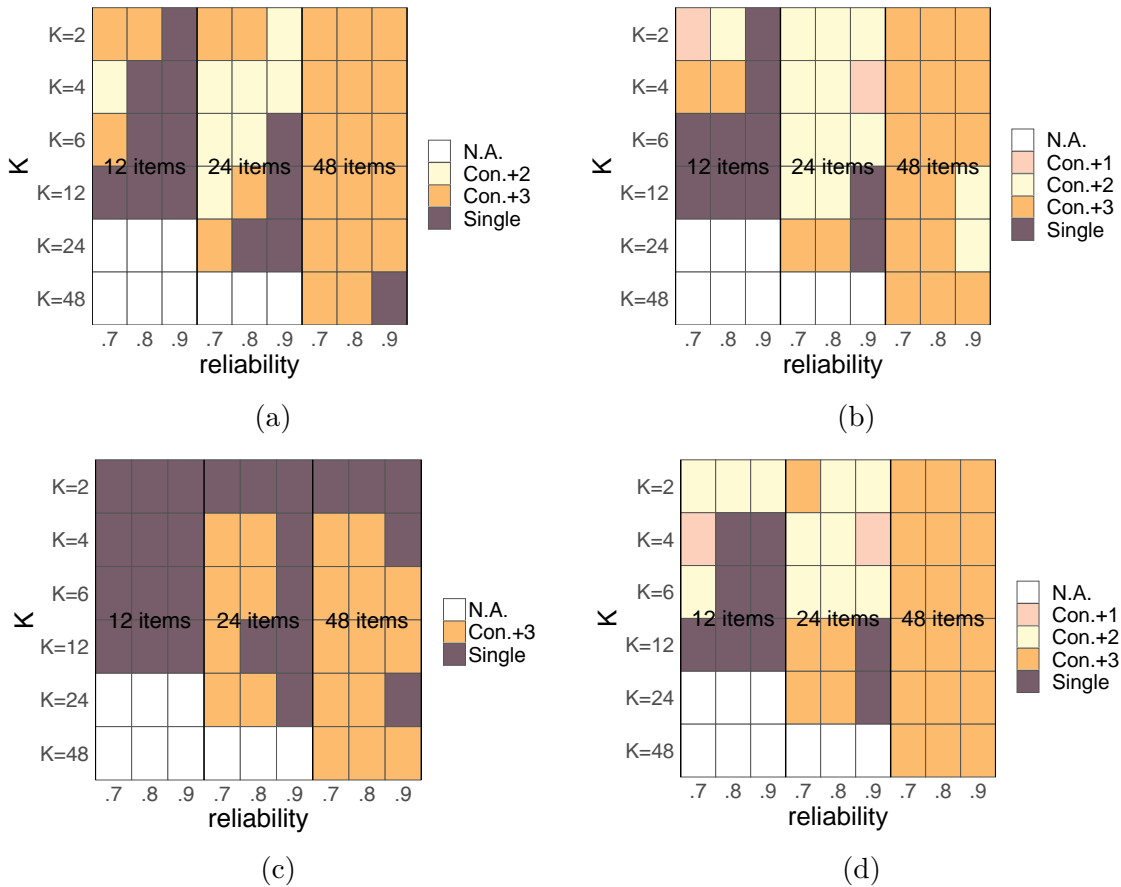


Figure 2.17: Preference for the single SEM (purple) and the conditional SEMs +1 (peach), +2 (yellow) and +3 (orange) based on the MSE for different K 's, number of items and overall reliability. Note that the person-specific SEM (in the previous Figure 2.16 in green) and the conditional SEM (in the previous Figure 2.16 in pink) were also included as options, but never resulted in the lowest MSE in any of the conditions. Plot (a) corresponds to the relationship between 'true' scores and error variances as depicted in Figure 2.5a, (b) to the relationship of Figure 2.5b, (c) to the relationship in Figure 2.5c and (d) to the relationship in Figure 2.5d. All four plots are based on rounded and truncated parallel test scores.

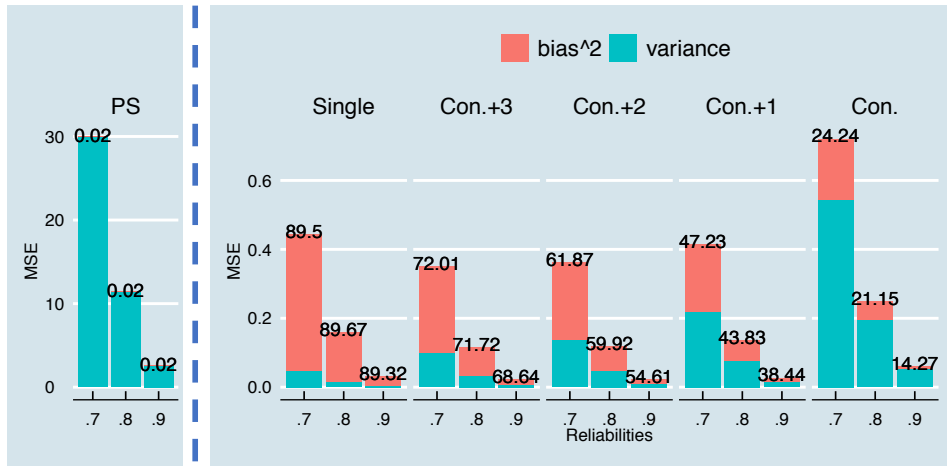
the difference between the total and true scores and thus the less bias is introduced in the grouping. The results shown so far are all based upon rounded and truncated parallel test scores. Appendix 2.C contains similar plots as in Figure 2.18, but then for continuous parallel test scores. Generally, the results with continuous and rounded and truncated parallel test scores are very similar. The MSE, bias² and estimation variance are, however, overall somewhat lower when no rounding and truncation have taken place. Also, rounding and truncation apparently increase the bias² notably when opting for a conditional SEM; in the continuous case (see Appendix 2.C) the MSE of the conditional SEM is almost completely due to estimation variance.

Table 2.1: Percentage bias accompanying Figures 2.18a–c.

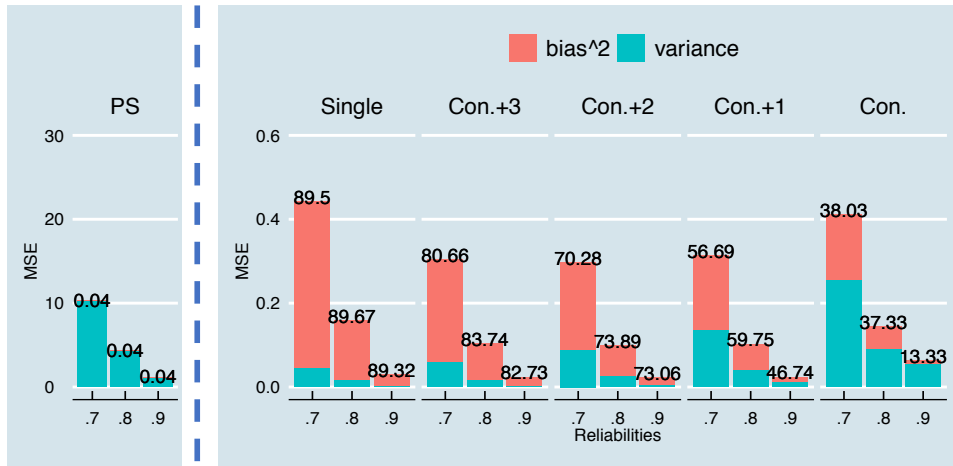
	$K=2$			$K=4$			$K=6$		
	0.7	0.8	0.9	0.7	0.8	0.9	0.7	0.8	0.9
PS	-0.01	0.01	0.00	0.00	-0.04	-0.08	0.01	-0.02	-0.01
Con.	-0.01	0.01	0.00	0.00	-0.04	-0.08	0.01	-0.02	-0.01
Con.+1	0.33	0.40	0.49	0.38	0.32	0.24	0.41	0.29	-0.22
Con.+2	0.82	0.96	1.02	0.95	0.87	1.05	1.01	0.79	0.72
Con.+3	1.24	1.44	1.55	1.46	1.34	1.41	1.55	1.27	1.98
Single	3.29	3.46	2.10

Discussion

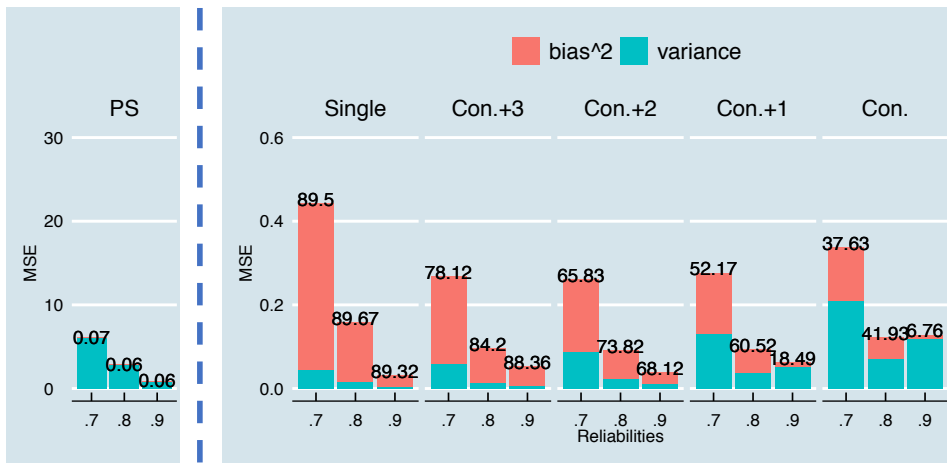
By convention, the single Standard Error of Measurement (SEm) is used to express measurement (un)certainty of examinees (see Equation (2.8)). By using this single SEm, two strict assumptions are made regarding measurement (un)certainty. First, it is assumed that all persons are measured with the same accuracy; i.e., measurement error variance is assumed constant. Second, the person's intra-individual variation in measurements is assumed equal to the inter-individual variation in measurement of the norm-population examinees; i.e., intra-individual variation and inter-individual variation are interchangeable (Molenaar, 2004). To circumvent these assumptions, one can also opt for the conditional SEM. The conditional SEM relaxes the two assumptions in that only examinees with the same (expected) true scores are assumed to have constant measurement error variance and interchangeability of inter- and intra-individual variation. Instead of basing the estimate of SEM on all examinees in the normpopulation, like the single SEM, the conditional SEM takes the expected value of the person-specific measure of SEM over all examinees with the same obtained total score (see Equation (2.16)). In the conditional SEM literature, this person-specific measure of SEM is treated as an intermediate step but – as this chapter showed – can also be used as a result on its own. Using this person-specific SEM has the advantage that no assumptions are made regarding measurement (un)certainty (see Equation (2.11)). Every relaxation of the two assumptions of the single SEM – however - comes with a price: in practice, the estimation variance of the conditional SEM and person-specific SEM always exceed the estimation variance of the single SEM. In this chapter, we have illustrated how the mean squared error (MSE) - a measure that is based on both estimation variance and bias - can be used to make a choice for either the single, the person-specific or the conditional SEM



(a) $K = 2$



(b) $K = 4$



(c) $K = 6$

Figure 2.18: bias², variance and MSE for a test with 24-items and $K = 2$ (a), $K = 4$ (b) and $K = 6$ (c) parallel test parts. The overall reliability of the test is varied on the x-axis. The numbers on top of the bars show the percentage of the MSE ‘caused’ by the bias², which is also reflected in the size of the stacked pink bar. Note that these plots are based on the relationship in Figure 2.5a.

in realistic test situations. This ‘optimal’ SEM guarantees that the estimated SEM is as close as possible to the true error variance of examinees. We also showed, using a simulation study, how the person-specific SEM, the conditional SEM and the single SEM are influenced by six characteristics of the test situation: (1) whether the parallel test scores are continuous and unrestricted or rounded to integers and truncated, (2) the number of repeated test takes, (3) the number of items of the test, (4) the number of parallel test parts K , (5) the relationship between the 1,000 ‘true’ examinee scores and their error variances and (6) the overall reliability of the test. Refraining from repeating all findings for each of these six characteristics here, we like to highlight three important overall conclusions from the simulation study. First, it is important to realize that the three different SEMs are developed for a situation with continuous test scores. When scores are rounded and truncated – as they typically are in practice – we face a number of challenges in the estimation of the person-specific and the conditional SEM that we should keep in mind. Second, it is important to stress that even though the MSE helps to balance long run unbiasedness and estimation variance, it does not guarantee that the estimates for a single test take make sense. Rounding and truncation can for instance lead to sparseness in the possible SEM-estimates (see Figure 2.11), such that no single estimate can be close to the real error variance after one test take. One single estimate also relatively often equals zero, since zero is the between variance that most often occurs for any underlying error variance (see Appendix 2.E). It would be dangerous to conclude that this examinee is thus measured very accurately, since the underlying error variance can still be large. Third, it is advisable to think about the purpose of the SEM-estimation, and to choose accordingly. The conditional SEM, for instance, can be suitable in certain test situations to get an idea of how variable we expect the test scores to be. It might not be so suitable, however, to construct a confidence interval and to see whether this confidence interval overlaps with or is smaller/larger than the confidence interval of another examinee. Since both the total score and the SEM on which the confidence intervals are based vary, we don’t want to draw conclusions that might be based on random error in one or all total scores and one or all error variances. It might be fairer to use the single SEM in that case. As a last example: the person-specific SEM might not be suitable in a certain situation as a direct measure of error variance but it could be suitable to select examinees for which the estimate is relatively large. Since large between variances are rare (or non-existent) for small and medium error variances (see Figure 2.13), a large between variance points to a large underlying error variance. This can be a reason to be careful with using the single SEM for this specific examinee. In order to help making a choice between the single, conditional and person-specific SEM, we end with a set of practical recommendations.

Practical recommendations

This chapter shows that there is not one type of SEM (i.e., single, conditional with different sizes of intervals or person-specific) that is superior in every test situation. Therefore, in order to choose one must first come to a realistic estimation of the ‘truth’ (i.e., how strong is the anticipated relationship between true score and error variance? How ‘unique’ do we expect the error variances of different examinees to be?) and a realistic idea of the limits and possibilities of the test (i.e., how reliable is the test for the examinees that you have/had in mind? How many parallel or tau-equivalent parts

K are feasible, taking for instance the item difficulties and item content into account?). Preferably, this endeavor is followed by a small simulation to find the optimal balance between bias and variance in the error variance estimate(s). Alternatively, one could use the following rules of thumb, based on the results of our simulation:

- When a test is based on a limited set of items and/or can only be divided in a limited set of parallel (or tau-equivalent) test parts K , the gain in (less) bias does not make up for the loss (increase in) estimation variance. In such a situation, either use the single SEM or a conditional SEM in which you make coarse intervals. When you do not expect a relationship between ‘true’ score and error variance (Figure (2.5c)), selecting the single SEM is most appropriate in this situation.
- When the overall reliability of the test for a group of examinees is low, be cautious in the interpretation of the conditional SEM. The single SEM is preferred in this case, except when you expect a strong relationship between true score and true error variance.
- Opting for a person-specific SEM is only encouraged when you either have continuous parallel test scores plus a large number of items and a large K (see Figure (2.16)) or when it is possible to test an examinee multiple times (see Hu et al., 2016).
- Be sure to have at least three items in every parallel test part if your test consists of rounded and truncated scores, to minimize the influence of truncation and rounding. Where possible, check the limitations that rounding and truncation put on the possible between variances (see Figure (2.11)) and see whether taking an average over similar examinees (conditional SEM) solves these limitations.
- When in doubt, choose a conditional SEM with coarse intervals (i.e., “con.+3”). Especially when you anticipate a relationship between ‘true’ score and error variance, the conditional SEM with coarse intervals is able to capture the main trend whilst being a fairly stable estimate.
- When a test is based on a limited set of items and/or can only be divided in a limited set of parallel (or tau-equivalent) test parts K , the gain in (less) bias does not make up for the loss (increase in) estimation variance. In such a situation, either use the single SEM or a conditional SEM in which you make coarse intervals. When you do not expect a relationship between ‘true’ score and error variance (Figure 2.5c), selecting the single SEM is most appropriate in this situation.
- When the overall reliability of the test for a group of examinees is low, be cautious in the interpretation of the conditional SEM. The single SEM is preferred in this case, except when you expect a strong relationship between true score and true error variance.
- Opting for a person-specific SEM is only encouraged when you either have continuous parallel test scores plus a large number of items and a large K (see Figure 2.16) or when it is possible to test an examinee multiple times (see Hu et al., 2016).

- Be sure to have at least three items in every parallel test part if your test consists of rounded and truncated scores, to minimize the influence of truncation and rounding. Where possible, check the limitations that rounding and truncation put on the possible between variances (see Figure 2.11) and see whether taking an average over similar examinees (conditional SEM) solves these limitations.
- When in doubt, choose a conditional SEM with coarse intervals (i.e., “con.+3”). Especially when you anticipate a relationship between ‘true’ score and error variance, the conditional SEM with coarse intervals is able to capture the main trend whilst being a fairly stable estimate.

Chapter 3

BAYESIAN APPROXIMATE MEASUREMENT INVARIANCE



Kimberley Lek, Daniel Oberski, Eldad Davidov, Jan Ciecuch, Daniel Seddig & Peter Schmidt

Based on a chapter published in OECD working paper No. 201, *Invariance Analyses in Large-scale Studies* and a chapter published in *Advances in Comparative Survey Methods: Multinational, Multiregional and Multicultural Contexts*, 2018

Chapter 3

Approximate measurement invariance

Introduction

When comparing data from different countries, time points, or groups, we run into at least two problems. First, we want to avoid large measurement artifacts that lead to erroneous substantive conclusions (Davidov et al., 2010, 2014). For example, when comparing Finnish to Columbian survey answers, we may want to account for any differences in exuberance. Second, we want to ignore the – likely plentiful – small measurement artifacts whose effect on substantive conclusions is negligible (Meuleman, 2012; Oberski, 2014). For example, when comparing Finns in 2002 with Finns in 2004 on an income question, most of the differences found are likely to be substantive; we would not want to spend an inordinate amount of time and modeling power on identifying all the small measurement differences between these already highly comparable groups. Tests for the presence or absence of measurement differences are typically called measurement invariance tests, sometimes also known as tests of differential item functioning (Holland & Wainer, 2012) or item bias (Mellenbergh, 1989; Shealy & Stout, 1993). Techniques to test for measurement invariance are numerous (Van De Schoot et al., 2015), but, for the purposes of this chapter, can be described as broadly falling into one of two categories: *exact* and *approximate*.

In the *exact* methods (see Vandenberg & Lance, 2000; Vandenberg, 2002; Brown, 2015), the researcher looks for a measurement model in which any small measurement differences are assumed to be exactly zero, while large differences are left completely free to be estimated from the data (termed partial measurement invariance; Byrne, Shavelson & Muthén, 1989). Methods to establish the fit of such models include chi-square difference testing (Steenkamp & Baumgartner, 1998), comparative fit index (CFI), root mean square error of approximation (RMSEA), and other fit measure comparisons (Cheung & Rensvold, 2002; Chen, 2007) and examination of local fit measures such as modification indices (MI) and the expected parameter changes (EPC; Byrne, Shavelson & Muthén, 1989), or the EPC of interest (Oberski, Vermunt & Moors, 2015). One way or another, all of these methods ultimately aim to find a model that balances two strategies, namely, accounting for large measurement differences while ignoring the small ones.

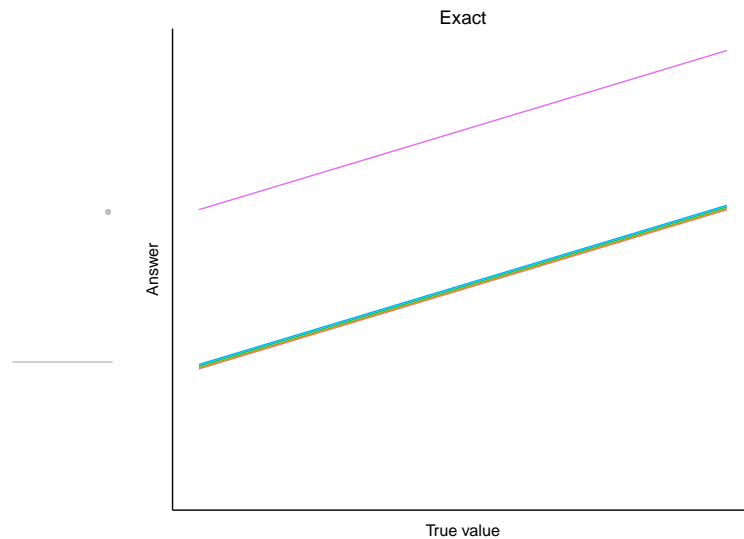
An alternative to the family of exact methods, and the focus of this chapter, is the *approximate* approach. In this approximate measurement invariance model, large and

small differences alike are assumed to follow a known distribution of nonzero values. Random effects distributions (Fox & Verhagen, 2010), multilevel models (Davidov et al., 2012, 2016), and strong Bayesian priors (Muthén & Asparouhov, 2013; Van De Schoot et al., 2013) have all been used for this purpose. The idea in all of these techniques is that any smaller differences are automatically accounted for in the model. Thus, approximate measurement invariance is primarily designed to deal with the second strategy — that of ignoring small differences automatically. The first strategy – dealing with large measurement artifacts – is problematic, although several existing proposals are discussed at the end of this chapter.

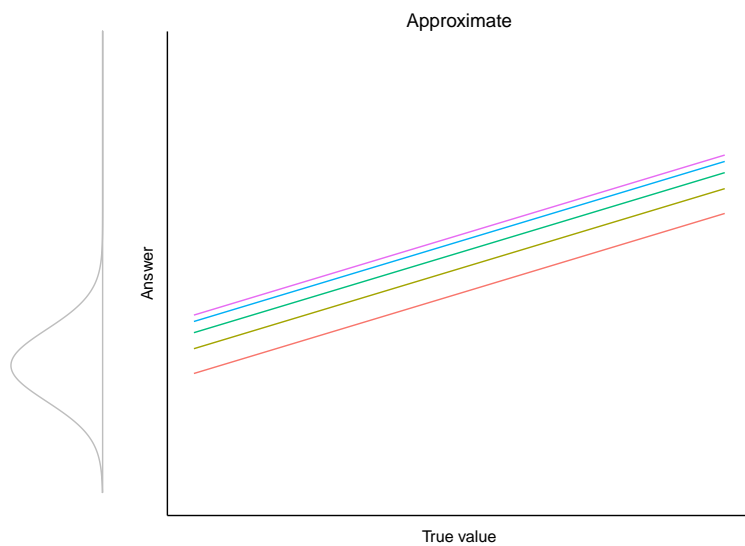
According to the advocates of approximate measurement invariance, exact zero constraints are overly strict, especially when there are many groups or time points involved (e.g., Davidov et al., 2015). One consequence is a frequent rejection of the exact invariance model, even when the parameter differences are ignorable (i.e., the second strategy). Another consequence is often a large series of model modifications that may capitalize on chance (MacCallum et al., 1992). In approximate measurement invariance, small differences in parameters are allowed. Moreover, the mind-boggling search through all possible combinations of measurement restrictions is replaced by a relatively simple estimation procedure. With many groups and measurement parameters, this practical advantage is considerable. For example, even in the simplest testing setup, a 10-factor analysis of 21 items over 19 countries (e.g. Davidov et al., 2008; Davidov 2008, 2010) yields 380 possible univariate violations of intercept equalities alone. The number of models resulting from all possible combinations of equality restrictions on intercepts and loadings is in the tens of millions. The corresponding approximate measurement invariance model aims to allow for measurement differences in these models by parameterizing them and imposing zero mean and small variance distributions in a more manageable procedure.

Figure 3.1 illustrates the difference between the exact (a) and approximate models (b). Each graph shows the theoretical, unobserved, true value to be measured on the horizontal axis and the obtained survey answer on the vertical axis. The lines thus correspond to the answers given by respondents with a particular true value: the response functions. These response functions may differ in intercept over groups to be compared (colors); if so, the same answer (point on the vertical axis) given by respondents from different groups (colored lines), could correspond to very different true values (corresponding points on the horizontal axis). Thus, comparing answers from these groups will compare not only true value differences but also differences in the intercepts of their response functions. If the true value differences are of the same order of magnitude as these measurement artifacts, the differences should be accounted for to prevent bias in the comparisons. Likewise, the slope of the response function may differ across groups (i.e. the loading of the survey item onto the latent factor; Brown, 2015). To keep matters simple, this chapter focusses on differences in intercepts only.

Figure 3.1a demonstrates the exact model: Most of the lines are held equal, while others (one in the example) are allowed to differ by any amount. How much it will differ is estimated from the data without restriction. The distribution of lines, shown in gray on the vertical axis, consists of a spike and a dot, since all intercepts are assumed equal except for some, which can differ by any amount. Figure 3.1b illustrates the corresponding approximate model. All lines are allowed to differ here, turning the spike into a normal distribution. This means that the lines that differed somewhat from the average are now



(a) Exact



(b) Approximate

Figure 3.1: Response functions (lines) for different groups (colors) under exact (a) vs. approximate (b) measurement invariance models.

allowed to differ by some amount. How much they will differ is determined in part by the data and in part by the restriction that the difference follows a normal distribution, shown on the vertical axis. This also implies that the pink group, which was estimated to differ considerably from the others in the exact model, is now pulled strongly toward the average by this prior. In other words, the strategy for allowing for small measurement differences is accomplished but traded off against reduced detection of large measurement differences.

The multigroup confirmatory factor analysis

This chapter discusses the use of measurement invariance testing as illustrated in Figure 3.1 in latent variable measurement models. In such models, the response functions are

estimated through presumed conditional independence assumptions, and investigation of measurement invariance proceeds through restrictions on the parameters of these estimated functions. The most common model for this test is the confirmatory factor model, but this framework also includes Item Response Theory (IRT) models, latent class models, and generalized multitrait-multimethod models (see Davidov et al. 2014). To simplify the discussion, we will limit ourselves to a multigroup confirmatory factor analysis (MGCFA) here.

Given a survey response y_{igj} for respondent i , group g , and item j , a MGCFA measurement model is

$$y_{igj} = \tau_{gj} + \lambda_{gj}\eta_{igj} + \epsilon_{igj}, \quad (3.1)$$

where

- η_{igj} is the unobserved true value (latent variable) for respondent i ;
- ϵ_{igj} is the unobserved measurement error value (latent variable) for respondent i ;
- τ_{gj} is the group-specific intercept for item j ;
- λ_{gj} is the group-specific loading (slope) for item j .

Measurement invariance then imposes cross-group restrictions on respectively the item structure (“configural invariance”), the factor loadings (“metric invariance”) and the intercepts (“scalar invariance”; Billiet, 2003; Millsap, 2011). Exact scalar invariance as in Figure 3.1a (for all groups except for the pink group), for example, may imply $\tau_{blue,j} = \tau_{green,j} = \tau_{yellow,j} = \tau_{red,j} \neq \tau_{pink,j}$. Since the intercept of the pink group ($\tau_{pink,j}$) is allowed to differ from the other groups, we speak of ‘partial’ rather than ‘full’ measurement invariance (Byrne, Shavelson & Muthén, 1989). We can test a similar assumption for the slopes, though we will simplify matters here by limiting ourselves to intercept differences, as in Figure 3.1, and assuming that all slopes are equal in the data. Approximate measurement invariance suggests that the intercept differences follow a certain probability distribution, often normal (Gaussian):

$$\tau_{gj} - \tau_{g'j} \sim N(0, \sigma_j) \quad (3.2)$$

for all differing pairs of groups $g \neq g'$. This distribution corresponds to the distribution of differences shown on the vertical axis of Figure 3.1b. As in the exact procedure, on average intercept differences are expected to be zero. Differences may vary, however, and the standard deviation of these differences for item j is denoted here as σ_j . When σ_j is estimated from the data, a random effect (Verhagen & Fox, 2010) or a multilevel model (Davidov et al. 2012; Jak et al. 2014a, 2014b) results. When it is fixed in advance by the researcher, a Bayesian approximate measurement invariance model results (Muthén & Asparouhov, 2013). An important question is how large the typical difference σ_j should be to appropriately balance the two strategies of measurement invariance analysis: accounting for the large measurement differences while ignoring the small ones.

In the remainder of this chapter, we will focus on a practical analysis of the Bayesian approximate measurement invariance model using standard software. The following section contains a worked example. We then discuss some of the outstanding pitfalls and issues with this technique in the discussion and conclusion section.

Illustration

For this illustration, we have simulated a simple dataset (dataset 1) consisting of continuous variables y_1 - y_4 , each believed to measure a certain continuous latent construct f_1 . Two groups are created, consisting of 500 respondents each. Mplus (Muthén & Muthén, 1998-2016) is used to apply the approximate measurement invariance testing procedure to this data. Together with the R package Blavaan (Merkle & Rosseel, 2016), Mplus is currently the only software package that allows you to test for approximate measurement invariance. The Mplus (Version 7.4) input file that is used to simulate the data can be found in Figure 3.2. Notice that the intercept differences are relatively small (0.1 vs. -0.1) and cancel each other out between as well as within groups. The latent mean difference between groups 1 and 2 is 0.5 (i.e. 0 in group 1 and 0.5 in group 2).

Using the MGCFA chi-square difference test procedure to test for measurement invariance (Vandenberg & Lance, 2000) – which is the default in Mplus - one would conclude that exact measurement invariance does not hold in dataset 1. This can be seen in Figure 3.3, which shows that the chi-square difference test of scalar versus metric equivalence is statistically significant ($\alpha = .05$). Since chi-square tests are known to be sensitive to sample size and violations of the normality assumption (Brannick, 1995), some authors (e.g., Brown, 2015; Chen, 2007) have suggested to take into account commonly used fit indices such as the comparative fit index (CFI; Bentler & Bonett, 1980) and the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) in the judgment of measurement invariance. Following the guidelines of Chen (2007; p. 501), also based on the CFI and RMSEA differences, we would conclude that scalar invariance does not hold (Δ CFI \geq -0.01; Δ RMSEA \geq 0.015, Table 3.1). Ignoring the absence of scalar invariance leads to an underestimation of the f_1 mean difference between groups 1 and 2 (i.e., 0.399 instead of 0.500).

Table 3.1: RMSEA and CFI differences between the configural, metric and scalar model.

	Configural	Metric	Scalar
CFI	0.991	0.995	0.873
RMSEA	0.047	0.025	0.112
	(0.00 - 0.092)	(0.00 - 0.064)	(0.088 - 0.137)

Instead of forcing the differences in intercepts to be exactly zero, we could opt for approximate measurement invariance by using the Mplus input file depicted in Figure 3.4 (based on Muthén & Asparouhov, 2013 and Van De Schoot et al., 2013b). This input file is a special application of Bayesian structural equation modeling (BSEM) in which strict zero constraints are replaced by probability distributions with zero mean and small variance (see Muthén & Asparouhov, 2012; Van Erp, Mulder & Oberski, 2018). These probability distributions are called “priors” in the Bayesian terminology.

Montecarlo:

```
names = y1-y4;
ngroups = 2;
nobs = 500 500;
nreps = 1;
save = dataset 1.dat;
```

Model montecarlo:

```
f1 by y1@0.7 y2@0.6 y3@0.5 y4@0.4;
      y1@0.51 y2@0.64 y3@0.75 y4@0.84;
! 1 - factor loading^2
```

```
[y1@-0.1 y2@0.1 y3@-0.1 y4@0.1];
```

```
[f1@0];
```

```
f1@1;
```

Model montecarlo-g2:**! group 2**

```
[y1@0.1 y2@-0.1 y3@0.1 y4@-0.1];
```

```
[f1@0.5];
```

Figure 3.2: Mplus input file containing the population parameter values for the intercepts, factor loadings, latent means and latent variances.

Models Compared	Chi-square	Degrees of	
		Freedom	P-value
Metric against Configural	0.797	3	0.8502
Scalar against Configural	63.928	6	0.0000
Scalar against Metric	63.131	3	0.0000

Figure 3.3: Mplus output of the MGCFA chi-square comparisons. The scalar equivalence model fits significantly worse than the metric equivalence model, hence exact measurement equivalence does not hold.

The prior distributions are confronted with the data, reflected in the “likelihood”, to come to a “posterior” distribution that is essentially a compromise of the prior and the likelihood (for a more thorough discussion of Bayesian statistics, see e.g. Gelman et al., 2003; Kaplan & Depaoli, 2013; Kruschke et al., 2012; Lee, 2007). Thus, when we place a small variance prior with zero mean on the difference in intercepts, the posterior balances model fit on the one hand (i.e., the likelihood) and measurement invariance restrictions (i.e., the prior) on the other. The smaller the prior variance, the more the posterior will be influenced by the prior measurement invariance restrictions. The key part of the input file in Figure 3.4 is

```
MODEL: [y1 - y4] (nu#_1 - nu#_4)
MODEL PRIOR: DO(1,4) DIFF (nu1_# - nu2_#) ~ N(0,0.01)
```

; the part where the small variance prior is specified. Let us disentangle this part of the input file step by step, beginning with the last comment “N(0,0.01)”. This comment shows that in this input file, the prior follows a normal distribution with mean zero and variance 0.01. Remember that the choice of variance is important, since it is the variance that determines the wiggle room we allow in the intercept estimates of groups 1 and 2 (see Equation (3.2); Van De Schoot, Kluytmans & Tummers, 2013). In front of “N(0,0.01)” we find the statement “DIFF (nu1_#-nu2_#)”. Because of this “DIFF” statement, we place the small variance prior on the difference in intercepts between group 1 and group 2, instead of on the intercepts themselves. “nu1_#-nu2_#” between brackets are the labels referring to the intercepts for y1 to y4 in group 1 and group 2. These labels were attached previously to the intercepts of y1 to y4, under the “MODEL” statement. Since labeling can be cumbersome, in this input file automatic labeling is applied. Specifically, the # in the labels is automatically replaced by the number of the item (1, 2, 3, or 4). As such, Mplus automatically places the small variance prior on the difference between nu1_1 (the intercept of y1 in group 1) and nu2_1 (the intercept of y1 in group 2), the difference between nu1_2 and nu2_2, the difference between nu1_3 and nu2_3, and the difference between nu1_4 and nu2_4. Finally, the “DO(1,4)” comment makes sure Mplus

correctly replaces the # of the automatic labeling by 1, 2, 3, and 4.

When we run the input file of Figure 3.4, posterior draws of the parameters are generated over and over again in each iteration of the Bayesian algorithm. As an illustration, Figure 3.5 shows the posterior draws of iteration 1-20 for the intercept of y1 in group 1 (left) and group 2 (right). Posterior draws for groups 1 and 2 in a specific iteration are connected by a line. The steepness of this line – i.e. the difference between y1 in groups 1 and 2 – is restricted by the prior we have specified. If the parameters were equal in each draw, the lines would be horizontal; the steeper the lines, the larger the intercept differences between groups in each posterior draw. As can be seen in the Figure 3.5, these differences in each posterior draw are present but modest – exactly as the Gaussian prior on these differences stipulates.

When the Bayesian algorithm is completed, we first need to check whether this algorithm has converged to the appropriate posterior (see Depaoli & Van De Schoot, 2015). In Mplus, convergence can be assessed visually, by looking at the traceplot for every parameter in the model, and statistically, by checking the potential scale reduction factor which should be close to 1 (PSR, Gelman, 1996). Mplus stops the Bayesian algorithm when the PSR drops below $1 + \epsilon$ with a default ϵ between 0.05 and 1 for most of the models²³ (Asparouhov & Muthén, 2010). We choose a more stringent stopping rule by specifying `BCONVERGENCE = .01`; (Figure 3.5). We additionally force Mplus to run at least 100,000 iterations by specifying `BITERATIONS = (100000)`; . Mplus informs us that convergence has been reached according to the adjusted PSR criterion (`THE MODEL ESTIMATION TERMINATED NORMALLY`). Based on the traceplots of the intercepts, we would also conclude that the algorithm has converged (Figure 3.6), allowing us to turn to the Mplus output.

A part of the Mplus output resulting from the input in Figure 3.4 is shown in Figure 3.7. Notice first that most of the fit indices usually provided by Mplus (RMSEA, CFI) are not available anymore. To judge whether our Bayesian approximate measurement invariance model fits our data, we rely on a likelihood ratio test (LRT) between the approximate measurement invariance model and an unrestricted mean and (co)variance model (Asparouhov & Muthén, 2010). Specifically, in every iteration Mplus conducts two LRTs using the current parameter estimates. The first of these LRTs, (1), evaluates the fit between the current model and the original data. The second one, (2), confronts the current model with a newly generated dataset, simulated on the basis of the estimated model. This latter one shows LRT chi-square values that can reasonably be expected when approximate measurement invariance holds. Chi-square values of (1) that are systematically higher than those of (2) are an indication of model misfit. To determine whether this is the case, we can either look at the PPP-value (Gelman et al., 1996) or the 95% credibility interval provided in the Mplus output (Figure 3.6). The PPP expresses the proportion of chi-square values obtained with (2) that exceed (1). PPP-values around 0.5 are indicative of good model fit, and low PPP-values close to zero should be avoided. In this case, we would be fairly satisfied with a PPP-value of 0.269 (Table 3.2), although a PPP closer to 0.5 would be preferable. The 95% credibility interval is determined for the distribution of differences between (1) and (2). When (1) is not systematically higher than (2), zero is included in this 95% credibility interval, which is fortunately the case in the present example. Turning to the estimates (Table 3.2), we see that the intercepts of the two groups are estimated in line with their true values (Figure 3.2) but are generally pulled closer to zero. The `DIFFERENCE` section of the output shows the

```

DATA:      FILE = "dataset 1.dat";

VARIABLE: NAMES ARE y1-y4 group;
              KNOWNCLASS IS g(group=1 group=2);
              CLASSES ARE g(2);

ANALYSIS: TYPE = MIXTURE;
              ESTIMATOR = BAYES;
              MODEL = ALLFREE;

              BCONVERGENCE = .01;
              BITERATIONS = 500000(100000);
              bseed = 123;

MODEL:

%OVERALL%
f1 by y1 y2 y3 y4 (lam#_1-lam#_4);
[y1-y4] (nu#_1-nu#_4);

%G#1%
[f1@0];
f1@1;

%G#2%
[f1];
f1@1;

MODEL PRIOR:
DO(1,4) DIFF (lam1_#-lam2_#) ~ N(0, .01);
DO(1,4) DIFF (nu1_#-nu2_#) ~ N(0, .01);

```

Knownclass is used to describe the grouping variable; needed when "type is mixture" is specified in the analysis command

MODEL = ALLFREE is needed for DIFF and automatic labeling with # (see MODEL statement)

Stricter convergence guidelines than default to reduce any bias due to precision

Labeling; the # makes sure labels are automatically specified for group 1 and 2

DO(1,4) loop applies the DIFF statement to all 4 variables. DIFF statement is used to place a prior on the differences in intercepts and factor loadings.

Figure 3.4: Input file in Mplus for the Bayesian approximate measurement equivalence test.

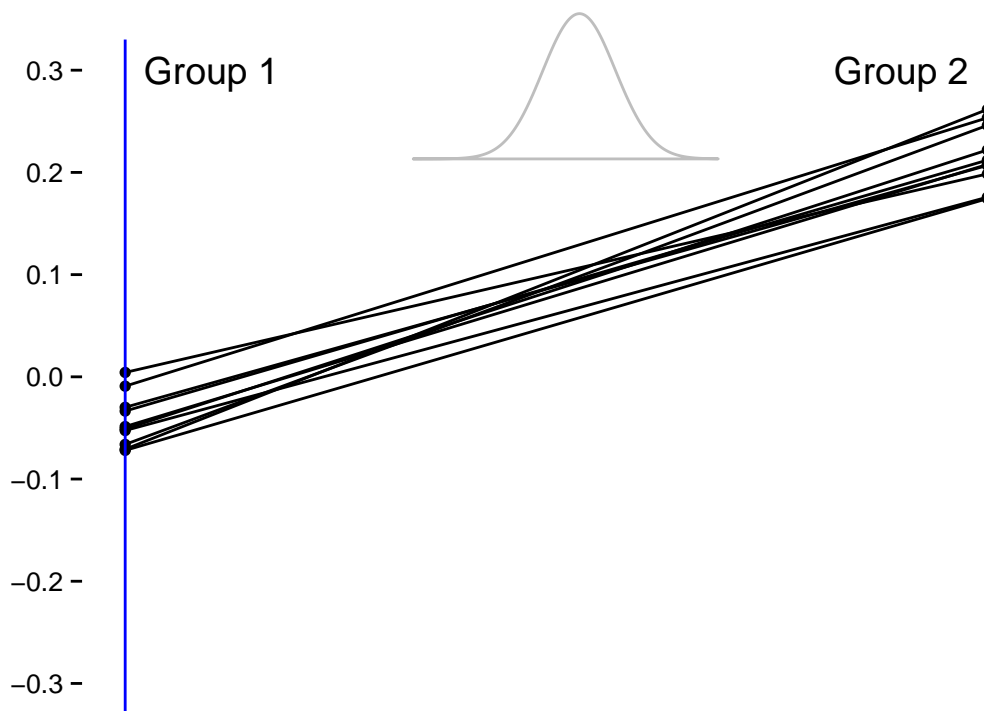


Figure 3.5: Visualization of the estimation of the intercept y_1 in groups 1 and groups 2.

mean intercept across groups and the amount by which every group-specific intercept deviates from this value. The latent mean difference is estimated to be 0.477, reasonably close to the true difference of 0.5. In this example, we initially allowed a prior variance of 0.01, taking into account the scale of the y_1 - y_4 variables. Since the choice for a suitable prior variance is crucial to the Bayesian approximate measurement procedure, it is good practice to perform a sensitivity analysis with multiple plausible prior variances, as displayed in Table 3.2 (Van De Schoot & Depaoli, 2014). In this way, it is possible to carefully balance model fit (i.e. PPP, 95% confidence interval) and the possibility to compare groups (i.e. keeping the prior variance as small as possible). When we increase the prior variance to 0.05 in this example, the PPP-value moves closer to 0.5, and the 95% credibility interval becomes more symmetric around zero. However, increasing the prior variance also enlarges the standard errors of the intercepts and the latent mean difference estimate. The resulting latent mean difference estimate (0.457) is slightly worse than the one we obtained with prior variance 0.01 (0.477). Increasing the prior variance to 0.1 does not yield a further improvement of the PPP / 95% credibility interval and only changes the parameter estimates slightly. Therefore, a prior variance of 0.01 or 0.05 seems the best choice here.

Altogether, Bayesian approximate measurement invariance seems to largely solve the problem of exact scalar noninvariance (Figure 3.3) in dataset 1. Indeed, Bayesian approximate measurement invariance is suggested to be useful in situations in which there are many small parameter differences that cancel each other out both within and between groups (De Boeck, 2008; Muthén & Asparouhov, 2013; Van De Schoot et al., 2013b; Wolvers & Lugtig, 2016). What if the differences between intercepts become

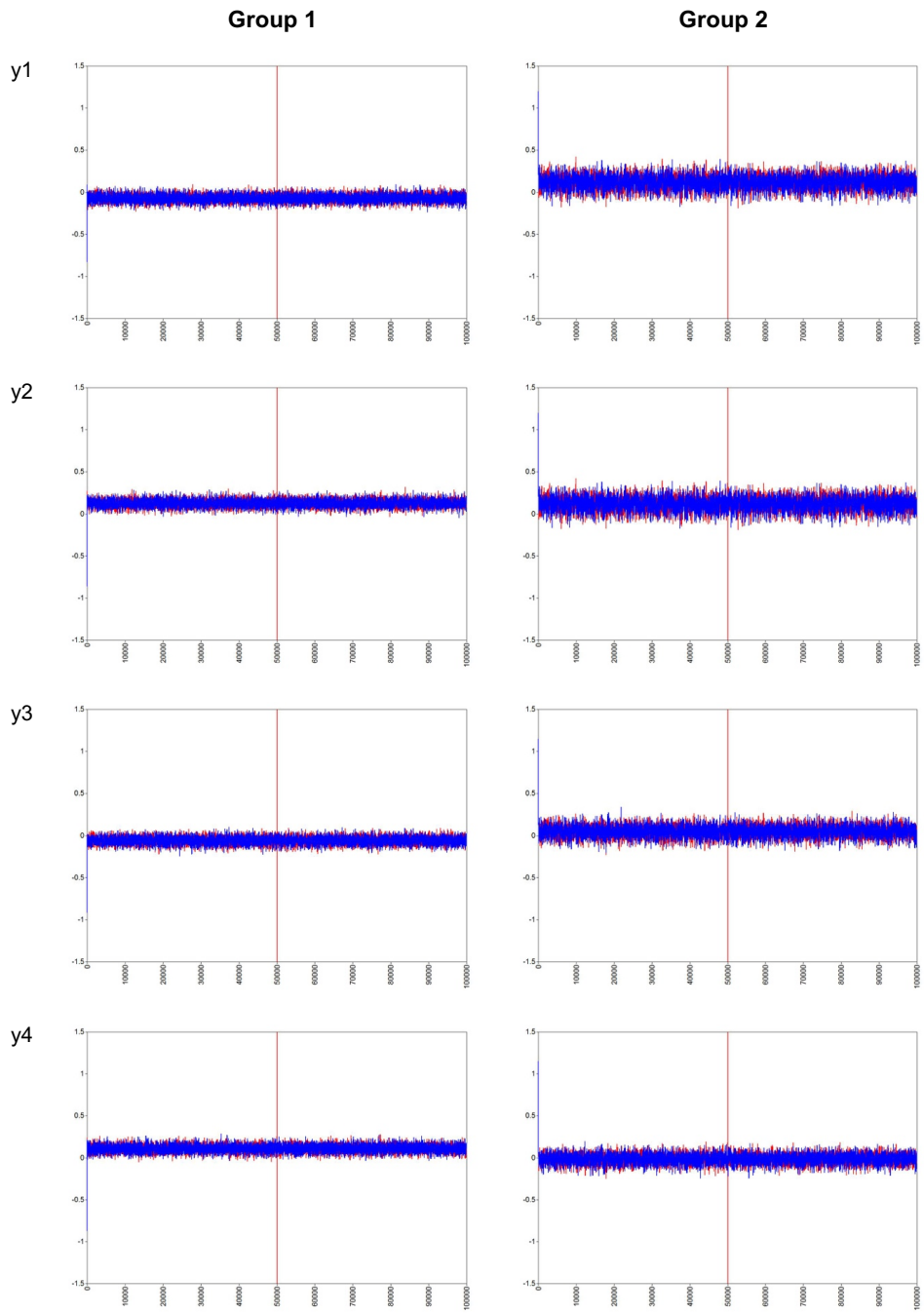


Figure 3.6: Traceplots to judge the convergence of intercept y_1 - y_4 in group 1 and 2. Note that only the last 50,000 (after the vertical line) are used for the parameter estimates.

MODEL FIT INFORMATION

Bayesian Posterior Predictive Checking using Chi-Square

95% Confidence Interval for the Difference Between
the Observed and the Replicated Chi-Square Values

-14.418 27.098

Posterior Predictive P-Value 0.269

MODEL RESULTS

	Estimate	Posterior S.D.	One-Tailed P-Value	95% C.I.		
Significance				Lower 2.5%	Upper 2.5%	
Latent Class 1 (1)						
Means						
F1	0.000	0.000	1.000	0.000	0.000	
Intercepts						
Y1	-0.069	0.044	0.056	-0.155	0.017	
Y2	0.127	0.045	0.002	0.039	0.215	*
Y3	-0.058	0.044	0.095	-0.145	0.029	
Y4	0.112	0.044	0.006	0.025	0.197	*

(a)

Latent Class 2 (2)						
Means						
F1	0.477	0.122	0.000	0.242	0.721	*
Intercepts						
Y1	0.121	0.080	0.069	-0.041	0.272	
Y2	-0.084	0.072	0.121	-0.227	0.056	
Y3	0.053	0.066	0.212	-0.079	0.177	
Y4	-0.011	0.056	0.421	-0.125	0.096	
DIFFERENCE OUTPUT						
			NU1_1	NU2_1		
5	0.026	0.053	-0.095*	0.095*		
			NU1_2	NU2_2		
6	0.022	0.049	0.105*	-0.105*		
			NU1_3	NU2_3		
7	-0.003	0.046	-0.055	0.055		
			NU1_4	NU2_4		
8	0.050	0.041	0.062*	-0.062*		

(b)

Figure 3.7: Part of the Mplus output resulting from the input file in Figure 3.4.

larger? Or if the differences between the groups are systematic (i.e., do not cancel each other out within groups)? To check the performance of Bayesian approximate measurement invariance in these situations, we altered the intercept values of dataset 1 in the way described in Table 3.3.

Regardless of prior variance choice, when intercept differences are systematic, the intercept estimates are no longer in line with their true values. With a prior variance of 0.01, the latent mean difference estimate, 0.789, is too high. Interestingly, the PPP fails to detect the misfit ($PPP = 0.368$). As stated by Muthén and Asparouhov (2013), recovery of parameters is not expected when the noninvariance is not in line with BSEM. Enlarging the intercept differences as in the first row of Table 3.3 leads to a PPP-value of 0.000 with prior variance 0.01. Increasing the prior variance to 0.05 yields a PPP-value of 0.186 and a latent mean difference estimate of 0.642. Increasing the prior variance even further to 0.1 changes the PPP to 0.278 and a more acceptable latent mean difference estimate of 0.566. In sum, when differences are systematic or relatively large, one should be cautious in applying the approximate measurement testing procedure.

Discussion and Conclusion

The increasing availability of large cross-cultural and cross-country surveys in the last several decades has significantly increased the possibilities for researchers to conduct comparative studies. However, they have also considerably increased the risk of drawing wrong conclusions that researchers may run into. Therefore, the methodological literature on cross-cultural and cross-country analysis has recommended testing for measurement equivalence to guarantee that differences across groups are due to substantive true differences and not methodological artifacts. This recommendation has been increasingly applied by researchers, who tested for the measurement equivalence properties of various scales (e.g. Cieciuch et al. 2014; for an overview, see Davidov et al. 2014). Unfortunately, a new problem has come up, namely, that many scales failed to display high levels of equivalence. In this chapter we have discussed approximate measurement invariance as a possible solution to this problem. Instead of restricting the differences between all measurement parameters (i.e., factor loadings, intercepts) to be exactly zero, approximate measurement invariance assumes that these differences follow a (normal) distribution with mean zero and small variance σ_j . This variance σ_j can either be estimated from the data (Verhagen & Fox, 2010; Davidov, 2012) or be fixed in advance by the researcher (Asparouhov & Muthén, 2013). The latter is known as Bayesian approximate measurement invariance and is illustrated in this chapter with standard software. Approximate measurement invariance seems especially advantageous when the number of groups or repeated measurements is large, there are many small differences in intercepts and factor loadings, and differences cancel each other out both within and between groups (Muthén & Asparouhov, 2013, Van De Schoot et al., 2013; Wolvers & Lugtig, 2016). Exact measurement invariance almost never holds in this scenario and is cumbersome to test. When additionally there are some large differences in intercepts or factor loadings, approximate measurement invariance may not establish equivalence. The small variance prior tends to pull strongly deviating parameter estimates toward the average across groups and/or time points. The result is that the deviating parameter will be smaller, while the invariant parameters will be

Table 3.2: The influence of prior variance on parameter differences.

	$\sigma_j = 0.1$		$\sigma_j = 0.05$		$\sigma_j = 0.01$	
	Est (se)	G1 Est (se) G2	Est (se)	G1 Est (se) G2	Est (se)	G1 Est (se) G2
Intercepts						
y1	-.09 (.04)	.17 (.19)	-.08 (.04)	.16 (.14)	-.07 (.04)	.12 (.08)
y2	.15 (.05)	-.10 (.19)	.15 (.05)	-.10 (.14)	.13 (.05)	-.08 (.07)
y3	-.07 (.05)	.09 (.15)	-.07 (.05)	.09 (.11)	-.06 (.04)	.05 (.07)
y4	.13 (.05)	-.02 (.12)	.13 (.05)	-.02 (.09)	.11 (.04)	-.01 (.06)
Δf_1	.447 (.296)		.457 (.216)		.477 (.122)	
Model fit						
95% CI	-15.78	25.19	-15.86	24.80	-14.42	27.10
PPP-value	.322		.326		.269	

Table 3.3: Alteration of the intercept values of dataset 1.

Differences	Group 1				Group 2			
	y1	y2	y3	y4	y1	y2	y3	y4
large	-0.5	0.5	-0.5	0.5	0.5	-0.5	0.5	-0.5
systematic	-0.1	-0.1	-0.1	-0.1	0.1	0.1	0.1	0.1

larger than their true values (Muthén & Asparouhov, 2012). This leads to a considerable bias in the latent mean estimates (Van De Schoot et al., 2013). As illustrated in this chapter, bias may also result from systematic differences between groups. A promising solution to reduce bias is to combine approximate measurement invariance testing with the newly developed alignment procedure of Asparouhov and Muthén (2014). This alignment procedure rotates the solution in such a way that there are many invariant parameters and a few (large) noninvariant parameters, using the same principles as used in CFA (see Jennrich, 2006 for technical details; for an application see Cieciuch et al. 2018, Munck et al. 2017). Another solution is to free noninvariant parameters and only apply approximate measurement invariance to the remaining parameters (see Muthén & Asparouhov, 2013).

Several studies have already applied the approximate measurement invariance test (e.g. Davidov et al. 2015, 2017, Zercher et al. 2015). These studies have demonstrated that approximate equivalence may be given also when exact equivalence is rejected by the data. However, as Davidov et al. (2015) mentioned, it “does not do magic”; there is a point at which one must conclude that measurement invariance simply does not exist (Lommen, Van De Schoot & Engelhard, 2015). The key question is when exactly that point is reached. More research into this key question, the role of large deviating parameters, and the size of σ_j is necessary²⁴.

Acknowledgments

The authors thank Rens Van De Schoot for comments on an earlier version of this chapter. The work of Eldad Davidov and Jan Cieciuch was supported by the University Research Priority Program Social Networks, University of Zurich. The work of Peter Schmidt was supported by the Alexander von Humboldt Polish Honorary Research Fellowship granted by the Foundation for Polish Science. Kimberley Lek is funded with a talent grant from the Netherlands Organization for Scientific Research (NWO): NWO Talent 406-15-062. Daniel Oberski is supported by NWO Veni grant 451-14-017.



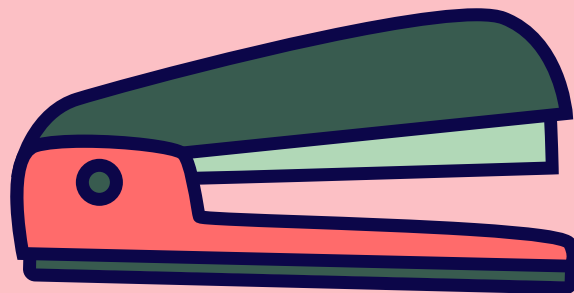
PART II

Teacher



Chapter 4

**DOES SOCIO-ECONOMIC STATUS,
ETHNICITY AND GENDER MATTER?
INVESTIGATION OF TEACHER
RECOMMENDATIONS IN THE
TRANSITION FROM PRIMARY TO
SECONDARY EDUCATION IN THE
NETHERLANDS**



Kimberley Lek, Trudie Schils & Remco Feskens

Chapter 4

Does socio-economic status, ethnicity and gender matter? Investigation of teacher recommendations in the transition from primary to secondary education in the Netherlands

Abstract

The present study investigated whether socio-economic status, ethnicity and gender influence teachers' track recommendations in the transition from primary to secondary education in the Netherlands. An average effect of socio-economic status and gender was found, with lower-SES pupils and girls being (slightly) disadvantaged. Schools appeared to vary greatly, however, in the extent to which these factors influence track recommendations. And although no average effect of ethnicity was found, the large variance component for this factor indicated that in some schools ethnicity does lead to either a profound advantage or disadvantage. The composition of the class (regarding SES, ethnicity and achievement) was found to influence the average track recommendation in classes, but was unable to explain differences in the relationship between SES, ethnicity and teacher recommendations over classes. A detailed examination of the teacher bias revealed that teacher bias was not generally of a statistical (based on achievement characteristics of the, for instance, ethnic group a pupil belongs to) or taste-based (a systematically over- or underestimation of certain groups of pupils) nature, underlining the need for a more detailed explanation of teacher bias.

Keywords: Teacher judgment, meritocracy, teacher bias, track allocation, track recommendations

Introduction

The transition from primary to secondary education is one of the most crucial and difficult transitions in a student's educational career (Zeedijk et al., 2003), as students who transition unsuccessfully risk unnecessary absenteeism, dropout, grade retention and runoff (Van Rens et al., 2018; Bosch et al., 2008). In many countries, during the transition students are sorted into multiple tracks. Since the track entered at the beginning of secondary education largely influences the rest of the pupil's educational career (Van Rooijen et al., 2017), track allocation needs to be performed with utmost care and attention. In some countries (e.g., Great Britain and the United States), track allocation is based on a standardized test, while in other countries (for instance, France, Germany, Luxembourg, Switzerland and Flanders [the northern part of Belgium]) allocation is largely or exclusively determined by the judgment of the teacher (see Boone & Van Houtte, 2012; Sneyers, Vanhoof & Mahieu, 2018). Despite the potential of such teacher judgments, a disadvantage is that these judgments are often a 'black-box'; the track recommendation is known but the process leading to this recommendation is not. Ever since the controversial Pygmalion study²⁵ (Rosenthal & Jacobson, 1968), researchers have attempted to investigate which (un)desirable factors affect teacher expectations, teacher judgments and teachers' track recommendations²⁶. In an early meta-analysis, Dusek and Joseph (1983) underlined the potential effects of pupils' socioeconomic status (SES), ethnicity and gender (among other demographic pupil characteristics). These pupil characteristics are interesting since pupils cannot change these characteristics while they do have the potential to influence pupils' school career. Since then, SES-, ethnicity, gender and similar teacher biases have been investigated in many real-life and experimental settings (see, for instance, Becker, Japel & Beck, 2013; Glock & Krolak-Schwerdt, 2013; Holder & Kessels, 2017; Rubie-Davis & Peterson, 2016; Van Den Bergh et al., 2010). Most of these studies have demonstrated a consistent effect of ethnicity and socio-economic status, with a smaller and more mixed effect for gender (see the systematic review of Wang, Rubie-Davies & Meissel, 2018).

Despite the vast body of literature on SES-, ethnicity- and gender-effects on teacher judgments, there are questions that still remain unanswered (Wang, Rubie-Davies & Meissel, 2018). In previous studies, the nature of the teacher bias is, for instance, often unclear. In Geven et al. (2018), a distinction is made between 'statistical' (rational) and 'taste-based' (irrational) bias. Taste-based bias reflects a situation in which the teacher systematically under- or over-recommends certain pupils based on their (for instance ethnic) group membership. Statistical bias occurs when the teacher takes into account existing achievement differences between (for instance ethnic) groups in the recommendation for an individual pupil who is part of this group. In this latter case, the recommendation might be inaccurate for an individual pupil, but the median recommendation matches the average achievement of the group. In the former case, the median recommendation is structurally higher or lower than the average group achievement would justify. Distinguishing between these two kinds of bias is important, as statistical and taste-based bias have different origins and ask for different interventions. Statistical bias, for instance, might occur when the teacher does not have enough information about specific pupils. Taste-based bias might be the result of stereotypes. Literature is also relatively limited with respect to the investigation of variation between schools/classes in teacher bias and possible school or class-level

factors (such as SES and ethnic composition) that might explain this variation (Wang, Rubie-Davies & Meissel, 2018; for an exception see Agirdag et al., 2013). Finally, from a methodological perspective, studies have often used a limited proxy for SES (only the highest attained educational level of parent(s) or the occupation of the father, for instance) and a notable portion of studies into teacher bias have failed to control for pupil's actual achievement, making it impossible to distinguish between teacher bias and manifest differences in achievement between different groups of pupils (Wang, Rubie-Davies & Meissel, 2018).

In the Netherlands, the sixth grade teacher became primarily responsible for pupils' track allocation after a policy change in 2015. Ever since the introduction of this new allocation policy, the discussion about possible teacher bias has gained renewed attention in the Netherlands (see Timmermans, De Boer, Amsing & Van Der Werf, 2018; Geven et al., 2018). Using the Dutch situation as a case study, the current article explores SES-, gender and ethnic biases in the context of the transition from primary to secondary education. The Netherlands offers an interesting framework for this exploration, due to the large variation in possible track allocations. To add to the existing knowledge base, we make a deliberate distinction between statistical and taste-based bias. We also investigate the variance in teacher bias over school classes and attempt to explain this variance by class-level factors such as the class' SES- and ethnic composition. Finally, we have access to the results of the Dutch 'End-of-primary-school (EPST) test to control for pupil achievement and to multiple indicators of SES (highest attained educational level of both parents, income of both parents and the estimated value of the house the pupil lives in). Taken together, we attempt to answer the following three questions in the present article:

1. do teacher recommendations systematically differ for pupils with a particular socio-economic status, gender and ethnicity, after the EPST-result is taken into account?
2. do primary schools systematically differ in their track recommendations and is it possible to explain these differences by the composition of the school (regarding ethnicity, socio-economic status and average EPST-result)?
3. if teacher recommendations differ systematically for pupils with a particular SES, gender and ethnicity, is this bias generally of a statistical ('rational') or taste-based ('irrational') nature?

Literature review

Focusing on research from the Netherlands, Luxembourg and Flanders²⁷ (the northern part of Belgium), multiple studies have indicated that a pupil's gender, socio-economic status (SES) and ethnicity might influence the judgment of the teacher (see, for instance, Timmermans, De Boer, Amsing & Van Der Werf, 2018). Especially SES is almost consistently found to influence track recommendations after performance on the EPST-test is taken into account (Geven et al., 2018). In Dutch literature, SES is often exclusively measured by the level of parental education. Timmermans, Kuyper and Van Der Werf (2015) for instance conclude that pupils from lower educated parents are

allocated to lower tracks than pupils with the same EPST-score with higher educated parents. According to the Dutch inspectorate of Education (2019), especially children with the lowest educated parents (at most ‘mbo-2’, see Figure 4.1) and without a migration background are at risk for under-recommendations (i.e., a systematically lower track recommendation from the teacher than the EPST). Other indicators and proxies also have been used in the literature to measure SES. The Flamish study by Boone and Van Houtte (2013), for instance, primarily looked at the parents’ occupation. Specifically, they classified the highest of the parents’ occupations following Erikson, Goldthorpe and Portocarero (1979) and then reclassified these into the four broader categories ‘working class’, ‘lower middle class’, ‘middle class’ and ‘upper middle class’. Based on this classification, they concluded that teacher recommendations were partly determined by pupils’ socio-economic backgrounds, even when prior scholastic achievement of the pupil were taken into account. Some studies investigate a combination of socio-economic characteristics. Feron, Schils and Ter Weel (2015), for instance, assessed the educational level of the father and mother (lower education, vocational education, higher education and university), the labor market position of the father and mother (employed, unemployed, sick/unable to work) and the number of workdays for the father. Based on these characteristics, they concluded that Dutch pupils from families with lower socio-economic status are less likely to receive a teacher assessment that is higher than the EPST-test score. Another example is the ISEI-scale (see Ganzeboom, De Graaf & Treiman, 1992) which relates parents’ education, occupation and income. Using this scale, Sneyers, Vanhoof and Mahieu (2018) find a small to moderate effect of parent’s socio-economic status on Flamish pupil’s track recommendations.

According to the longitudinal study by Timmermans et al. (2018), other than SES, the influence of ethnicity largely decreased between the 90s and 2014²⁸. In the 1980s and 1990s, minority children (mostly Turkish and Moroccan) were typically over-recommended in the Netherlands, meaning that their teacher’s track recommendations exceeded those of the EPST-test (Driessen et al., 2008; Driessen, 1991). According to Claassen and Mulder (2003), this over-recommending reverted to under-recommending around 2000. After that, different ethnic groups seemed to receive comparable educational recommendations given equal school performance (Driessen et al., 2008). In agreement with the conclusion by Timmermans et al. (2018), according to the inspectorate of Education in the Netherlands (2019), ethnicity nowadays indeed plays a lesser role in track allocation. Other studies, however, still report a disadvantage for non-native pupils (see Baysu, Alanya & Van De Valk, 2018; Scheerens & Van Der Werf, 2018). As for SES, gender biases have also been found consistently in the literature, although the effect of gender is generally small. Of the 18 studies considering gender biases in teacher evaluations reviewed in Geven et al. (2018), 12 for instance suggest that teacher evaluate girls more positively than boys. In other countries, the gender bias also has been shown repeatedly (see, for instance, Cornwell, Mustard & Van Parys, 2011; Mizala, Martinez & Martinez, 2015).

Although the effect of SES, ethnicity and gender on teachers’ track recommendations has repeatedly been found, there are two caveats. First, as briefly mentioned in the introduction, the potential bias of teacher recommendations is not universal. According to Geven et al. (2018) and Timmermans et al. (2015), there are large variations across Dutch primary schools in the extent to which track recommendations are influenced by students’ demographic traits. The composition of the school/class may play a role in this

school-specific bias. When classes are ‘mixed’ (containing multiple groups of ethnically similar students) teachers may base their expectations on general group characteristics rather than on individual characteristics, due to directly perceived contrast (Timmermans et al., 2015). Furthermore, it is well-known that the judgment of individual pupils can be influenced by the level of the other pupils in the class (Driessen et al., 2008). This is called the ‘frog-pond’ effect (see Davis, 1966), where pupils in classes with a high cognitive level receive a relatively lower educational recommendation.

Second, in the aforementioned studies no distinction is made between ‘statistical’ and ‘taste-based’ bias. According to Geven et al. (2018), statistical bias occurs when the teacher bases his or her recommendation of a certain pupil on the average performance of the social or ethnic group to which (s)he belongs (also see Becker, 2010). The average group level is thus used as a proxy for the pupil’s potential, with the advantage that the decision at the group-level is accurate (i.e., on average the recommendation for the pupils is correct). On the positive side, teachers may in this way correct for the phenomenon of ‘regression towards the mean’ (Borghans, Diris & Schils, 2018). According to this phenomenon, when a pupil’s EPST-score is exceptionally low or high for his social/ethnic group, it is probable that his EPST-score is closer to the group average upon retesting. By giving a teacher recommendation that is closer to the social/ethnic group average than the EPST-result, this possible over- or underachievement on the EPST-test is corrected. On the negative side, taking into account the group average might lead to inaccurate and unfair decisions for specific individuals belonging to that group. Taste-based bias is based on the personal preferences of teachers and occurs when teachers evaluate pupils more or less favorably when they belong to a certain group. Whereas the (un)fairness of statistical bias is debated, taste-based bias is per definition unfair.

Taken together, a pupil’s socio-economic status and - to a lesser degree - gender and ethnicity might partly determine the track advice (s)he receives from the teacher. The degree of teacher bias and the type of bias (statistical versus taste-based) might, however, vary over classes and schools.

The current study

The Dutch context

In the Netherlands, pupils typically transition from primary to secondary education at the age of 12, after the sixth grade of primary education. Since the policy change in 2015, the sixth grade teacher formulates a track recommendation for secondary education. As a ‘second-opinion’, an End-of-Primary-school test (EPST-test) is administered after the teachers’ judgment. Based on this EPST-result, the teacher judgment can be adjusted to a higher track, if the teacher believes this adjustment is justified (see <https://www.vanponaarvo.nl/>). General Dutch secondary education is divided into three main hierarchical tracks (see Figure 4.1): ‘vmbo’ (preparatory secondary vocational education; the lower track), ‘havo’ (senior secondary general education; the intermediate track) and ‘vwo’ (pre-university education; the higher track). Sometimes, combinations of the three tracks are offered by the secondary school as well (‘vmbo/havo’ and ‘havo/vwo’), mainly in the first year of secondary education. As each of the three hierarchical tracks differ in standards and qualifications, each of them leads to different

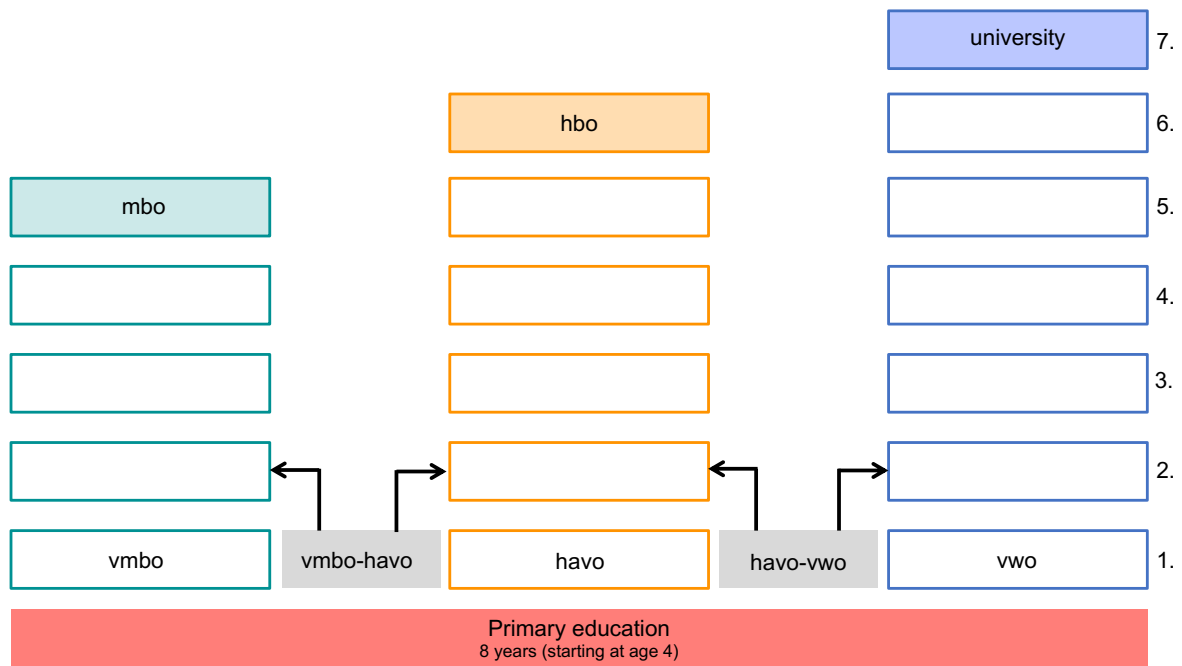


Figure 4.1: Overview of the Dutch educational system.

subsequent options for education. With a vwo-diploma, pupils can enter university (first a bachelor, then a master and eventually a PhD), whereas a havo-diploma provides entrance into universities of applied sciences ('hbo'; first a bachelor, then a master) and a vmbo-diploma into a senior secondary vocational education ('mbo'), subdivided into mbo1 (entry training), mbo2 (basic vocational training), mbo3 (professional training) and mbo4 (middle management training). Notice that despite the fact that the teacher recommendation is binding, the Dutch educational system has some flexibility and permeability between the displayed tracks.

Data

For the purpose of our research, we were authorised to perform analyses on microdata files in the CBS ('Statistics Netherlands') remote secured environment²⁹ (see Bakker, Van Rooijen & Van Toor, 2014). The CBS tested our results on identifiability before publication. One of the microdata files we used was the so-called 'CITOTab'. At the request of the College for Tests and Exams, the Cito provided the CBS with data of the 2014-2015 End of Primary School Test (EPST). The data comprise the results of the EPST as well as the teacher's recommendations. The CBS creates the CITOTab file based on these data each year. They anonymise the data and assign RIN identification numbers that enable links with other microdata. We combined the CITOTab file with seven other microdata files, including the KINDOUDERTab that made it possible to link the RIN identification numbers of pupils and their parents.

Primary schools can authorise Cito to share the EPST results with CBS. In 2015, 5,082 institutions for primary education did this for 163,794 pupils. Data about these pupils can be consulted in the CBS CITOTab file of 2015. We did not include all of these institutions and pupils in our study. We first selected 123,033 pupils based on the following two

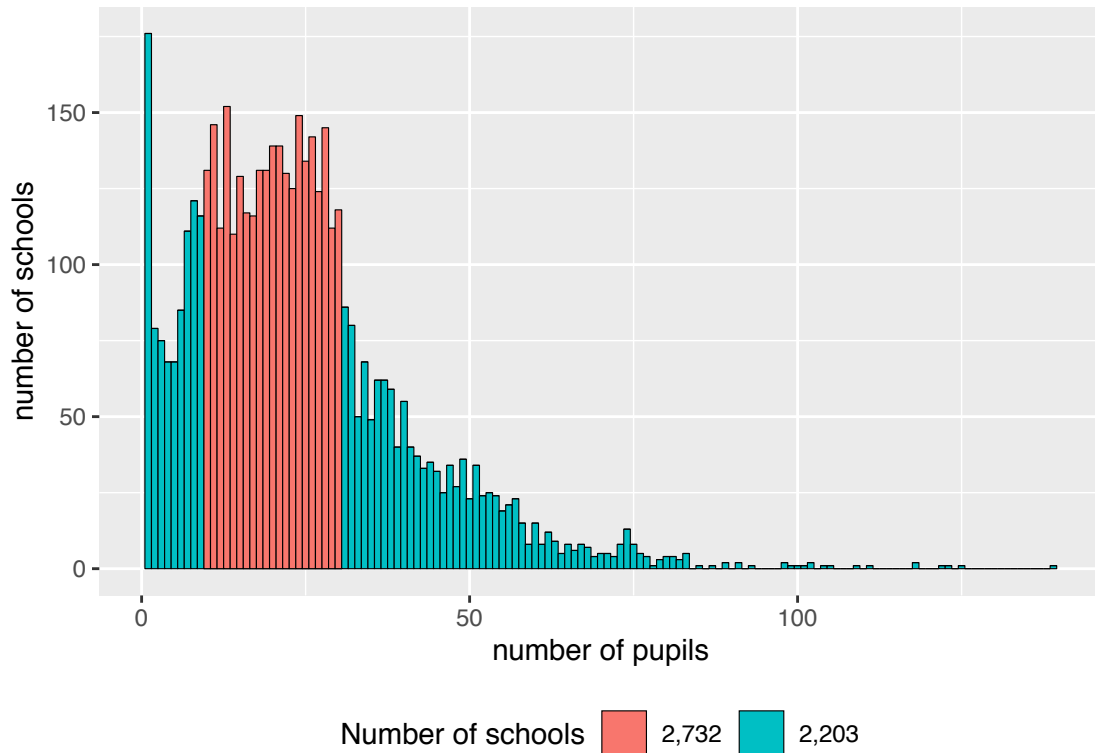


Figure 4.2: Overview of the selection of the 2,732 out of 4,935 schools with 10 to 30 sixth grade pupils.

criteria: (1) pupils occurred in the ‘Basisregistratie Personen’ (BRP; the Municipal Personal Records Database) and (2) both the teacher’s advice and the EPST-result in school year 2014-2015 were available for these pupils. We also applied some selection criteria on the level of the institutions. First, we selected the 4,935 institutions (containing 118,748 pupils) for primary education with only one location (i.e., one primary school). We took this selection step to avoid a mix of institution-level (for the institutions with more than one location) and school-level (for the institutions with only one location) effects. Unless stated otherwise, all figures and descriptive statistics in this article are based on these 118,748 pupils. For the multilevel analyses, we went one step further by subsequently selecting the 2,732 schools (containing 54,650 pupils) with a minimum of 10 and a maximum of 30 sixth grade pupils in 2014-2015. The selection of schools with a maximum of 30 pupils was conducted in an attempt to avoid a mix of school-level (for the schools with more than one sixth grade class) and class-level (for the schools with only one sixth grade class) effects. As most Dutch classes do not contain more than 30 pupils, we used 30 pupils as a cut-off between schools with probably one sixth grade class and schools with probably more sixth grade classes, as information on the number of sixth grade classes per school was lacking. We required classes with a minimum of 10 pupils to avoid estimation problems³⁰. Figure 4.2 shows a histogram of the number of pupils per school location; the red bars are the schools that were selected.

Table 4.1: Variable overview.

Variable	Scale	mean(sd)/percentage
Individual-level variables		
Highest attained educational level parents	< mbo2	11.17%
	mbo2, mbo3 or mbo4	30.11%
	hbo bachelor	27.97%
	hbo master or wo	30.75%
Average income	negative	0.81%
	<= 50,000	60.37%
	> 50,000	38.82%
WOZ-value	WOZ-value / 100,000	2.57 (2.02)
Gender	boy	49.79%
	girl	50.21%
Ethnicity	native Dutch	77.89%
	Moroccan	4.04%
	Turkish	3.09%
	Miscellaneous	14.98%
Teacher recommendation	vmbo	44.07%
	vmbo/havo	8.80%
	havo	18.92%
	havo/vwo	9.23%
	vwo	18.99%
EPST-score, language	0-13.50	9.76 (1.86)
EPST-score, mathematics	0-8.50	6.09 (1.50)
EPST recommendation	vmbo	36.06%
	vmbo/havo	13.77%
	havo	11.30%
	havo/vwo	18.80%
	vwo	20.10%
Class-level variables		
Average WOZ-value	average WOZ/100,000	2.46 (0.97)
Proportion \leq hbo bachelor		0.72 (min = .04, max = 1)
Proportion \neq hoogste inkomen		0.64 (min = 0, max = 1)
Proportion non-native Dutch pupils		0.23 (min = 0, max = 1)
Average EPST-score, language		9.68 (0.80)
Average EPST-score, mathematics		6.03 (0.60)

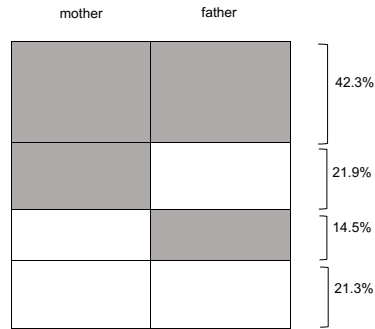
(In)dependent variables

Table 4.1 provides an overview of the (in)dependent variables included in the present study. Below, each of the variables is discussed in more detail.

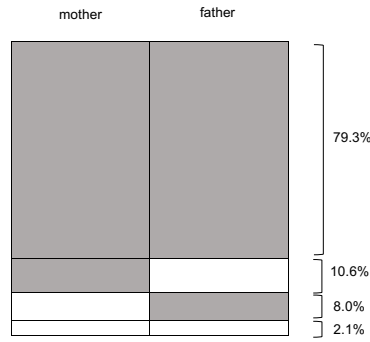
1. *SES*. As an indication of Socio-economic status (SES), we use three variables: 1) the highest attained educational level of the parents, 2) the average annual income of the parents and 3) the WOZ-value of the house the pupil is currently living in. Below, each of these SES-indicators are discussed in more detail. As Figure 4.3 illustrates, all the SES-indicators contain at least some missing values. As is done in for instance Chmielewski (2017), the missingness in the SES-indicators is dealt with using multiple imputation by iterative chained equations (see Van Buuren & Groothuis-Oudshoorn, 2011). Van Buuren's R package 'MICE' is used for this purpose, creating five imputation datasets. Only the SES-variables of Figure 4.3 are included in the imputation process, using 'predictive mean matching' for all these variables as the imputation technique (Little, 1988). Note that all subsequent

analyses in this article (see “Analytical approach”) are conducted on each of the imputed datasets separately and then pooled using Rubin’s rule to take into account variability due to the imputation process. In this study, the five imputed dataset yielded similar results.

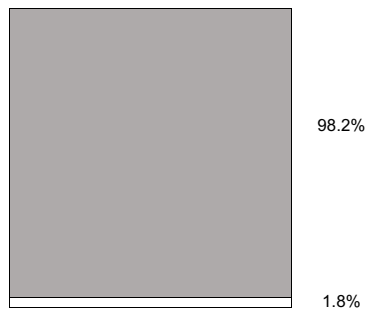
- *Highest attained educational level parents.* For the highest attained educational level of the father and mother (in 2015), we used the CBS microdata file ‘HOOGSTEOPLtab’. For those parents with a diploma of a government funded institution, information is automatically stored in ‘HOOGSTEOPLtab’ (i.e., it is register data). This information, however, is not collected automatically for private and foreign institutions and for (long) business trainings. Therefore, CBS collects data on these types of education on a sampling basis (see Lindler, Van Roon & Bakker, 2011). Because of the selectivity of the microdata file, 2% of the people registered in HOOGSTEOPLtab actually have a higher attained educational level than reflected by the originally collected register data of the CBS. CBS corrects for this underestimation by modifying the attainment variable, using imputation techniques. The resulting attainment variable consists of 18 possible educational levels, running from primary education (1) to PhD (18). After the multiple imputation (see above), we first divided the 18 possible educational levels of the mother and of the father into four broader categories (see Table 4.1): (1) lower than mbo2, (2) mbo2, mbo3 or mbo4, (3) hbo bachelor, (4) hbo master or university (bachelor, master or PhD). Thereafter, we compared the highest attained educational level of father and mother and selected the highest of the two. The resulting variable was recoded into three dummy variables with (4) hbo master or university (bachelor, master or PhD) as the reference category.
- *Average income.* For every parent, the various sources of income (income because of employment, entrepreneurship, government subvention) were accumulated over the period 01-01-2015 until 31-12-2015. Then, we averaged the income for the father and the mother of every pupil. Note that this (average) income can be negative or equal to zero for the entrepreneurs and self-employed parents. This negative or zero income can be misleading, as parents with such a negative income are not necessarily poorer than parents with a positive income over 2015. It might be, for instance, that the parent(s) invested money from their own capital to start a new business and did not make profit yet. To account for this negative income phenomenon, the variable ‘average income’ was divided in three categories including ‘negative income’. The other two categories are ‘ $\leq 50,000$ ’ (low to medium income) and ‘ $> 50,000$ ’ (high income) (see Table 4.1). The resulting variable was recoded into two dummy variables with ‘ $> 50,000$ ’ as the reference category.
- *WOZ-value.* The WOZ-value is the amount of money a house is currently worth according to an annual evaluation of the municipality. We used the WOZ-value of the house the pupil is currently living in and divided this value by 100,000. For the multilevel analyses (see section ‘Analytical approach’) We subsequently subtracted the average WOZ-value over all pupils from the absolute WOZ-values, to create a centered variable. Note that Table 4.1 shows the average WOZ and standard deviation of the uncentered variable.



(a) Highest attained educational level



(b) Annual income



(c) WOZ-value

Figure 4.3: Overview of missing data in the three SES-variables. A grey colored box denotes that the value for that variable is present in the data.

2. *Gender.* Gender was collected by the ‘GBApersoonstab’ for all pupils, with value 0 corresponding to boys and 1 to girls. As Table 4.1 shows, 50.21% of the pupils were boys.
3. *Ethnicity.* To describe someone’s ethnicity, CBS advises to use two variables of the ‘GBApersoonstab’. The first variable describes whether a person has a native Dutch background (0), a first generation migration background (not born in the Netherlands; 1) or a second generation migration background (at least one of the parents not born in the Netherlands; 2). The second variable lists the country a person is connected to based on the country of birth of his parents or himself. Someone with a score 0 on the previous variable, has ‘the Netherlands’ as value on the second variable. For someone with a score 1 or 2, the country is respectively the country in which the pupil is born or the country in which the mother is born (unless this is the Netherlands, in that case it is the birth country of the father). The large list of countrycodes of the second variable was recoded into different ethnical groups (i.e., native Dutch, Moroccan, Turkish, Surinamese, Dutch Antilles and Aruba, miscellaneous (non-)Western countries). For our purposes, we then combined the generation variable and the country variable in a variable with the following categories: 1) native Dutch, 2) Moroccan, first or second generation, 3) Turkish, first or second generation, 4) Miscellaneous (non-)Western countries, first and second generation. Note that this last category is rather heterogeneous and is therefore only included for completeness. By taking into account the immigration background of pupils on top of their nationality, we avoid the risk that effects of ethnicity that last longer than one generation are masked (i.e., in other studies all second generation immigrant are sometimes coded as Dutch; Hachfeld et al., 2010). The resulting ethnicity variable was recoded into dummy variables, with “1) native Dutch” as the reference category.
4. *Teacher recommendation.* In the CITOTab file, the teacher recommendations are available as they were written down on the EPST-form. Note that these recommendations were formulated and written down *before* the results of the EPST-test were known. We recoded these teacher recommendations into five different options: 1) vmbo³¹, 2) vmbo-havo, 3) havo, 4) havo/vwo, and 5) vwo.
5. *EPST-score.* the EPST-result consist of a language part and a mathematics part. In this study, we include the EPST-score for language and mathematics separately. The EPST-score for language is the number of items correct out of 135 for this part. For mathematics, the EPST-score is the number of items correct out of 85. Both language and mathematics scores are divided by 10 and centered by subtracting the mean mathematics and language scores over all pupils in the study. Note that Table 4.1 shows the mean and standard deviation of the uncentered scores.
6. *EPST-recommendation.* Based on the language and mathematics part (see “EPST-score”), a composite EPST-score is calculated by Cito, running from 501-550. Cito divides the score range 501-550 into five non-overlapping score-intervals, corresponding to the five track possibilities: 501-532 (vmbo), 533-536 (vmbo-havo), 537-539 (havo), 540-544 (havo-vwo), 545-550 (vwo).
7. *SES composition: highest attained educational level parents.* For every included school (see “Data”), we calculated the proportion of pupils of which the parents

had at most hbo bachelor as their highest attained educational level. A value of 0 thus corresponds to a class with only pupils of parents who completed a hbo master, university bachelor, master or PhD.

8. *SES composition: average income.* For every included school (see “Data”), we calculated the proportion of pupils of which the parent’s average income was less than 50,000. A value of 0 thus reflects a class with only relatively rich pupils.
9. *SES composition: WOZ-value.* We averaged the centered WOZ-value of the pupils in all included schools (see “Data”). A value of 0 thus indicates a class with pupils living in ‘average-priced houses’ on average.
10. *Ethnic composition.* Following Thys and Van Houtte (2016), we calculate the proportion non-native pupils as an indication of the ethnic composition of the school. A score of 0 indicates a class with solely native Dutch pupils; a score of 1 a class without any native Dutch pupils.
11. *EPST-score composition.* Based on the individual pupil’s EPST-scores, we determined the average EPST-score for every school included in the dataset. We did this for the centered mathematics and centered language part separately.

Analytical approach

To investigate the influence of SES, gender, ethnicity and class composition on teacher recommendations, hierarchical regression analysis was conducted taking into account the EPST-score of pupils (Finch, Bolin & Kelley, 2014). Following Driessen, Slegers and Smit (2008), teacher recommendation was treated as an interval variable, with (1) vmbo, (2) vmbo/havo, (3) havo, (4) havo/vwo and (5) vwo (see Table 4.1). Seven hierarchical models were run, following the recommendations of Hox et al. (2017): 1) null-model, 2) model with only pupil-level, fixed effects of socio-economic status (highest attained educational level, average income, average WOZ-value), 3) model with pupil-level, fixed effects of gender and ethnicity added, 4) model with pupil-level, fixed effects of the EPST-scores for mathematics and language included, 5) model with aggregate class-level effects included, 6) model which allows random effects of socio-economic status, gender and ethnicity over classes and 7) model with cross-level interactions included to explain differences between classes. To compare these models, we based ourselves on the conditional R^2 -measure of Nakagawa, Johnson and Schielzeth (R_c^2 ; 2017; also see Johnson, 2014), expressing the variance explained by the entire hierarchical model. When systematic differences in teacher recommendations were found for pupils with different demographic characteristics, these differences were subsequently inspected to distinguish between statistical and taste-based bias (see Geven et al., 2018). In all analyses, the emphasis is placed on the strength of associations and the interpretation of coefficients rather than significance. The rationale for this is that with a large number of pupils and a large number of primary schools, associations with little or no relevance may quickly reach significance (Driessen, Slegers & Smit, 2008). Models are fitted using the R package ‘lme4’ (Bolker, version 1.21). The R package ‘MuMIn’ (Bartón, version 1.43.6) is used to estimate R_c^2 and the package ‘merTools’ (Knowles, Frederick & Whitworth, version 0.5.0) to fit the hierarchical models to all imputed datasets and extract the random effect statistics.

Results

Teacher recommendations

The results of the seven hierarchical models are shown in Table 4.2 and 4.3. As the null-model (**Model 1**) indicates, much of the variance in teacher recommendations is located at the pupil instead of the class-level (ICC = 9.2% of the variance on class level). **Model 2** shows that socio-economic status, as indicated by the highest attained educational level of parents, their average income and the WOZ-value of the pupil's house, influences the teacher recommendation. Specifically, pupils with the lowest educated parents on average receive a recommendation that is almost one track lower than the pupils with the highest educated parents (-0.95; see Table 4.2). Pupils with (slightly) higher educated parents (mbo2-4, hbo bachelor), also receive a track recommendation that is on average lower than the track recommendation of pupils with the highest educated parents (respectively 0.81 and 0.26 track 'points' lower; see Table 4.2). A lower than 50,000 average income also negatively impacts the average track recommendation: pupils from parents with a relatively high salary receive a track recommendation that is 0.19 (income \leq 50,000) or 0.15 (negative income) 'track points' higher on average. Parents with negative income over 2015, however, comprise a relatively mixed group, which explains the relatively large standard error for the "negative income versus $>$ 50,000" effect. Additionally, every 100,000 euro a house is worth extra, increases the average track recommendation by about 0.08 'track point'. That this effect is relatively small might be because of some extremely high WOZ-values in our dataset. To investigate the effect of WOZ-value further, Figure 4.4 illustrates the effect of WOZ-value (divided by 100,000) on the five track recommendations of the teacher. This Figure shows a clear, general decrease in the lower track recommendations between the relatively low WOZ-values and WOZ-values around 500,000 to 700,000 euros. After that, the number of lower track recommendations seem to slightly increase/stabilize. Note that the last bar contains all pupils with WOZ-values over 1,000,000 euros, to avoid bars that are based on few pupils. Adding the fixed effect of the SES-indicators to the model (model 2) explains about 13% of the total variance (according to Nakagawa, Johnson and Schielzeth, 2017; see Table 4.2).

When a pupil has a Moroccan or Turkish first or second generation background, his average track recommendation goes down by about 0.10 and 0.35 'track points', respectively (see **Model 3**, Table 4.2)³². In a model with socio-economic status and ethnicity included, gender shows no influence on the average track recommendation. Compared to model 2, model 3 does not lead to a further increase of the explained variances (see Table 4.3).

Model 2 and 3 indicate that track recommendations are generally lower for pupils with a lower SES and a non-Dutch ethnicity. However, in these models it is unclear to which extend these different track recommendations reflect 'true' achievement differences. Therefore, **Model 4** includes the EPST-score for language and for math. After inclusion of these EPST-scores, most coefficients drop. Especially the effect of ethnicity becomes much smaller. The effect of the highest education level of parents also becomes smaller, but the effect is still relatively large. Taking into account the EPST language and math scores, having low-educated parents still reduces the average track recommendation by about a quarter (-0.23 'track points' for <mbo2, -0.26 'track points' for mbo2-4; see

Table 4.2: Fixed effects of the seven hierarchical models.

	Models						
	1	2	3	4	5	6	7
Intercept	2.39 (.01)	3.00 (.02)	3.01 (.02)	2.71 (.01)	3.47 (.05)	3.25 (.05)	3.39 (.06)
Pupil level							
Highest education parents							
<mbo2		-0.95 (.02)	-0.92 (.02)	-0.23 (.02)	-0.21 (.02)	-0.20 (.02)	-0.43 (.08)
mbo2-4		-0.81 (.02)	-0.81 (.02)	-0.26 (.01)	-0.24 (.01)	-0.22 (.01)	-0.44 (.08)
hbo bachelor		-0.26 (.01)	-0.34 (.02)	-0.14 (.01)	-0.12 (.01)	-0.11 (.01)	-0.24 (.05)
Average income parents							
negative		-0.15 (.07)	-0.14 (.07)	-0.06 (.05)	-0.06 (.05)	-0.05 (.01)	-0.05 (.05)
<= 50,000		-0.19 (.01)	-0.19 (.01)	-0.08 (.01)	-0.06 (.01)	-0.05 (.01)	-0.05 (.01)
WOZ-value		0.08 (.01)	0.05 (<.01)	0.02 (<.01)	0.01 (<.01)	0.02 (<.01)	0.02 (<.01)
Gender			-0.01 (.01)	-0.10 (.01)	-0.10 (.01)	-0.10 (.01)	-0.10 (.01)
Ethnicity							
Moroccan			-0.10 (.03)	0.01 (.02)	-0.02 (.02)	<.01 (.03)	<-0.01 (.06)
Turkish			-0.35 (.04)	-0.02 (.03)	-0.05 (.03)	-0.02 (.03)	0.02 (.06)
Other			-0.01 (.01)	0.03 (.01)	0.02 (.01)	0.02 (.01)	0.02 (.01)
EPST, language				0.39 (<.01)	0.39 (<.01)	0.40 (<.01)	0.40 (<.01)
EPST, math				0.35 (<.01)	0.36 (<.01)	0.37 (<.01)	0.37 (<.01)
Class level							
prop.< hbo master					-0.88 (.06)	-0.69 (.06)	-0.90 (.08)
prop.< 50,000					-0.34 (.06)	-0.31 (.05)	-0.31 (.06)
Average WOZ-value					.05 (.01)	0.05 (.01)	0.04 (.01)
prop. non-native					0.21 (.03)	0.24 (.03)	0.23 (.03)
Average language					-0.16 (.02)	-0.20 (.02)	-0.20 (.02)
Average math					-0.20 (.02)	-0.25 (.02)	-0.25 (.02)
Cross-level interactions							
<mbo2 * prop.< hbo master							0.32 (.11)
mbo2-4 * prop.< hbo master							0.32 (.08)
hbo bachelor * prop.< hbo master							0.20 (.08)
Moroccan * prop. non-native							0.00 (.08)
Turkish * prop. non-native							-0.07 (.08)

Table 4.3: Random effects and R_c^2 of the seven hierarchical models.

	Models						
	1	2	3	4	5	6	7
σ_e^2	2.16	2.05	2.04	0.90	0.90	0.83	0.83
$\sigma_{u,0}^2$	0.22	0.11	0.11	0.14	0.08	0.13	0.13
$\sqrt{\sigma_{u,language}^2}$						0.08	0.08
$\sqrt{\sigma_{u,math}^2}$						0.08	0.08
$\sqrt{\sigma_{u,WOZ}^2}$						0.02	0.02
$\sqrt{\sigma_{u,Moroccan}^2}$						0.21	0.20
$\sqrt{\sigma_{u,Turkish}^2}$						0.17	0.17
$\sqrt{\sigma_{u,Other}^2}$						0.09	0.09
$\sqrt{\sigma_{u,gender}^2}$						0.12	0.13
$\sqrt{\sigma_{u,<mbo2}^2}$						0.14	0.15
$\sqrt{\sigma_{u,mbo2-4}^2}$						0.16	0.17
$\sqrt{\sigma_{u,hbobachelor}^2}$						0.16	0.16
$\sqrt{\sigma_{u,negative}^2}$						0.17	0.17
$\sqrt{\sigma_{u,<=50,000}^2}$						0.08	0.08
R_c^2		.13	.13	.64	.62	.65	.65

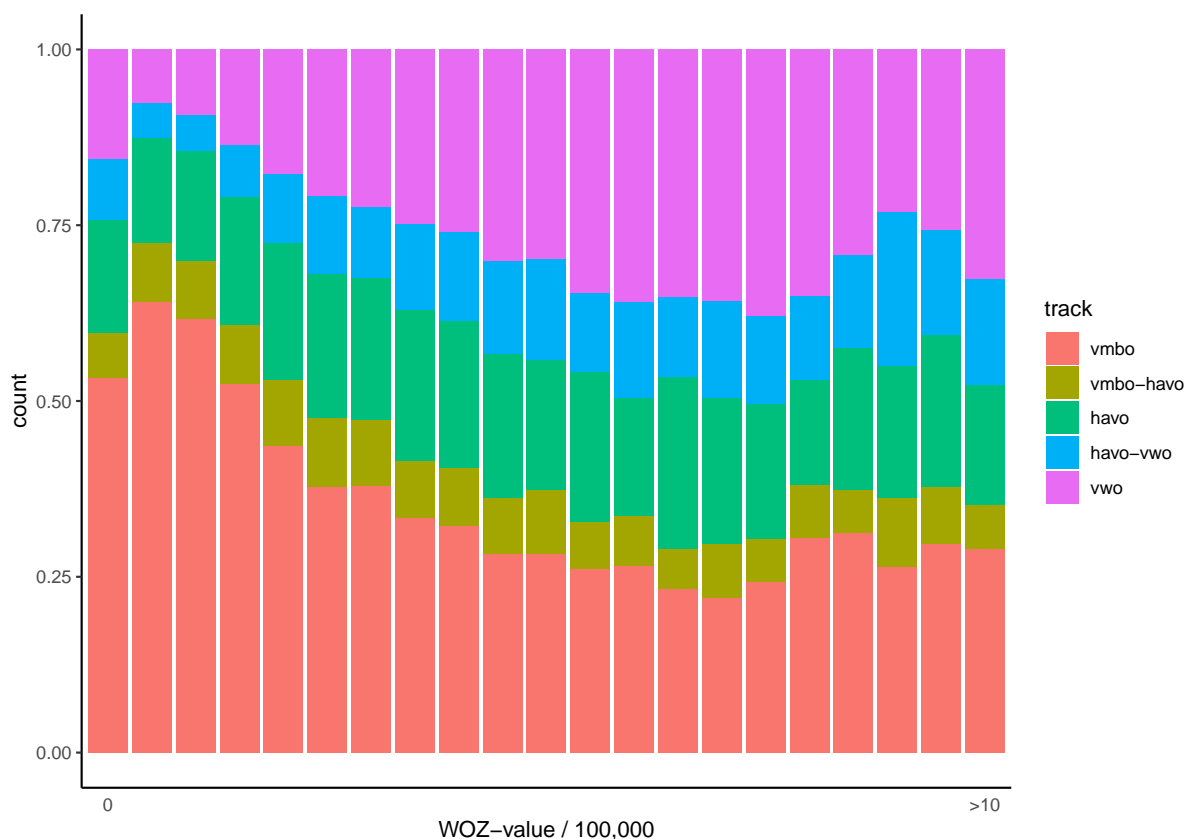


Figure 4.4: Relationship between WOZ-value/100,000 and teachers track recommendations. Note that the last bar is based on all WOZ-values above $10 \cdot 100,000$.

Table 4.2). Interestingly, the effect of gender *increases* (-0.01 to -0.10). When taking their actual achievement into account, girls thus receive lower track recommendations on average than boys. As Table 4.3 shows, including the EPST-scores we are able to explain 64% of the total variance.

The models so far have only considered fixed effect, pupil-level variables. **Model 5** shows that aggregate, class-level variables also generally influence the average teacher recommendation in a class. In a class with only highly educated parents (\geq hbo master), for instance, the average teacher recommendation is almost one track higher than in a class with only parents without such an educational background (-0.88; see Table 4.2). And in classes where all pupil's parents have an average income below 50,000, the average track recommendation is 0.34 'track points' lower than in classes where all pupil's parents earn over 50,000 on average. Being in a class with only non-native Dutch pupils generally leads to an advantage: a track recommendation of 0.21 'track points' higher, compared to a class with only native Dutch pupils. Finally, being in a class with high average EPST-scores generally leads to a lower track recommendation.

In **Model 6**, the pupil-level effects are allowed to vary over classes. Especially interesting are the variation in ethnicity effects (Moroccan versus native Dutch, Turkish versus native Dutch and 'other' versus native Dutch) and variation in the effect of highest attained educational level of the parents on teacher recommendations (see Table 4.3). Whereas the fixed effect of ethnicity largely disappeared after including achievement scores (see model 4), classes seem to vary greatly to the extent that ethnicity influences the track recommendation. As reflected by the standard deviation for the "Moroccan versus native Dutch" and "Turkish versus native Dutch" effects over classes (respectively .21 and .17; see Table 4.3), part of the classes notably favor non-Dutch over Dutch pupils whereas in other classes the opposite is true. As shown by the standard deviation for the "< mbo2" and "mbo2 - mbo4" versus hbo bachelor or higher effects (.14 and .16 respectively), in some classes the effect of parent's educational level is more extreme. Another part of the classes has effects closer to zero or even an effect in the opposite direction.

Model 7 attempts to explain some of the random variation in the relationship between pupil-level variables and teacher recommendations by including cross-level interactions. Specifically, the proportion of relatively low-educated parents (i.e., proportion of parents with less than a hbo master as the highest attained level of education) is included in an attempt to explain the variance in the relationship between 'having parents with less than mbo2 as their highest attained level of education' and teacher recommendation and in the relationship between 'having parents with mbo2, mbo3 or mbo4 as their highest attained level of education' and teacher recommendations (see Table 4.2). The resulting coefficients (both .32; Table 4.2) imply that the relationship between '<mbo2-' or 'mbo2-4-parents' and teacher recommendations on average changes by +0.32 'track points' if the pupil is in a class with *only* relatively low-educated parents. This value '0.32' is difficult to interpret without taking into account the direct effects of the dummy variables '<mbo2| \geq hbo master' and 'mbo2-4| \geq hbo master' and the class-level variable 'proportion parents with education < hbo master' (see Table 4.2). Specifically, the predicted teacher recommendation for a pupil with the lowest educated parents (< mbo2) in a class with only relatively low educated parents (< hbo bachelor) is $-0.43 + -0.90 + 0.32 = -1.01$ (see Table 4.2, model 7), keeping all other variables constant at zero³³. The predicted teacher recommendation for a pupil with the lowest educated parents in a class

without any other pupils with relatively low educated parents is $-0.43 + (\frac{1}{n_j})-0.90 + (\frac{1}{n_j})0.32$, where n_j is the class size of class j . In a class with 10-30 pupils, the resulting predicted teacher recommendation varies between -0.49 and -0.45. Comparing these values to the -1.01 indicates that being in a class with only relatively low-educated parents when the pupil's parents have the lowest attained educational level might have a negative effect on teacher recommendations. This cross-level effect, however, should be interpreted with caution as the coefficient is relatively small compared to the accompanying standard errors (respectively .11 and .08, see Table 4.2). Additionally, the random variances in the effects of 'mbo2' and 'mbo2, 3 or 4' on teacher recommendations do not drop as a result of including the cross-level effect. Finally, adding the cross-level effect does not add to the proportion of explained variance (i.e., R_c^2 of model 6 and 7 is .65).

Another cross-level effect that is included in model 7 is aimed at explaining the relatively large variation in the relationship between ethnicity (i.e., having a Moroccan or Turkish migration background) and teacher recommendations. Specifically, it is checked whether this relationship is moderated by the proportion of non-native Dutch pupils in the class. These cross-level interaction effects are zero (for Moroccan pupils) or close to zero (for Turkish pupils), indicating that the proportion of non-native Dutch pupils is not able to explain why classes differ in the influence of ethnicity on teacher recommendations.

Statistical versus taste-based bias

The models in Table 4.2 indicate that demographic characteristics of pupils (especially SES) influence teachers' track recommendations. The question is whether this bias is of a statistical or a taste-based nature. To get an idea of the bias, we first compared the histogram of recommendations based on the EPST-result (composite) to the teacher track recommendations (see Appendix 4.A; <https://osf.io/qrg4e/>), for a variety of demographic characteristics. For bias to be generally 'statistical' in nature, we would expect the medians of the EPST- and teacher recommendations for a specific demographic group to overlap. Additionally, we would expect tracks further away from the median EPST-recommendation to be chosen less often by the teacher than the EPST. The histograms in Appendix 4.A do not universally support the idea of statistical bias. Generally, the median of EPST- and teacher recommendations are not perfectly aligned as teachers tend to give lower recommendations than the EPST. In the group of native Dutch pupils, for instance, the median track recommendation of the EPST is havo (3) whereas the teacher gives a vmbo-havo (2) recommendation or lower in 50.7% of the cases (see Appendix 4.A). Similarly, when parents' highest attained educational level is mbo2, mbo3 or mbo4, the median EPST track recommendation (vmbo-havo; 2) exceeds the median teacher recommendation (vmbo; 1). Furthermore, teachers are less likely to opt for a combination of tracks (i.e., vmbo/havo (2) and havo/vwo (4)) than the EPST. Using the group with average parent's income above 50,000 euros as an example, we see that although the median track recommendation is the same (i.e., havo; 3), the lower track alternative 'vmbo' is chosen more often by the teacher than the EPST, whereas the track in between (vmbo/havo) is chosen less often. Note that in Appendix 4.A, we also included a density plot with the EPST-scores leading to the five EPST-track recommendations³⁴.

Although the histograms in Appendix 4.A give an idea of the type of teacher bias, they do not show the relationship between the median group EPST-recommendation and the EPST-recommendation versus teacher recommendation for specific pupils. Additionally, they do not provide information on the interplay between SES, ethnicity and gender with regard to teacher recommendations. Figure 4.5a-4.5c therefore include more detailed information about the relationship between EPST- and teacher recommendations for pupils with all possible mixes of SES, ethnicity and gender. Specifically, in Figure 4.5a-c, pupils are divided (see rows) on the basis of their gender, highest attained educational level of their parents (<mbo2, mbo2-4, hbo bachelor and hbo master/university) and ethnicity (having an immigration background or not). Subsequently, pupils are divided in pupils with an EPST-recommendation below their group median (left) and pupils with an EPST-recommendation above their group median (right). Then, the percentage of pupils is shown for every possible scenario in which the EPST-recommendation and teacher recommendation are not the same (e.g., teacher recommendation < EPST, teacher recommendation is in between EPST-recommendation and the median group EPST-recommendation, teacher recommendation > EPST, et cetera). All groups with the same median group EPST-recommendation are shown in the same figure, to ease comparison.

Figure 4.5a, for instance, compares all groups with a median EPST-recommendation ‘havo’, which turn out to be all pupils with the highest educated parents. Generally, when the EPST-recommendation and teacher recommendation do not match, for these pupils teachers opt mostly for a track in between the EPST-recommendation and the EPST group median (i.e., option ‘B’; approximately 50-60% of the time). This finding is in line with the idea of ‘statistical bias’, in which the teacher recommends closer to the group median. However, comparing option ‘A’ (< EPST-recommendation (left), < EPST group median (right)) and option ‘C’ (> EPST group median (left), > EPST-recommendation (right)), it is clear that the teacher recommendation is often even more extreme than the EPST-recommendation (either even lower when the EPST-score is already below the group median or even higher when the EPST-score is already above the group median). This is the opposite of what we would expect if teacher bias was mostly statistical in nature. Comparing the percentages over the four groups, we see that girls are generally slightly advantaged over boys in the teacher recommendations. In the scenario in which the obtained EPST-recommendation is lower than the group median (left), teachers tend to more often give a recommendation lower than the EPST-recommendation for boys than girls. In the scenario in which the obtained EPST-recommendation exceeds the group median, teachers more often opt for an even higher track than the EPST-recommendation for girls. Having an immigration background also leads to slightly higher percentages for the higher option ‘C’.

Figure 4.5b compares all groups with a median EPST-recommendation ‘vmbo/havo’. These groups contain all pupils with hbo bachelor educated parents, plus native Dutch boys with mbo2-4 educated parents. Again, in general the option in between the group median and the EPST-recommendation is chosen the most by teachers (option ‘B’; approximately 50-70% of the time). Comparing option ‘A’ to option ‘C’ (on the right side; in the left scenario the teacher recommendation cannot be below the EPST-recommendation), we, however, do not observe the universal tendency for more ‘extreme’ recommendations as we saw for the groups in Figure 4.5a. The percentages for option ‘A’ and option ‘C’ are much closer in the groups of Figure 4.5b. Comparing

the groups of Figure 4.5b, the group of boys with relatively low educated parents (mbo2-4 educated parents) stands out. For this group, teachers less often opt for a track above the group median when the EPST-recommendation is below the group median (left side; 37.9% versus 41.3%-46.5% in the other groups). Additionally, teachers more often opt for a teacher recommendation below the group median (option ‘A’) when the EPST-recommendation exceeds the group median (right side; 26.3% versus 16.5%-20.1% in the other groups). Native Dutch girls again have a slight advantage over native Dutch boys (both left and right, the percentages for option ‘C’ are higher). Having an immigration background sometimes lead to an advantage (i.e., the percentage for option ‘C’ is slightly higher under non-native than native Dutch boys) and sometimes to a disadvantage (i.e., the percentage for option ‘A’ on the right side is slightly higher under non-native than native pupils).

Figure 4.5c shows the last groups with a median EPST-recommendation ‘vmbo’. These are almost all pupils with the lowest educated parents (< mbo2 and mbo2-4; except for the group with native Dutch boys with mbo2-4 educated parents, which was shown in Figure 4.5b). Most of these pupils receive a teacher recommendation that is in between their EPST-recommendation and their group median (around 80% in all groups). The percentages for option ‘C’ are largely comparable over groups, although the percentage for boys with the lowest educated parents (< mbo2) is slightly lower than for the other groups (16.8% versus 19.8%-22.8% for the other groups). Comparing Figure 4.5c to Figure 4.5a and Figure 4.5b, we see that the pupils of higher educated parents (Figure 4.5a) have a higher probability of obtaining a track recommendation that exceeds the recommendation based on their EPST-result (option ‘C’; right side). With a few exceptions, these probabilities do not differ that much between pupils whose parents completed any other, lower educational level (see Figure 4.5b and 4.5c).

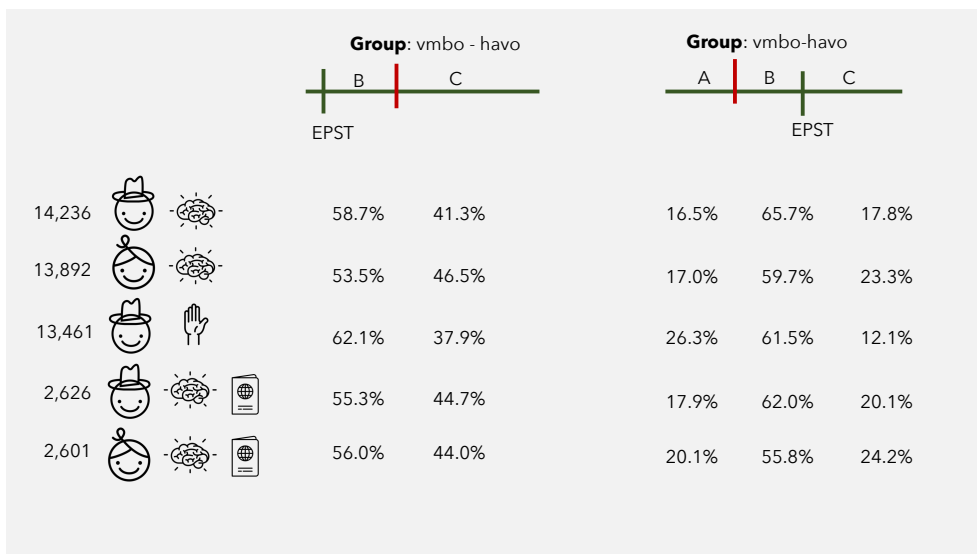
Taken together, based on Appendix 4.A and Figure 4.5a-c, teacher bias does not seem statistical in nature. Although teachers relatively often opt for a track recommendation in between the EPST-recommendation and the group average (when teacher recommendations \neq EPST-recommendations), they also often opt for a track recommendation that is even further away from the group median than the EPST-result. The ‘taste-based’ explanation for bias is only visible in the groups with the highest educated parents. When their EPST-recommendation exceeds their group average, pupils in these groups are more likely to receive an even higher track recommendation from their teacher than any other pupils with lower educated parents. But also this ‘taste-based’ explanation is too simplistic to explain the patterns in Figure 4.5a-c. Looking again at the groups in Figure 4.5a, boys with the highest-educated parents relatively often receive an even higher track recommendation than their EPST-result (right side). However, they also often receive an even lower track recommendation than their EPST-result (left side). We therefore cannot conclude that native Dutch boys with highly educated parents are always advantaged by teachers.

Discussion

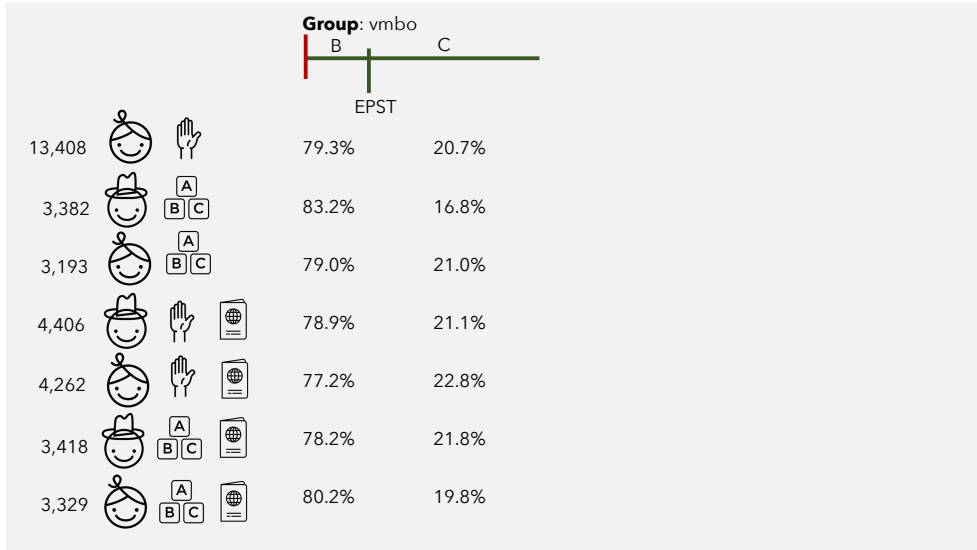
In this study, we have investigated whether teacher recommendations are influenced by a pupil’s socio-economic status (SES), ethnicity and gender in the Dutch transition from primary to secondary education. Additionally, we have investigated the nature of this



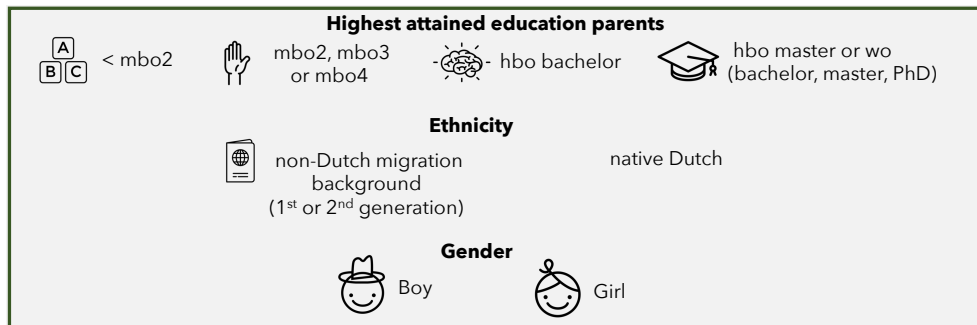
(a)



(b)



(c)



(d) Meaning of the icons.

Figure 4.5: The relationship between EPST- and teacher recommendations for pupils with all possible mixes of SES, ethnicity and gender. Note that the icons are created by Path Lord (passport), ST (hand, brain), Vectors Point (alphabetic blocks), Zuzanna Nebes (boy with hat, girl) and Seb Cornelius (university hat) from Noun Project (<https://thenounproject.com/>).

influence (statistical versus taste-based bias) and the variability of the SES-, ethnicity- and gender effects over classes. As was shown in Table 4.2, controlling for pupil's actual achievement, teacher recommendations are generally lower for pupils with lower-educated parents, parents with a lower annual income and lower WOZ-value (the three indicators of SES). Especially the effect of the highest attained educational level of parents is noticeable: keeping all other factors constant, the average track recommendation is about a quarter track level lower for pupils of lower educated parents (i.e., \leq mbo4), compared to the highest educated parents (\geq hbo master). Gender effects seem also present, but the hierarchical analysis and visual inspection of teacher recommendations (see 'Statistical versus taste-based bias') led to a mixed picture. Based on the hierarchical analysis, we would conclude that girls on average receive lower track recommendations than their male counterparts, after taking their EPST-achievement into account. Figure 4.5a-c, however, also indicate a possible advantage of girls over boys, when their EPST-result is comparable. For ethnicity, there was no indication of teacher bias on average. This finding is in line with Driessen et al. (2008) and Timmermans et al. (2018), who concluded that ethnicity effects largely disappeared over the past years. The effect of ethnicity, however, does seem to vary greatly over schools. In some schools, there is indeed no effect of ethnicity, while in others there is a reasonably large positive or negative effect present. The variation around the gender and SES-effects is also relatively large, indicating that the earlier presented effects do not hold for all schools and classes. We attempted to explain some of the variation in the SES- and ethnicity effects between school/classes by including aggregate class-level variables (i.e., the proportion of parents with less than hbo master as their highest attained level of education and the proportion of non-native Dutch pupils). These aggregate variables, however, could not satisfactorily explain why schools/classes differ. Regarding the nature of teacher bias, there was no clear support for either the notion of statistical bias (i.e., teachers take into account the average achievement of the, for instance, ethnic group a pupil belongs to) or taste-based bias (i.e., teacher systematically over- or underrecommend certain groups of pupils). Apparently, there is a need for a more detailed understanding and explanation of teacher bias.

There are numerous possible explanations for the found effects of SES and gender. Regarding SES, it is possible that parents with a higher socio-economic status put more pressure on teachers for a high track allocation (Dumont et al., 2019). Relatedly, low-SES parents might have less power in the interactions with teachers, or they might object more rarely to lower track recommendations (see Timmermans et al., 2018). The effect might also be more indirect. Possibly, non-cognitive pupil attributes that teachers often consider in their track recommendations, might be unevenly distributed across SES-classes (e.g., punctuality; Boone & Van Houtte, 2013; Geven et al. 2018). According to Timmermans et al. (2018) an often mentioned explanation is that teachers take into account the resources and support parents can offer during the secondary school period (see Ditton, Krüsken & Schauenberg, 2005). While researchers generally agree that a difference in track recommendations due to pressure of the parents is unfair, it is debatable how unfair the latter two explanations are. With regard to gender, teachers might provide higher track recommendations for girls because of their generally better work habits and engagement in primary education (Timmermans et al., 2016). It would be informative if future research would focus on verifying these different explanations.

As with any other study, this study is not without limitations. Although the CBS-database offered us the opportunity to analyze a large number of schools and

pupils, we were limited by their data collection methods. Information on schools, for instance, was only available on institution level (every institution can consist of multiple schools for primary education, and every school can contain more than one sixth grade class). As class-level effects are the most interesting, we only selected those institutions with one school for primary education and (probably) one sixth grade class. As information on sixth grade classes is lacking, we cannot guarantee that all included schools have in fact only one sixth grade class nor can we guarantee that no single sixth grade classes were excluded. As the size of institutions (i.e., the number of schools) and schools (number of sixth grade classes) is generally larger in the larger cities of the Netherlands than outside of these cities, our resulting sample might contain more pupils from the countryside than from the large cities. Additionally, the CITOtab dataset of CBS was limited to pupils who completed the ‘Centrale Eindtoets’ as EPST-test. There are, however, other options available. The choice for one of the EPST-tests differs per region in the Netherlands and size of schools, again leading to a sample that might not be entirely representative for all schools in the Netherlands (Staat van het primair onderwijs, 2019). Furthermore, we did not have access to information about the sixth grade teacher. This information would have been valuable, since teachers have been shown to vary greatly in their judgment accuracy (Südkamp et al., 2014). Possibly, variation in the effects of SES, gender and ethnicity might be caused by differences between teachers rather than class composition. Explaining this variance might also have been difficult in the present study, as we were necessarily limited to a finite set of variables. In real-life, probably many student traits are considered by the teacher that were not available to us (see Geven, Batruch & Van De Werfhorst, 2018; Timmermans, De Boer & Van Der Werf, 2016). Research into teacher recommendations would benefit from more mixed-method studies that have the advantage of qualitatively disentangling these traits without losing the advantage that large datasets and quantitative studies offer (Gross, Gottburgsen & Phoenix, 2016).

Finally, many studies into teacher recommendations, including the present one, are limited to the period of transition between primary and secondary education. It would be interesting, however, to also investigate the years prior and after this transition. SES-, ethnicity- and gender-differences might develop early on in the educational career of pupils and can either be enlarged or mitigated by teachers in the years thereafter. After the teacher recommendation, pupils have the possibility to ‘correct’ an erroneous teacher recommendation by track switching (see dashed lines, Figure 4.1). Research into track mobility in secondary education, for instance, suggests that SES, gender and ethnicity might influence pupil’s initial track choice and the degree to which track switching, grade retention and drop-out occur (see, for instance, Tieben, 2009; Tolsma, Coenders and Lubbers, 2007). When there is an effect of SES, gender and ethnicity on track mobility, we currently do not know whether this effect relates to the initial track recommendations of the sixth grade teacher. A danger of focusing solely on the teacher recommendation is furthermore that logistics of the transition between primary and secondary education are not taken into account. Teacher recommendations can be influenced, for instance, by the tracks that are provided in adjacent schools for secondary education, as some schools offer ‘bridge-years’ (i.e., vmbo/havo and havo/vwo) whereas others do not. Actual track placement of pupils also differs from the teacher recommendations, as secondary schools need to properly divide the pupils into classes. With a ‘havo’ teacher advice, for instance, pupils can end up in a first year vmbo/havo, havo or havo/vwo class. The arguments secondary schools use to place ‘havo recommended pupils’ in any of these classes are

interesting to investigate, as a havo/vwo class offers other educational opportunities to pupils than a vmbo/havo or havo class does. Thus, only if we take the whole primary and secondary education period into account, we can fully understand the impact of SES, ethnicity and gender on educational opportunities.

Chapter 5

**DEVELOPMENT AND EVALUATION OF
A DIGITAL EXPERT ELICITATION
METHOD AIMED AT FOSTERING
ELEMENTARY SCHOOL TEACHERS'
DIAGNOSTIC COMPETENCE**



Kimberley Lek & Rens Van De Schoot

Chapter 5

Development and evaluation of a digital expert elicitation method aimed at fostering elementary school teachers' diagnostic competence

Abstract

Expert elicitation - an approach to systematically consult experts and quantify their insights - has been successfully applied in fields as risk assessment, health and environmental research. Unfortunately, it has never been used within the Educational sciences, while it offers ample opportunities for educational practice, especially when used to foster the accuracy of teacher judgments; generally referred to as their 'diagnostic competence'. The current chapter is the first to explore expert elicitation in an educational context and has two major goals. The first goal is to develop a digital expert elicitation method suitable to be used by elementary school teachers for self-reflection purposes. The second goal is to extensively test the expert elicitation method, using a test panel of 24 primary school teachers for 503 pupils in total. Results regarding the development of the elicitation method and its reliability, construct validity, face validity and feasibility are discussed as well as ideas how this elicitation method can be a valuable self-reflection instrument for teachers. The results are promising: all measures of reliability, feasibility, face validity and construct validity show positive results and teachers are enthusiastic about the possibilities of the method.

Keywords : Expert elicitation, Bayesian statistics, Educational practice, prior, Bayesian updating

Introduction

Sometimes, experts³⁵ possess unique knowledge, that is impossible or impractical to attain using traditional data collection methods. In those instances, expert elicitation can be used to 'obtain' this knowledge. Specifically, the purpose of expert elicitation is to "construct a probability distribution that properly represents expert's

knowledge/uncertainty” (O’Hagan et al., 2006, p. 9), such that this expert knowledge can be used in – for instance - research, engineering projects and decision-making (O’Hagan et al., 2006). The resulting probability distribution can be analyzed on its own or can be combined with data. Regarding the latter option, Bayesian statistics can be applied (see Van De Schoot et al., 2013; Kaplan & Depaoli, 2013; Kruschke, 2011). The expert knowledge distribution – called a ‘prior’ in Bayesian terminology – is then joined with data (‘likelihood’) in a ‘posterior’, which is an optimal compromise of the expert’s knowledge and the data.

Expert elicitation has gained a lot of attention in for instance risk assessment (e.g., Clemen & Winkler, 1999; Edlmann et al., 2016), health (e.g., Hald et al., 2016; Havelaar et al., 2008; Kirk et al., 2015; Knol et al., 2010) and environmental research (e.g., Siegel et al., 2017; Singh et al., 2017). But, as König and Van De Schoot (2017) show in their systematic review, although Bayesian statistics is slowly gaining attention in the Educational sciences, expert elicitation is not. This is unfortunate, since we believe expert elicitation offers many opportunities for educational practice and research. Using expert elicitation, we could elicit teachers’ judgments with regard to the ability of the pupils they teach. Since teachers spend much time with their pupils, they have a unique perspective on their development, educational needs, et cetera. Or, as Machts et al. (2016) put it: “because of this broad exposure, teachers are often expected to be able to provide differential diagnostic information on their students that go far beyond the measure of their performance on academic tasks in specific domains” (p. 85). Expert elicitation can make the intangible, implicit judgments of teachers explicit.

An advantage of making these judgments explicit is that the elicitation tool can function as a feedback instrument for the teacher. When used on multiple occasions, for instance, the teacher can see how his view on the child’s development has changed and he can evaluate what (rational and/or irrational) events have led to this change. Another advantage: when multiple teachers teach the same class, it is possible to quantitatively compare the judgments of these teachers, making differences in judgments directly apparent and open for discussion. Furthermore, the process of completing the elicitation tool can provide useful feedback as well. For example, when a teacher finds the elicitation difficult for a certain pupil, he knows that his view on this pupil’s development is still a bit vague. Another advantage for the teacher is that - using the Bayesian toolbox - it is possible to combine the elicited teacher expertise with other data on the development of the pupil (i.e., observational and test data). Doing so will lead to a posterior that reflects all the information available on the pupil’s development. The ultimate goal of making teachers aware of their implicit judgments, how these judgments change over time and how they compare to those of colleagues or to other data sources is to foster the accuracy with which teachers judge their pupils on a daily basis, referred to as “teachers’ diagnostic competence” (Artelt & Rausch, 2014; Pit-Ten Cate et al., 2014).

The above ideas of making teacher insights explicit fit within the increasing focus on data-driven decision making (Espin et al., 2017; Van Den Bosch et al., 2017) and evidence-based practice (Boudett et al., 2013) within education. Instead of basing classroom decisions on “[...] anecdotes, gut feelings, or opinions” (Mandinach, 2012, p. 71), teacher insights can be formally assessed and evaluated and be combined with other data, such as data from tests or classroom observations, to drive educational practice. In this way, expert elicitation can assist teachers that typically “[...] process information in their heads” (Mandinach, 2012, p. 72). The increasing focus on rational

data collection is in marked contrast to the long held belief that ‘informed intuition’ should be accepted as the primary basis of teacher judgments (see Vanlommel et al., 2018; Creighton, 2007).

The present chapter is the first to explore expert elicitation in an educational context and has two major goals. The first goal is to present an expert elicitation method that is specifically tailored to elementary school teachers. This tailoring was necessary since elementary school teachers differ from experts in fields as Risk assessment and Health in three important ways: (1) elementary school teachers typically have little knowledge of statistics, (2) teachers have to elicit many priors (one for every pupil) instead of one or a few, (3) the elicitation procedure has to be self-explanatory and time efficient to be of practical value. The second goal – following recommendations of Johnson et al. (2010a) – is to assess measurement properties of this elicitation tool (i.e., face validity, feasibility, intra-rater reliability and construct validity).

The remainder of this chapter is ordered as follows. First, we provide a theoretical background on the developed elicitation method and procedure. After pilot-testing, this elicitation method/procedure was applied and evaluated with 24 primary school teachers in an expert meeting. Background characteristics of the primary school teachers and specifics of the expert meeting are discussed in the section ‘methods’. Thereafter, the elicited priors are illustrated and the measurement properties of our elicitation method/procedure are discussed. The chapter ends with a general discussion of the usability of our elicitation method/procedure. This study received ethical approval from our internal Ethics Committee of the Faculty of Social and Behavioural Sciences of Utrecht University (nr. FETC17-046; the letter of approval can be found on the OSF page <https://osf.io/qrg4e/>).

Background Expert Elicitation

Expert Judgments

The general goal of expert elicitation is to capture and quantify experts’ implicit and intangible judgments. Applied to the context of elementary school teaching, our purpose is to make explicit teachers’ implicit judgments of their pupils ability. These judgments are a reflection of the teacher’s diagnostic competence; his or her ability to judge achievement characteristics of his or her pupils correctly (Artelt & Rausch, 2014). In this article, we focus specifically on the ‘math ability’ of pupils. Although judgments with regard to other areas such as language development can also be elicited, the advantage of mathematics is that it is a relatively unambiguous area. As Artelt and Rausch (2014) put it: “Given that mathematical skill is mainly acquired in school and that the curriculum is well-defined according to age, teachers quite likely have a shared understanding of what constitutes mathematical proficiency” (p. 35).

Within the literature, there is a broad consensus that diagnostic competence is a central aspect of teachers’ professional competence (Artelt & Rausch, 2014; Pit-Ten Cate et al., 2014), since many decisions are based on teachers’ judgments such as school placement, tracking decisions, grade allocation, adaptive teaching, ability grouping, instructional decision-making and the creation of tests in the classroom (Gabriele & Park, 2016;

Baumert & Kunter, 2013). Despite its importance, previous literature shows that teacher judgments are often quite subjective and do not always meet reliability and validity criteria (see, for instance, Südkamp, Kaiser & Möller, 2012; Kaiser, Retelsdorf, Südkamp, & Möller, 2013). One reason for this is that teacher judgments are not always based on a systematic and deliberate collection, weighing and integration of informational cues. Rather, teachers often switch from such a complex judgment process (called ‘attribute-based judgment; Krolak-Swerdt et al., 2012) to a quicker and more efficient judgment process based on a minimum of information cues and stereotypical information (called ‘category-based judgment; Krolak-Swerdt et al., 2012; a similar distinction can be found in Kahneman, 2011 and Evans, 2008). The danger of this latter processing strategy is that judgments can be flawed because of well-documented expectancy effects, Pygmalion effects, halo effects, fundamental attribution errors, and any other errors that are the result of selective information processing (Artelt & Rausch, 2014, Kahneman & Frederick, 2005; also see Meissel et al., 2017 and Kaiser, Südkamp, & Möller, 2017). In the expert elicitation method (see next section), we simply ask teachers to express their judgments, without guiding their judgment process. The elicited judgments can therefore be the result of a resource-intensive, rational strategy, a strategy that is based on mere impressions and intuition or a combination of both. Next to eliciting experts’ judgments, expert elicitation also focusses on the confidence of these experts in their judgments. Only recently, teacher judgment confidence has gained attention in the literature (see, for instance, Gabriele & Park, 2016). A promising theory suggests that when teachers are confident in making accurate judgments and not confident when they make inaccurate judgments, they are “effectively monitoring their judgment accuracy and are making calibrated judgments about their students” (Grabriele & Park, 2016, p. 51; also see Dunlosky & Metcalfe, 2009). Because of this meta-cognitive monitoring, the calibrated teacher knows when he/she needs to collect more information on certain pupils, increasing their judgment accuracy and hence improving their diagnostic competence.

Elicitation Method

The idea of expert elicitation is to express experts’ judgments and their confidence in these judgments in a so-called prior distribution. Applied to the context of elementary school teachers, this means that we obtain a prior distribution for the judgment of ‘math ability’ for every pupil in a teacher’s class. One of the first choices to make in the development of an elicitation method is the scale to be used. Since ‘math ability’ is not something that can be observed directly, it does not have an inherent, natural scale. Consequently, to keep track of the mathematical abilities of their pupils, schools use different tests and observational systems that (may) differ in the scale on which math ability is expressed. One of the challenges was thus to create a universal, intuitive scale that could be used by every teacher and school. We have tackled this challenge by creating a scale that is based on percentile scores (Crawford & Garthwaite, 2009). The advantage of percentile scores is that they are relatively simple to interpret and that every scale can easily be converted into percentile scores (Lezak et al., 2004). To make it even more simple and intuitive, we use a scale with either 5 (Figure 5.1a), 10 (Figure 5.1b), 25 (Figure 5.1c) or 50 (Figure 5.1d) ‘puppets’. Depending on the number of ‘puppets’, every ‘puppet’ represents a certain percentage of pupils (e.g., percentile score). When 5 ‘puppets’ are used, for instance, the first puppet on the scale represents the 20% pupils with the lowest

math ability in the age range of the pupil, the last puppet on the scale the 20% pupils with the highest math ability, et cetera.

The idea is that teachers position their pupils using 1-5, 1-10, 1-25 and/or 1-50 ‘puppets’. Thus, if a teacher believes a pupil to be part of the 20% best math students of his or her age, the teacher places this pupil at puppet 5 on the 1-5 scale. Using the scale(s) in this way, teachers can easily express their judgments regarding the math ability of their pupils at once, maintaining the relative differences between their pupils.

In order to obtain a distribution for every pupil, we also need to have an indication of the uncertainty of the teacher with regards to the positions chosen (i.e., the teachers’ judgment confidence). Obtaining such an estimate of uncertainty is a delicate matter, since people are known to generally underestimate their uncertainty (Lichtenstein et al., 1982; see also Speirs-Bridge et al., 2010; Bier, 2004). Additionally, most elicitation procedures ask the experts to state their uncertainty using precise probabilities (e.g., “90% certain”), something that is hard for people who are layman with respect to statistics. With the scales in Figure 5.1, however, obtaining an indication of uncertainty is rather intuitive and simple. Teachers simply choose the scale (Figure 5.1a, b, c or d at which they feel certain enough to position their pupil(s)). The scale with 5 ‘puppets’, for instance, is coarser than the scale with 25 ‘puppets’, and thus the teacher who chooses the latter scale is inherently more certain than the teacher who chooses the 1-5 scale. By using this approach to eliciting the teachers’ uncertainty, we avoid the necessity to ask for precise probabilities.

The idea of letting the number of ‘puppets’ decide the uncertainty of the teacher is related to the Equivalent Prior Sample (EPS)-method of the classical, often cited work of Winkler (1967). In the EPS-method, an expert gives an estimate of a proportion and an estimate of the sample size he is basing the proportion on (O’Hagan et al., 2006). Like the number of ‘puppets’, the larger the sample size, the more certain the expert is about his proportion. An important difference between our elicitation procedure with ‘puppets’ and the EPS, however, is that the ‘puppets’ are visualized. By visualizing the ‘puppets’, the teacher can experience what it means to choose a position with 5 or 50 ‘puppets’. This will help avoiding that teachers choose too many ‘puppets’ (i.e., in the EPS-method experts tend to select relatively large sample sizes) and, thus, avoids that the teachers underestimate their uncertainty.

A parametric distribution that naturally fits the idea of the scales with ‘puppets’ is the Beta distribution, with parameters alpha and beta. Linked to the scales in Figure 5.1, alpha is the chosen position and beta the total number of ‘puppets’ minus the chosen position. Figure 5.2a shows the translation of a few chosen positions into Beta distributions. Indeed, when keeping the position consistent while moving from scale 1-5, to 1-10, 1-25 and 1-50, the Beta prior becomes smaller and smaller, taking into account the increasing certainty of the teachers. Figure 5.2b also shows how the Beta prior is rather flexible; it is highly skewed when teachers choose a low position while reassembling a normal distribution around the middle positions.

Taken together, teachers only need to position each of their pupils on a chosen scale to define a prior distribution. This positioning takes little time, is intuitive, asks for little statistical knowledge and can be done for their whole class at once.

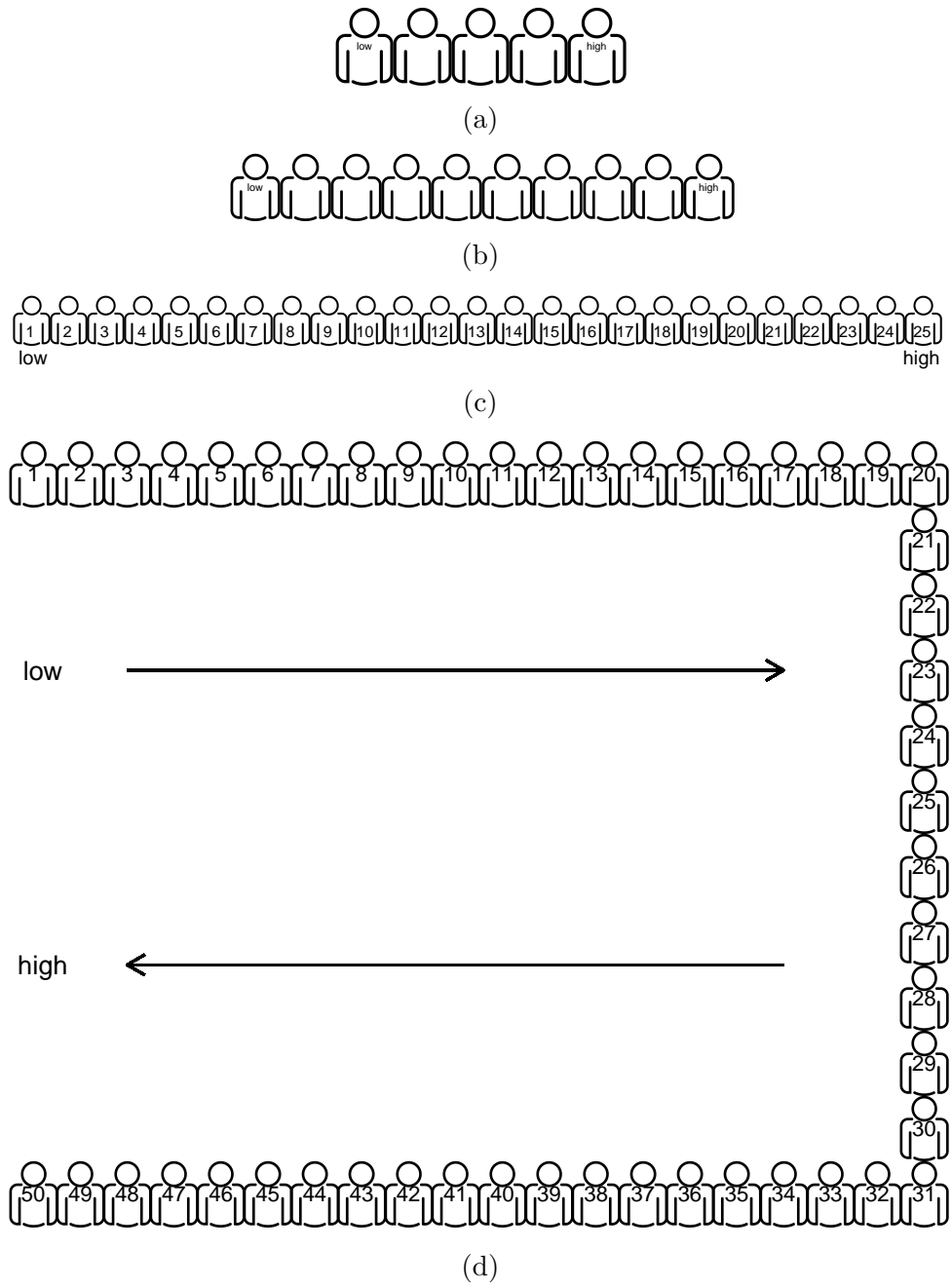
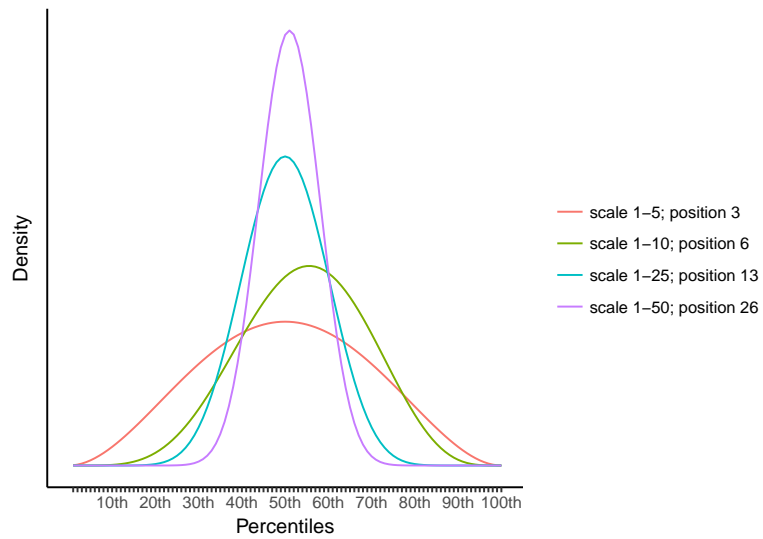
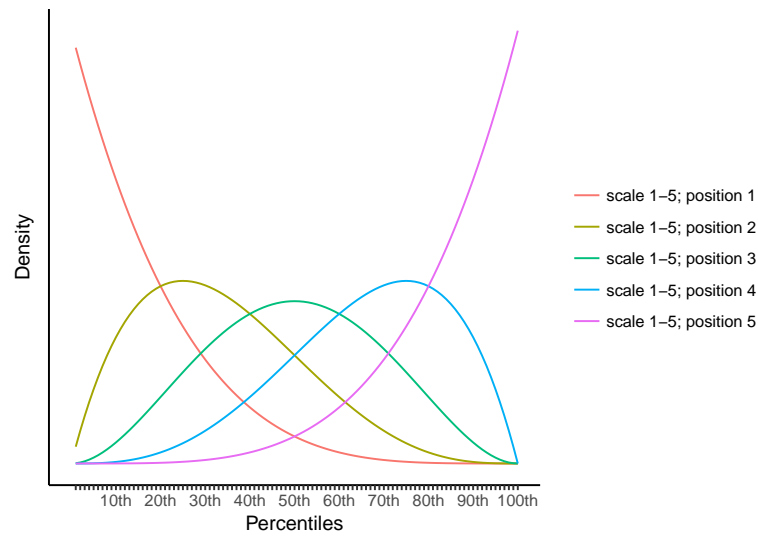


Figure 5.1: Scales used for elicitation, with varying number of ‘puppets’ between 5 (a), 10 (b), 25 (c) and 50 (d). Note that the icon is created by Tommy Lau from Noun Project (<https://thenounproject.com/>).



(a)



(b)

Figure 5.2: Examples of the translation of chosen positions on the scales (Figure 5.1) to Beta distributions. (a) shows the Beta distributions belonging to comparable positions on the 1-5 to the 1-50 scale; (b) shows the Beta distributions for all possible positions of the 1-5 scale.

Elicitation Procedure

We developed a digital elicitation instrument that walks through the elicitation method in six steps. This digital instrument was developed in Dutch and can be found on <https://osf.io/qrg4e/>. We programmed the digital version using R (R Core Team, 2017) and the Shiny package (Chang, Cheng, Allaire, Xie & McPherson, 2017). The digital elicitation instrument was developed to be self-explanatory; teachers' judgements can be elicited without the intervention of a researcher. Since the digital version in its current form requires a (stable) internet connection, we also developed a paper-based elicitation instrument that mimics the digital instrument as closely as possible, for when such an internet connection is not available.

Step 1.

Both the digital (start screen) and paper-based version (first page) start with a motivational text about the purpose of the expert elicitation. The goal of this motivational text is to engender the experts enthusiasm for the project (Clemen & Reilly, 2001) and to make sure they take the elicitation procedure seriously.

Step 2.

In both the digital and paper-based version, teachers were asked to divide their pupils in groups, based on their math ability. This step was included to ease step 5; the step in which the teachers position their pupils. Without the groups, teachers would have to position their pupils all at once. Now, teachers had the possibility to do the positioning for groups of pupils separately. Since most teachers work with math subgroups in their class, this was anticipated to be a relatively easy step to start with.

Step 3.

To maintain pupils' anonymity, in the digital version teachers needed to download a document with codes (a random combination of one letter and 2-3 numbers)³⁶. This document contained as much random codes as pupils in the teachers' class, for each of the defined groups separately. In the document, there was a blank space after every random code such that the teachers could fill in pupil names for their own administration and use this document as key. Teachers were instructed to keep the key private. In the paper-based version, teachers received the same document on paper, although they had to select the appropriate number of random codes themselves.

Step 4.

In the fourth step - both in the digital and the paper-based version - teachers could read all necessary information to continue to the actual elicitation. This is what O'Hagan (2006) calls a 'probability training'. Since (1) the elicitation instrument needed to be as time efficient as possible and (2) the elicitation method is deliberately kept as non-technical as possible, the probability training entailed only the basics that teachers needed to

comprehend. Specifically, the meaning of percentile scores and representation by the ‘puppets’ was explained and the Beta distribution was illustrated.

Step 5.

In the fifth step, teachers were asked to position their pupils using the scales in Figure 5.1. This positioning was done for each of the defined groups separately, using the random codes of step 3. Teachers started with 5 positions (Figure 5.1a). After they completed the positioning they were presented with 2 options: either continue to a more fine-grained scale (in this case, 10 positions; Figure 5.1b) or state that they did not feel certain enough to move to a more fine-grained scale and stop the positioning for this pupil/this group of pupils. Every time a teacher chose the option to move to a more fine-grained scale, it was presented with the next scale until the scale with 50 ‘puppets’ was reached. In the digital version, teachers could position their pupils by first selecting one of the student codes and then right-clicking on the ‘puppet’ of choice (Figure 5.3a). By doing so, the random student label appeared above the selected puppet and the corresponding Beta distribution was printed below the scale as feedback (Figure 5.3b). Based on this feedback, the teachers could except this representation of their beliefs or change the positioning of the selected pupil, by simply right-clicking on another ‘puppet’. When moving from one scale to the next, an orange box was shown for the selected pupil around the ‘puppets’ that would lead to a consistent positioning with respect to the previous choice of position (Figure 5.3c). In the paper version, teachers could simply write the pupil code above the puppet of choice. The interactive elements of the digital version (i.e., the Beta distribution and orange box) were absent in the paper-based version. Note that after completion of step 5, all the ingredients are present for the construction of a prior.

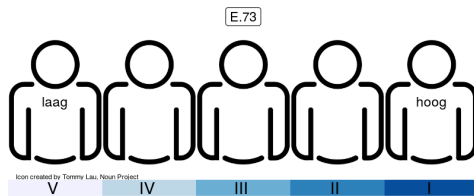
Step 6.

After the teacher positioned all of his pupils, he/she moved on to the final step: answering ‘check’ questions. In this step, teachers either selected (digital version) or wrote down (paper version) two student codes that were, according to the teachers, closest and furthest apart in terms of math ability. The answers to these questions are later on used to assess construct validity.

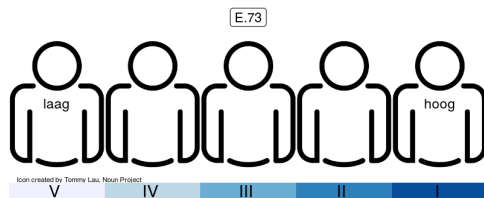
Elicitation result

After completing all steps of the elicitation procedure described above, the result of the elicitation can be visualized as in Figure 5.4a. In one glance, the teacher can see his or her judgments (the peak of the distributions), how confident he/she is in these judgments (the width of the distribution) and how his/her judgments and judgment accuracy differ over pupils. Now that it is visualized, these judgments can easily be shared with others, such as colleague teachers, the headmaster, parents, et cetera. According to Pit-Ten Cate et al. (2014), two promising ways of improving teachers’ judgment accuracy and hence their diagnostic competence are to raise awareness of their judgments and to increase accountability. As explained in Pit-Ten Cate et al. (2014), raising awareness and accountability boost teachers’ motivation to be highly accurate. Research

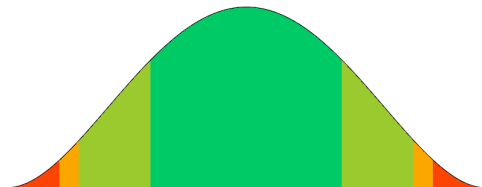
Selecteer een leerling
E.73



(a)



Feedback



(b)



(c)

Figure 5.3: Screenshots of the digital elicitation instrument. (a) illustrates the positioning of the pupil code “E.73” on a 1-5 scale; (b) shows the digital feedback for the chosen position of “E.73” and (c) shows which positions on the 1-10 scale would be consistent with the chosen position for “E.73” in (a).

has shown that teachers are likely to shift from a generally less-accurate category-based judgment process to the generally more-accurate attribute-based judgment process when their motivation to be accurate is high (see, for instance, Krolak-Schwerdt, Böhmer & Gräsel, 2009 and Kunda & Spencer, 2003). The idea is that when teachers look at the result of the elicitation and share this result with others, both their awareness of their implicit judgments is raised and their accountability is increased since they (feel the) need to justify their result to others. Pit-Ten Cate et al. (2014; also see Wahl, Weinert, & Huber, 2007) also stress the importance of comparing teachers' judgments with actual student achievement and showing possible discrepancies to the teachers as feedback. The result of the elicitation makes it fairly easy to make such a comparison, as visualized in Figure 5.4b. The discrepancy between judgment and actual achievement can be used, for instance, to find sources of errors and to test implicit hypotheses and judgments teachers have.

Methods

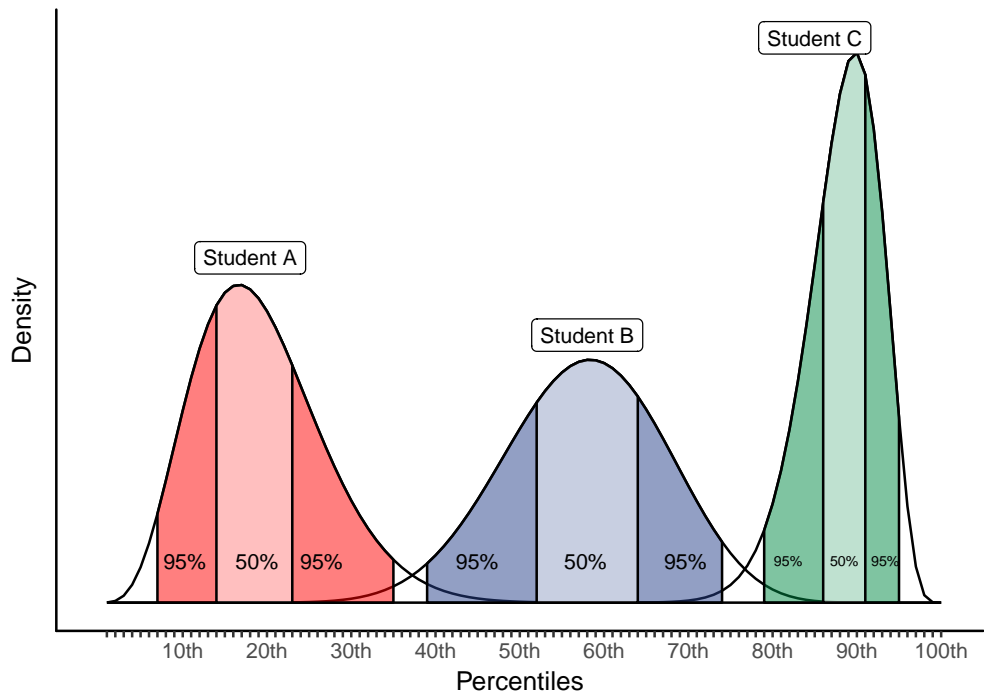
To test the usability of our elicitation method/procedure in practice, we first conducted a pilot test with two primary school teachers from the personal network of the first author, using the digital version of the elicitation instrument. Thereafter, we organized two expert meetings in which primary school teachers were invited to apply the elicitation method to the pupils of their own class(es). The primary school teachers participating in these expert meetings were asked to complete the elicitation procedure once more at home. In this section, specifics of the expert meetings are given. In the next sections, results of these endeavors are discussed, with a focus on measurement properties of the elicitation method/procedure.

Participants

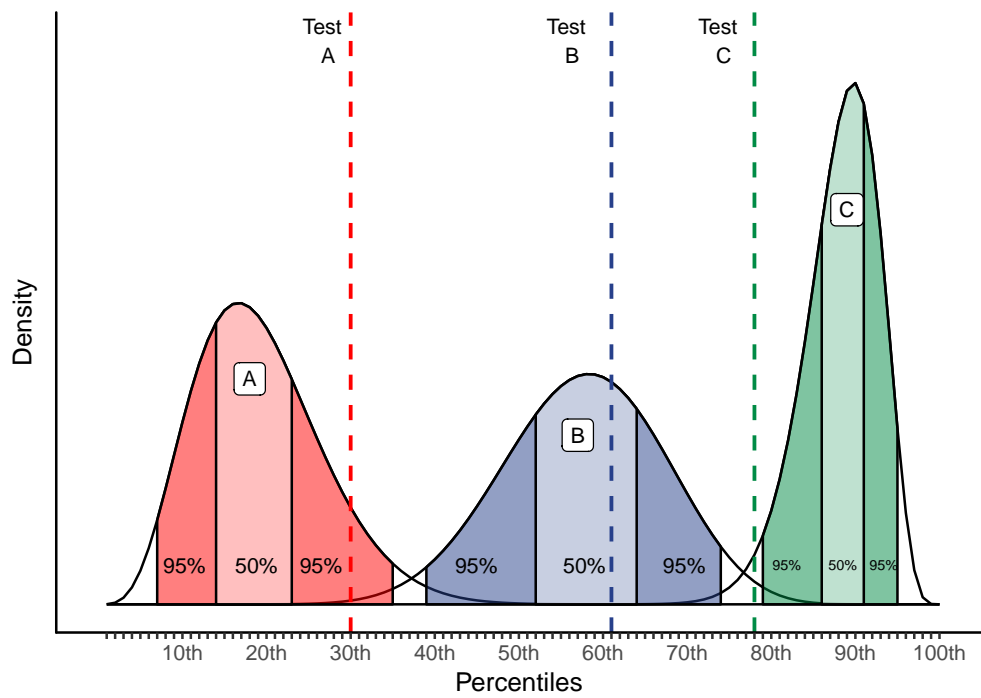
Primary school teachers were recruited via social media (i.e., Twitter and Facebook). Note that the focus on math ability was not mentioned in the call, to prevent a bias towards teachers with a specific interest in math. Participation of teachers was based on one expert meeting and one home assignment, which were rewarded with 100 euro. Two expert meetings were held (April the 14th and 19th, 2017) to accommodate different agendas. 24 primary school teachers elicited prior distributions during both of these expert meetings, for 503 pupils in total. Most of the teachers were female (20 of the 24) and the average age was 30.38 (sd = 8.18; min = 20, max = 55) with an average of 7.88 years of experience (sd = 8.14; min = 0/first year of teaching, max = 34). The teachers taught a variety of classes.

Design

Each of the expert meetings lasted for approximately 2.5 hours. During these 2.5 hours, teachers received an introduction into the goal of our research. Furthermore, they received a document with instructions for the home assignment, which simply entailed following the steps of the elicitation instrument once more, 2-5 weeks later in order to assess



(a)



(b)

Figure 5.4: Hypothetical example of the result of the expert elicitation. Each distribution is for one specific pupil. The dotted lines in (b) show the result of a test.

reliability. The most important instruction on this document was to keep the key and use the same student codes for the home assignment. After the introduction, the first author showed the teachers the digital version of the elicitation instrument, step by step. Then, teachers were invited to use the digital elicitation instrument. They were encouraged to follow the steps of the elicitation instrument independently. However, if questions did emerge, the first author was there for assistance. Unfortunately, due to issues with the server, not all teachers could complete the elicitation instrument digitally and had to switch to the paper version. Specifically, during the expert meetings 3 of the 24 teachers used the digital version whereas for the home assignment all teachers were able to complete the digital version. Upon completion of the elicitation instrument, teachers were asked to fill in an evaluation form. As described earlier, these questions were used to assess face validity and feasibility. During the expert meeting, written informed consent was obtained from all participating teachers. Parent(s)/caretaker(s) of pupils were not asked to sign an informed consent form, since no data of the pupil were collected except for the opinion of teachers.

Results

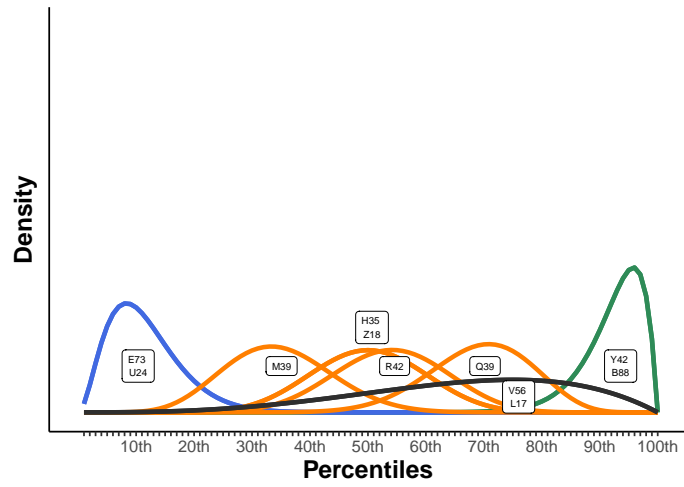
Elicited Distributions

The result of the expert elicitation is an expert knowledge distribution for every of the 503 pupils, in 24 classes. Here, we discuss the distributions for three of our teachers' classes as an example, using Figure 5.5. In Figure 5.5a, the distributions are shown for 11 pupils rated by one of our teachers. This teacher made four groups in step 2 of the elicitation procedure: a low math ability group (with pupils "E73" and "U24"), an average math ability group (with "M39", "R42", "Q39", "H35" and "Z18"), a high math ability group ("Y42" and "B88") and a group with relatively high math ability pupils ("V56" and "L17"). For this teacher, his (or her) division of pupils in groups corresponds with his elicited distributions as we see a clear separation of the four groups. Interestingly, the teacher is less certain about pupil "V56" and "L17" than about his other pupils. Figure 5.5b shows the distributions of a teacher with a less clear separation between groups. Possibly, this teacher experienced some difficulty with dividing his pupils into ability groups (in which case the elicitation instrument helps to practice this) or the groups are based on something different than solely math ability. Finally, whereas in Figure 5.5a and 5.5b there is some overlap between all probability distributions, in Figure 5.5c pupil "Y42" stands completely on its own, indicating that this pupil's math ability is not comparable to any other pupil in this class, at least according to his/her teacher.

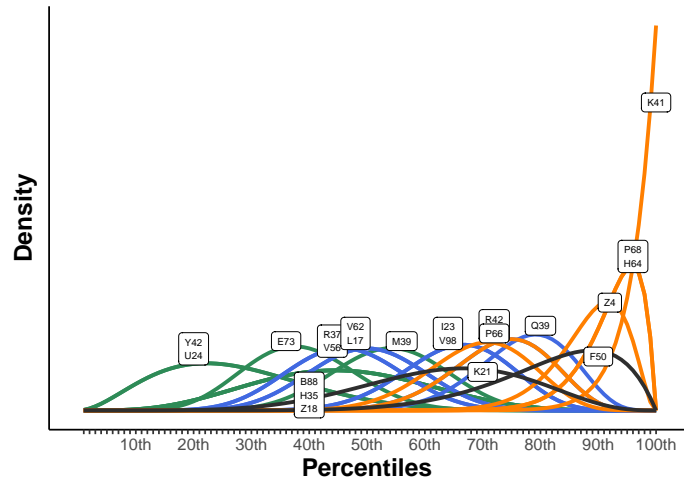
Measurement Properties

Reliability

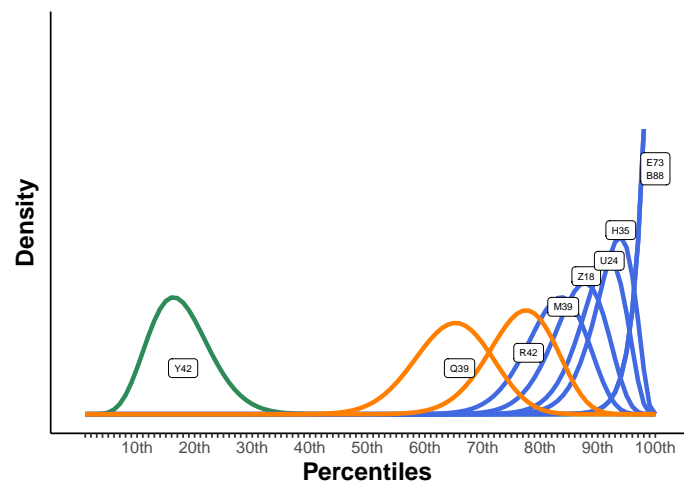
Intra-rater, test-retest reliability of the elicited probability distributions was assessed based on the results of the expert meeting and the home assignment. 23 out of the 24 primary school teachers succeeded in completing the home assignment, for 470 out of the 503 pupils. In the assessment of reliability, there were a few unique characteristics of our



(a)



(b)



(c)

Figure 5.5: a, b and c show three examples classes with elicited probability distributions.

elicitation method we needed to take into account. First, different teachers can choose different scales to rate their pupils on and it is possible that teachers use a different scale for a certain pupil at the expert meeting and at the home assignment (see Table 5.1). Second, for our elicitation method, an estimate of reliability for every teacher or even every pupil separately would be most informative, rather than a single reliability coefficient. Third, in the assessment of reliability we need to keep in mind that the result of the elicitation is a distribution instead of a single value.

Table 5.1: Scale chosen during the expert meeting and the home assignment.

	Home assignment				
	1-5	1-10	1-25	1-50	Row total
Expert meeting					
1-5	53	0	2	0	55
1-10	8	53	30	0	91
1-25	19	57	130	16	222
1-50	31	0	29	42	102
Column total	111	110	191	58	470

To take all these characteristics into account, we decided to use two ways of assessing reliability. First, we estimated a Cohen’s Kappa adjusted for the interval/ordinal nature of the positions chosen by the teachers (κ_A ; formula 22 in Gwet, 2008; also see Janson & Olsson, 2001). The advantages of this Cohen’s Kappa are that it can be calculated for every teacher separately and that it has known cut-off values that aid interpretation. A disadvantage is that it, like other common measures of reliability, ignores the fact that the elicited positions describe distributions (Johnson et al., 2010a). Another disadvantage is that it is unsuited for the scenario of different scales for the pupils of one teacher and discrepancies in scale chosen at the expert meeting and the home assignment. To deal with the latter disadvantage, the κ_A estimate is based on the highest scale on which all of the teacher’s pupils were rated.

Second, we calculated the overlap between the resulting Beta distribution at the expert meeting and the resulting Beta distribution at home for every pupil, using the Hellinger Distance (HD; see Nikulin, 2001), which takes into account that the result of the elicitation is a distribution. This HD can be estimated for every teacher and for every pupil separately. Furthermore, the HD can be estimated even when teachers choose a different scale at the expert meeting and home assignment. A downfall is that there are no guidelines for the HD as an indicator of reliability. To ease interpretation, Appendix 5.A (see <https://osf.io/qrg4e/>) visualizes the size of the HD for different chosen positions at the expert meeting and home assignment.

Overall, the Kappa coefficients and the Hellinger distances indicate satisfactory reliability. The resulting Kappa coefficients are high for 20 of the teachers (minimum = 0.75, maximum = 1.00, mean = 0.89) but low for 3 of them (i.e., 0.26, 0.54, 0.50). It is likely that the latter three teachers have not used the pupil coding consistently. Figure 5.6 illustrates the resulting HD for all pupils. The lower plot shows the HD when the highest scale is used on which a particular pupil is rated at both the expert meeting and at the home assignment. The upper plot shows the HD when different scales at the expert meeting and at the home assignment are taken into account. Here, the distances

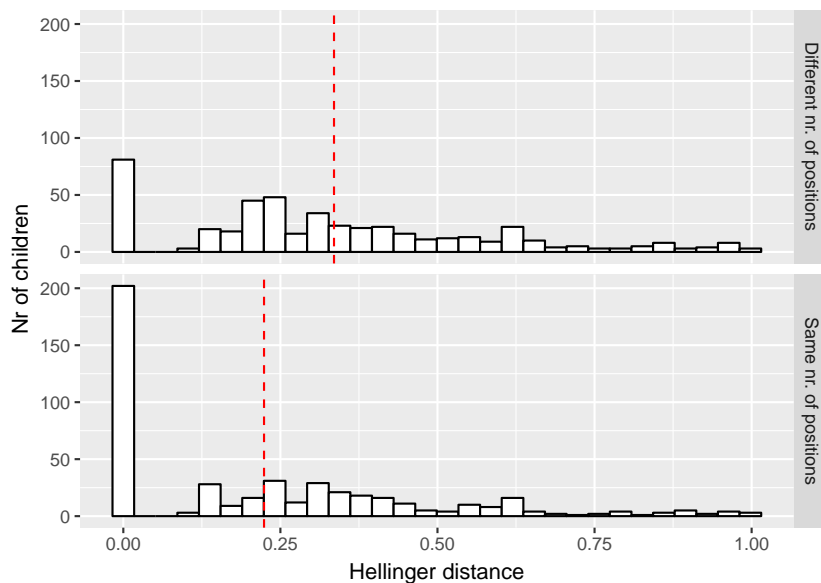


Figure 5.6: Hellinger distances (HD) for all pupils, when allowing different rating scales (upper part) or not (lower part). The red dashed line indicates the mean HD.

are generally bigger than at the lower part (mean difference = .11) since differences in the variance of the distributions are now taken into account as well. Generally, the HDs are indicative of a reasonable match between the Beta distributions for most of the pupils (in the upper part of Figure 5.6, 78% of the HDs are above the midpoint of the HD-scale; in the lower part this percentage is 86%).

Validity

Face validity. To get an indication of face validity – i.e., the appropriateness, sensibility and relevance of the expert elicitation for teachers (Holden, 2010) - we asked the teachers to rate two statements (in Dutch; statement 1 and 2 in Table 5.2) and asked to answer two questions (in Dutch; question 3 and 4 in Table 5.2). The teachers had the following answering options: (1) not at all, (2) not, (3) a little bit, (4) affirmative, (5) totally affirmative. Teachers could provide an explanation for the latter two questions as well. Generally, the answers of the teachers are indicative of a high face validity (all averages above 4; see Table 5.2.). In the explanations, the teachers merely stated that they believed their positions to be an accurate representation of their class (question 3) and that combining these chosen positions with for instance test or observational data would be valuable, in their opinion (question 4). Some examples (translated from Dutch):

- “I believe I have mapped my pupils in an accurate way”
- “The image (of my class) is clear and insightful”.
- “I believe that another teacher who is unfamiliar with my class, can get a good overview of my class by looking at my positioning”.
- “ [these steps] make you aware of the choices you make as a teacher about the level of the pupils.”

- “ [...] the distributions show clearly the position of the pupil without focusing too much on one single position”.
- “ this stimulates teachers to think very carefully about the abilities of their pupils without blind trust in test results”.

Feasibility. To obtain an indication of feasibility or ease of usage (Johnson et al., 2010b), we asked the teachers to rate the ease and clearness of the steps in the application (see questions/statements 5-9, Table 5.2). With the exception of the sixth and the seventh question, the answering options were again (1) not at all, (2) not, (3) a little bit, (4) affirmative, (5) totally affirmative. Again, feasibility is generally highly rated (all above 4 on average; see Table 5.2).

Construct validity. To evaluate the construct validity of the elicitation instrument, at each elicitation occasion, teachers were asked two ‘check’ questions:

- Which two pupils are closest to each other with respect to their mathematical ability?
- Which two pupils are furthest from each other with respect to their mathematical ability?

When the answers on these check questions and the ratings on the position scales did not contradict, this was used as an indication of construct validity. At the expert elicitation meeting, both for question 1 and 2 there was a 95.8% match between the teacher’s answers on the control questions and their ratings. This percentage was lower for the elicitation at home: 69.6% (question 1) and 82.6% (question 2; note that for one teacher, scores for this elicitation occasion were missing).

When the control questions and the ratings on the position scales did contradict, this usually indicated a minor contradiction. For question 1, usually two pupils were found who were placed slightly closer on the position scale(s) and for question 2, usually two pupils were found who were placed slightly further apart on the position scale(s); see Appendix 5.B (see <https://osf.io/qrg4e/>), Figure B.1-B.2. There were a few exceptions; see Figure B.3-B.4; Appendix 5.B.

In addition to the ‘check’ questions, we compared our elicitation method with the trial roulette method (Gore 1987; also see Johnson et al., 2010b; Goldstein et al., 2008; Goldstein & Rothschild, 2014; Zondervan-Zwijnenburg et al., 2017); a commonly used elicitation method. In our version of this method, teachers were asked to place digital coins on the scale with 25 ‘puppets’, for two of their pupils. These coins were presented in a hypothetical gamble situation: if you would earn money for correctly estimating the position of a pupil, on which positions would you bet (i.e., place a coin)? The teachers were instructed to place coins on at least 3 positions and were told that they could use as many coins as they liked, knowing that the total value of the coins was 1,000 euros. Thus, with 10 coins the value of one coin would be 100 and their hypothetical profit would be the total value of the 100-coins placed on the correct position. Other than in our elicitation method, for the trial roulette method there is not a one-to-one correspondence between the elicitation and the parametric prior distribution. Instead, the R-package ‘SHELF’ (Oakley, 2017) was used to fit a parametric beta distribution

Table 5.2: Average scores (sd) for face validity and feasibility ($n=24$).

	Mean (sd)
Face validity	
1. To me, it is clear what the rows with “puppets” represent	4.63 (0.49)
2. To me, it is clear what a “statistical distribution” is	4.17 (0.56)
3. Do you feel that your knowledge and insights with respect to your pupils’ math ability have been accurately represented by this application?	4.17 (0.56)
4. In this meeting, you have learned about the goal of our research. Do you think that our research can be valuable for primary education?	4.42 (0.58)
Feasibility	
5. After reading the information, it was clear to me what was expected	4.08 (0.65)
6. To me, positioning pupils was easiest with*	Frequency
5	9
10	11
25	5
50 puppets.	0
7. To me, positioning pupils was the most difficult with*	Frequency
5	2
10	1
25	2
50 puppets.	21
8. For me it was easy to...	
a. Divide the pupils in smaller groups	4.50 (0.59)
b. Answer the “control questions” (see “construct validity”)	4.21 (0.66)
9. For me it was clear what was expected of me...	
a. When asked to divide the pupils in smaller groups	4.88 (0.34)
b. When answering the “control questions”	4.38 (0.71)

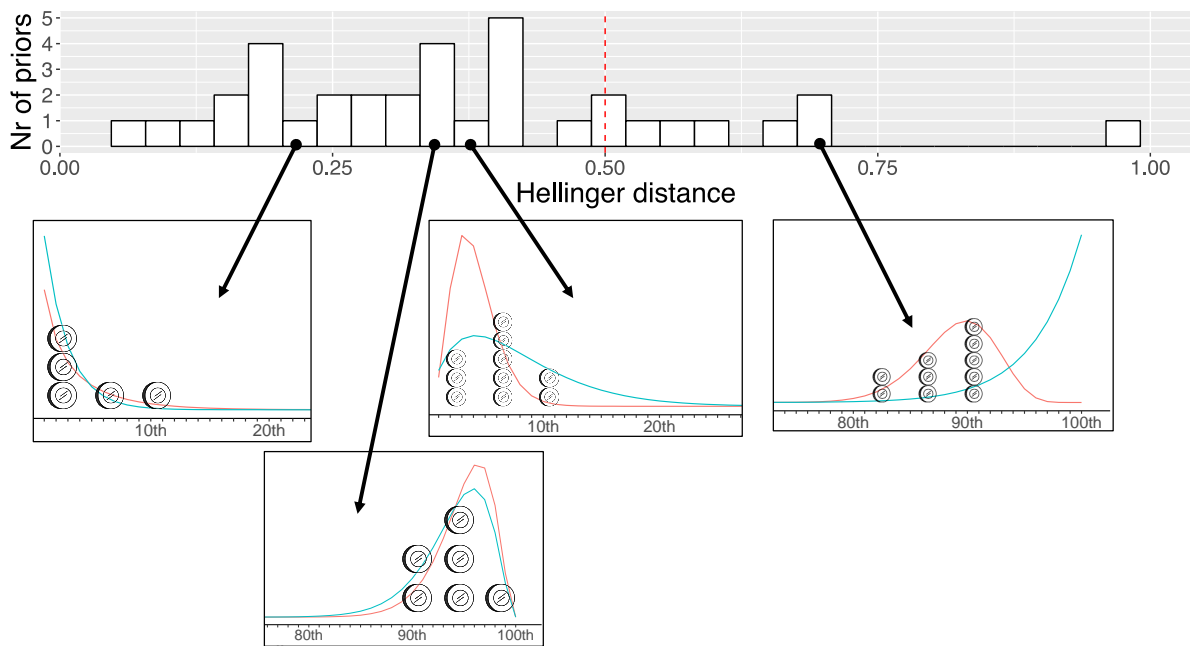


Figure 5.7: Histogram with Hellinger Distances ($n = 36$; on top) and four examples of differences between the trial roulette method (coins + in red) and our elicitation method (in blue; bottom).

that would reassemble the placed coins as precisely as possible. After completing the trial roulette method, teachers were asked to state whether they thought our elicitation method was easier to comprehend or the trial roulette method.

Eighteen teachers successfully completed the trial roulette method, yielding 36 elicited priors with this method. To compare these priors with the priors resulting from our elicitation method, we again used the Hellinger Distance. As Figure 5.7 shows, 29 of the Hellinger Distances were at or below the midpoint of the HD-scale, indicating a reasonable agreement between our elicitation method and the trial roulette method. This Figure also showcases a few examples of small, reasonable and exceptional Hellinger Distances. 17 of the 24 teachers stated to prefer our elicitation method, based on ease of usage. According to these teachers, our elicitation method provided a clearer overview, made it easier to rank order pupils and was easier to complete with a single mouse click. Given this preference and the fact that the trial roulette method would be time consuming to complete for every pupil, our elicitation method seems a feasible alternative.

Discussion

In this chapter, an easy and intuitive expert elicitation method is discussed which is suitable to be used by elementary school teachers. Specifically, this expert elicitation method can be used by teachers to quantify their insights regarding the math ability (or any other ability) of their pupils. In essence, the elicitation method entails the placement of pupils on a scale with ‘puppets’ (see Figure 5.1), representing a proportion of pupils. The scale the teachers choose determines the width of the resulting Beta

prior. The position of the pupil on this scale determines the prior's mean. Following the recommendations of Johnson et al. (2010a), we extensively evaluated the reliability and validity of the expert elicitation method. The results are promising: all measures of reliability, feasibility, face validity and construct validity show positive results. One of the additional challenges of this chapter was to create an elicitation instrument that could be applied reliably in educational practice without the assistance/presence of a researcher. Generally, the elicitation tool discussed in the current chapter meets this challenge. We did, however, notice some differences between the results of the expert meeting (with assistance of a researcher) and the results of the home assignment (without assistance of a researcher). For example, the percentages correspondence between the check questions and the ratings were lower for the home assignment than at the expert meeting, which might indicate sloppiness. Additionally, teachers tended to choose more detailed scales in the expert meeting than at home (see Table 5.1), which might indicate sloppiness or decreased motivation. Whatever the reason may be, when teachers stop at a relatively coarse scale, the resulting prior has less of an influence. In this way, the step-by-step procedure protects the teachers from allocating too much confidence to elicited positions that they have invested relatively little time in.

There are some limitations that should be taken into account when using the expert elicitation method. First, the simplicity and ease of usage of the elicitation instrument come at the expense of some flexibility. Since there are only four scales to choose from (Figure 5.1), the resulting Beta distribution can only take a limited set of variances/widths. In this study, we however found no indication that teachers wanted to use a coarser or more fine-grained scale which was not available. During a discussion at the end of the expert meeting, the teachers stated that they felt at least certain enough to place their pupils on a 1-5 scale, indicating that a coarser scale was not necessary. Additionally, when asked whether teachers would consider a scale from 1-100, only 3 of the 24 teachers said 'yes', indicating that the given scale options were generally sufficient.

Another way in which flexibility is limited is the type of distribution. With our elicitation instrument, only a Beta distribution can be elicited. The advantage of the Beta distribution is that it is a standard distribution that is well-understood, convenient and relatively easy to work with (O'Hagan et al., 2006), with a rather flexible form and an intuitive interpretation with respect to the scales in Figure 5.1. A downfall may be that a Beta distribution always has a peak, meaning that a teacher cannot express that two or more positions are equally likely for a certain pupil. But, as a partial justification of the Beta distribution, O'Hagan et al. (2006) state: "[...] experience indicates that people's knowledge about uncertain quantities is usually well represented by smooth unimodal densities".

As mentioned under 'design', not all teachers could complete the elicitation procedure using the digital version of the elicitation instrument. Although we tried to make the paper and pencil and digital version as comparable as possible (see 'Elicitation procedure'), there is one important difference between the two: the paper and pencil version lacks the interactive elements of the digital one. This difference might affect the chosen positions and scale (see Figure 5.1). Unfortunately, we were not able to assess whether the use of the digital versus the paper and pencil version has influenced the elicitation results. Since all teachers used the digital version for the home assignment and the majority of the teachers used the paper and pencil version at the expert meeting, any difference between the expert meeting and home assignment could be caused by the

different modes (digital versus paper and pencil) or different settings. Future research is needed to assess the consequences of the interactive elements of the digital version for the elicitation results.

The limitations notwithstanding, our elicitation method can have some clear advantages when applied in elementary school setting, both with and without (test) data. Without (test) data, the elicitation method can be interesting for feedback- and communication purposes. Figure 5.5 illustrated this argument. Looking at Figure 5.5a – 5.5c, the rank ordering of the pupils according to the teachers becomes immediately clear. We can also easily see how comparable pupils are – according to the teachers – with regard to math ability by looking at the degree of overlap between the distributions. Additionally, the uncertainty a teacher has regarding the math ability of his/her pupils becomes apparent, looking at the width and height of the distributions. On the evaluation form, one of the questions was whether teachers found the elicitation procedure useful (and why). Teachers were generally very enthusiastic about the elicitation procedure (see Table 5.1, question 4) and for instance stated that the procedure helps to put their insights and observations on paper in an insightful way. As described before under ‘Elicitation result’, having these insights on paper helps to raise awareness of teachers’ implicit judgments (Pit-ten Cate et al., 2014). Being aware of their implicit judgments, teachers generally become more motivated to collect data and form judgments in a systematic, deliberate way, especially when their elicitation result shows low confidence (Gabriele & Park, 2016). Having them on paper, some of the teachers also predicted their insights would be taken more seriously (by parents, colleagues, et cetera) instead of being perceived as ‘gutfeelings’. Others saw possibilities to use the elicitation method for instance when discussing a pupil with a remedial teacher. Another possibility is when multiple teachers teach the same class. In that case, both teachers can elicit a prior for the same pupil and compare the results quantitatively. The elicitation method then provides an easy and time efficient way to detect differences in insights and ideas about pupils. In all these examples, the elicitation result helps to increase (perceived) accountability, since teachers defend their judgments to others (Pit-ten Cate et al., 2014). Just as raising awareness, raising accountability can lead to a higher motivation to be accurate, leading to generally more accurate teacher judgments (see Kunda & Spencer, 2003).

The elicited distributions can also be compared with collected data, such as observations or test results. Figure 5.4b illustrated this possibility, for three hypothetical pupils. The hypothetical test results of these pupils are visualized with a dashed line and the dark and lighter colored areas respectively show the 95% and the 50% credible interval (i.e., the Bayesian version of a confidence interval, see Van De Schoot et al., 2013) of the hypothetical probability distributions as formulated by the teacher(s). In this hypothetical example, the teacher insights and the test result correspond for the pupil in the middle (i.e., the test result falls nicely within the 50% area). This correspondence confirms the insights of the teacher and he or she might want to specify a more peaked probability distribution next time, indicating his or her increased certainty. For the pupil at the right side, the test result is notably lower than what the teacher would expect. This discrepancy might indicate that the teacher overestimates this pupil’s math ability or that the test underestimates the pupil (or a combination of both). Figure 5.4b can in this case be used as a starting point to formally assess what caused the discrepancy between teacher and test. Finally, for the pupil at the left side of Figure 5.7, there is somewhat correspondence between the teacher and the test (i.e., the test result falls within the 95%

area). Based on Figure 5.7, the teacher can start to investigate whether the relatively high test result is a random, to be expected fluctuation or that this pupil might be capable of more than the teacher expects. According to Pit-ten Cate et al. (2014), this deliberate comparison of teacher judgments with test results is another promising way of improving teacher judgment accuracy.

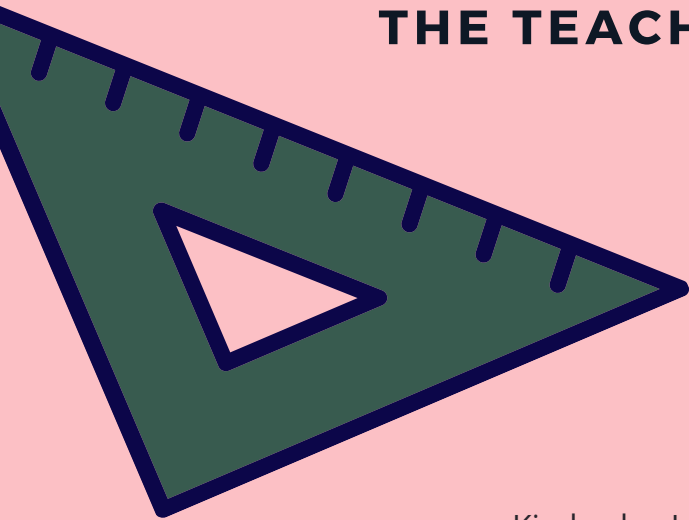
Now that we have created an elicitation procedure that is intuitive, easy to work with, reliable and valid, future research can start to explore all of the stated possibilities above. An interesting question, for instance, is how exactly the expert elicitation method can foster teachers' diagnostic competence. In order to answer this question, we need to understand which cognitive processes underlie teachers' ability to form accurate judgments. Promising work in this area is done by Herppich et al. (2018). From a statistical perspective, we want to investigate further how ideas of the teachers can be combined and contrasted with test data, using Bayesian statistics. Furthermore, we think it is interesting to find characteristics of the pupil(s), class, teacher(s) and/or school that influence the discrepancies found between the test data and the teacher judgments. This can be done in a quantitative way – using Bayesian multilevel analysis (see for instance Praetorius et al., 2017) – but also qualitatively, asking the teachers to comment on any discrepancies found. All in all, we believe the elicitation procedure opens up many opportunities to investigate tacit differences between 'subjective' teacher judgments and 'objective' data such as test results and to raise the diagnostic competence of teachers.

PART III

Teacher & Test

Chapter 6

**THE TRANSITION FROM PRIMARY TO
SECONDARY EDUCATION IN THE
NETHERLANDS: WHO KNOWS BEST,
THE TEACHER OR THE TEST?**



Kimberley Lek & Rens Van De Schoot

A Dutch version of this article was published in De Psycholoog, 2019, volume 54, issue 4

Chapter 6

The transition from primary to secondary education in the Netherlands: Who knows best, the teacher or the test?

Abstract

Before 2014, the End of Primary School Test (EPST) was leading in the transition from primary to secondary education in the Netherlands. With the start of the 2014-2015 school year, however, a change of law put the teacher's advice first. The new policy has been heavily disputed in academia and has led to fierce debate in the media. The central question is: who knows better, the teacher or the End of Primary School Test? Using longitudinal data from the CBS (Statistics Netherlands), the authors establish that neither is superior to the other in all situations. They conclude that using a combination of teacher's advice and test advice is the best option.

Keywords: End of primary school test, transition to secondary education, teacher judgment

Introduction

In advising for the best fitting educational level with the transition from primary education (PE) to secondary education (SE), the advice resulting from the Cito End of Primary School Test (EPST) used to be leading in the Netherlands. This changed in the 2014-2015 school year, with new legislation stipulating that the teacher's advice should be leading in the transition to secondary education, with an end-of-school test as secondary requirement.³⁷ The new policy meant a diminished role for the End of Primary School Test in determining the advice for pupils leaving primary school.

The new legislation was not welcomed by everyone. Niessen and Meijer, for example, wrote that the teacher is 'no measuring instrument,' arguing that teachers, like all human beings, 'often are not that good at weighing information to arrive at the right decision'

(NRC.next, 15 April 2016). To prevent ‘underadvising’³⁸, parents had the possibility to request a reconsideration of the teacher’s advice if the EPST indicated a higher level of secondary education than the original teacher’s advice. Soon, however, it transpired that only 1 out of 6 of these requests resulted in a modification of the teacher’s advice (See the House of Representatives Letter ‘Eerste inzichten Wet Eindtoetsing PO’ (4 December 2015) and Korpershoek et al., 2016). The result? Tension between parents, PE schools and SE schools (See the House of Representatives Letter ‘Overgang van primair naar voortgezet onderwijs’, 2 July 2015) and mounting inequality between children of assertive parents with degrees in higher education and children who were ‘just as smart’, but whose parents had no such degrees or leverage (‘De staat van het onderwijs’, Inspectorate of Education, 2018; Van Spijker, Van Der Houwer & Van Gaalen, 2017).

Still, others welcomed the new policy. Kamerman and Vasterman, for example, referred to the oft-mentioned argument that the test measured just one moment in time of a pupil’s cognitive development, whereas the teacher has a more overall idea of a pupil’s capacities (‘De leerkracht weet het vaak écht beter dan de Citotoets’, Kamerman & Vasterman, 30 March 2015). And research by Feron, Schils & Ter Weel (2015) indicated that in comparison to the test, the teacher’s advice was the better predictor of the pupil’s position in the third year of secondary education. Their research, however, was carried out in the period before the new policy, when teachers could actually consult the test results before writing their advice. This is no longer possible in the new situation, since the EPST was moved forward, meaning that teachers have no access to the test results in time for the admission procedures. The question therefore remains whether the teacher’s advice is a better predictor in the current situation, where teachers have no access to the test results.

Now that three school years have passed since the new policy was implemented, speculation and debate can give way to more objectively informed analysis. Relevant data have become available to investigate the predictive power of the teacher’s advice. We now have data on third-year placement of former sixth grade pupils and, different from Feron et al. (2015), we also know what teachers advised prior to the EPST in 2014-2015. This chapter addresses specifically the following questions:

1. Can we speak of a (systematic) discrepancy between the teacher’s advice and the EPST advice in 2014-2015?
2. In retrospect, how do EPST and teacher’s advice compare for predicting placement at a certain education level?
3. How often have pupils switched between the different SE levels and can these switches be explained by the teacher’s advice and/or the test advice?
4. Would pupils have benefitted from a mandatory rather than an optional adjustment of the teacher’s advice, if the test results indicated a higher SE level?

These questions shall be answered with the help of longitudinal data from the CBS (Statistics Netherlands). For the sake of readability, all kinds of technical matters are presented in four online appendices that can be consulted at the open science framework (see <https://osf.io/qrg4e/>).

Description of the data

For our analyses we used the CBS data files with data on EPST results, teacher advices, and figures on SE placement for a total of 119,751 pupils.³⁹ See Appendix 6.A (see <https://osf.io/qrg4e/>) for a detailed description of origin and availability of these data. The teacher advice consists of the advice that the sixth-grade teacher fills out on the EPST form. This advice is either single (indicating one segment: vmbo, havo, or vwo) or a composite (e.g. vmbo-havo, havo-vwo). See Appendix 6.A for definitions of these advices and how they are used in our analyses.

With respect to placement, we have decided to subsume the four different vmbo levels (bb, bbkb, kb, and gt) into one overall category, ‘vmbo.’ We chose to do so because moving up a level is generally easier within the vmbo segment, than from vmbo to havo (Onderwijsraad, 2018) or from havo to vwo (Exalto et al., 2019). In other words, the distinction between the four vmbo levels is less decisive for a pupil’s educational career than the vmbo-havo-vwo distinction.

Discrepancy EPST and Teacher’s Advice

Can we speak of a systematic discrepancy between the teacher’s advice and the EPST advice in 2014-2015? A rough distinction for vmbo, vmbo-havo, havo, havo-vwo, and vwo advices yields an exact match of teacher’s advice and EPST advice for 56.1% of the 119,751 pupils included in our study (see Figure 6.1. For 29.4%, the EPST advice was at least half a level higher than the teacher’s advice. For our purposes, ‘half a level’ refers to a composite advice by either teacher or EPST, or by both, and with a partial overlap of teacher’s and test advice. For example, the teacher advises havo, the EPST havo-vwo. For only 14.5% of the pupils the teacher’s advice was at least half a level higher. When wielding the stricter requirement that for the different vmbo levels (bb, bbkb, kb, gt) the advice of teacher and test should be exactly the same, the agreement percentage drops from 56.1% to 37.6% (see Appendix 6.B for additional information, available at the open science framework <https://osf.io/qrg4e/>).

As for the type of advice, teachers in the 2014-2015 school year were less inclined to use a composite advice (vmbo-havo, havo-vwo) than the EPST. For example, for 16,566 pupils (13.8%) the EPST advice was vmbo-havo, while the teachers gave such an advice only for 10,576 (8.8%) pupils. Similarly, 22,756 pupils received a havo-vwo EPST advice (19.0%), with that number dropping to 11,159 (9.3%) in teachers’ advices. In addition, the EPST had a relatively high number of vwo advices (24,297 pupils, 20.3%) when compared to the teacher vwo advices (23,037 pupils, 19.2%).

Match of advised level and the level eventually taken

The previous section shows the lack of agreement between EPST advice and teacher’s advice for quite a large number of cases in the third year of SE. This raises the following question: In retrospect, how do EPST and teacher’s advice compare for predicting placement at a certain education level? We answer this question by comparing the

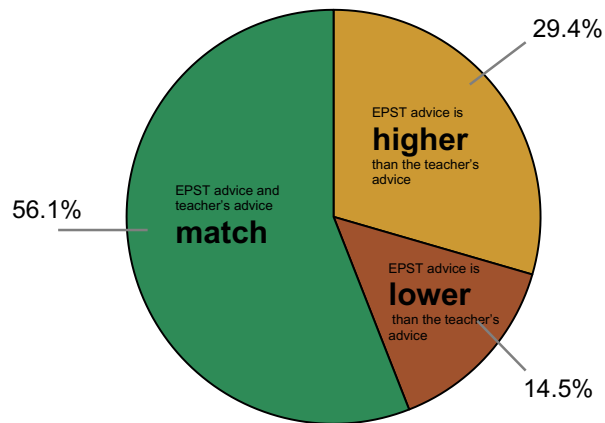


Figure 6.1: Percentage of the 119,751 pupils with matching EPST and teacher's advice in 2014-2015 (green), with the EPST advice at least half a level higher than the teacher's advice (yellow) and the EPST advice at least half a level lower than the teacher's advice (brown).

predictive value of EPST and teacher's advice with respect to placement of pupils three years later. We decided on the third year of secondary education, because in the first two years it is still relatively easy to switch levels.

Match with both teacher's and EPST advice

Three quarters of the pupils were assigned a level that matched both teacher's and EPST advice (See 'Average', Figure 6.2). Relatively speaking, this match occurred least often with havo pupils. This does not come as a surprise, since, given the middle position havo takes up in the three levels, there is a higher chance of either overestimating or underestimating by teacher and EPST. With vmbo and vwo pupils there is only the chance of either overestimating or underestimating. The percentage for 'matching both teacher and test' is highest for vwo pupils.

Match with either teacher's advice or EPST advice

In those cases where EPST and teacher's advice did not match in 2014-2015, the pupil of 2017-2018 was, on average, more often placed at a level that agreed with the teacher's advice than with the EPST advice (See 'Average', Figure 6.2). Especially with vmbo pupils, the predictive value of the teacher's advice is higher: 13.4% of the pupils are placed at a level that matches only the teacher's advice, compared to 3.8% matching only the EPST advice. For havo pupils these percentages are less divergent. For vwo pupils, in contrast, it is the EPST that has greater predictive value: 4.7% of the pupils are assigned a level that matches only the teacher's advice, compared to 10.7% matching only the EPST advice.

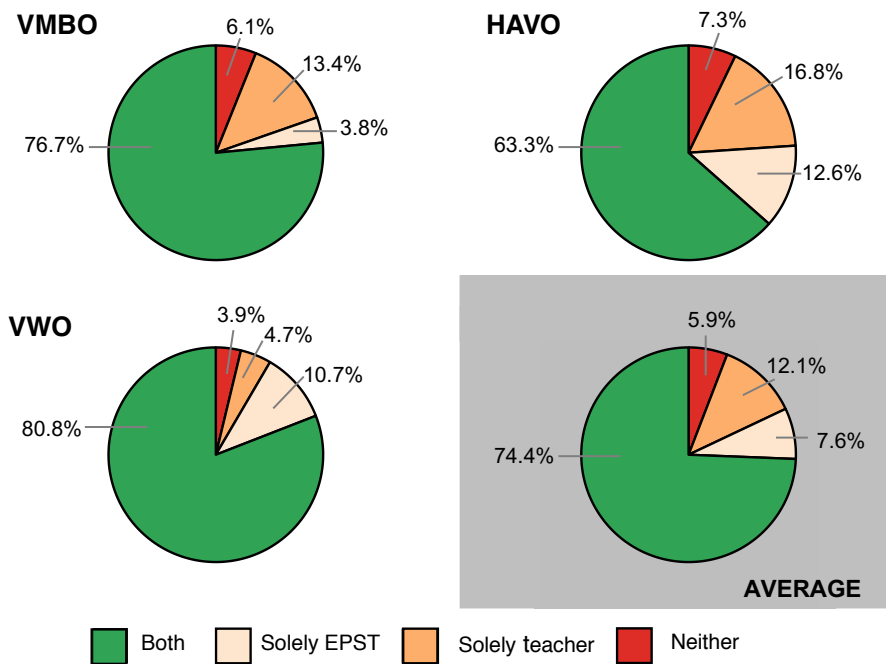


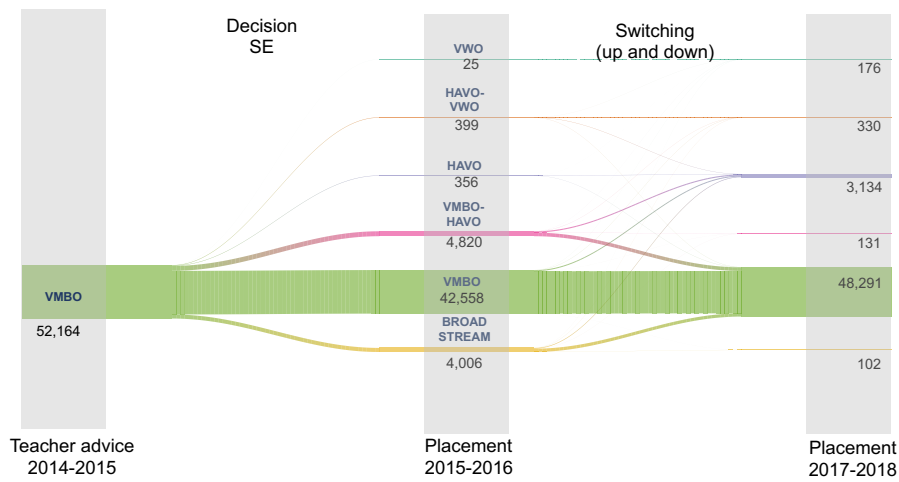
Figure 6.2: The pie charts show the percentages of pupils who in 2017-2018 were taking an educational level that ‘matched’ either their teacher’s advice or the EPST advice. ‘Matching’ here means: exact agreement or overlap of the advice and the level eventually taken. A separate pie chart shows the percentages of pupils enrolled in vmbo, havo, or vwo in 2017-2018.

Pupils’ switching-behaviour

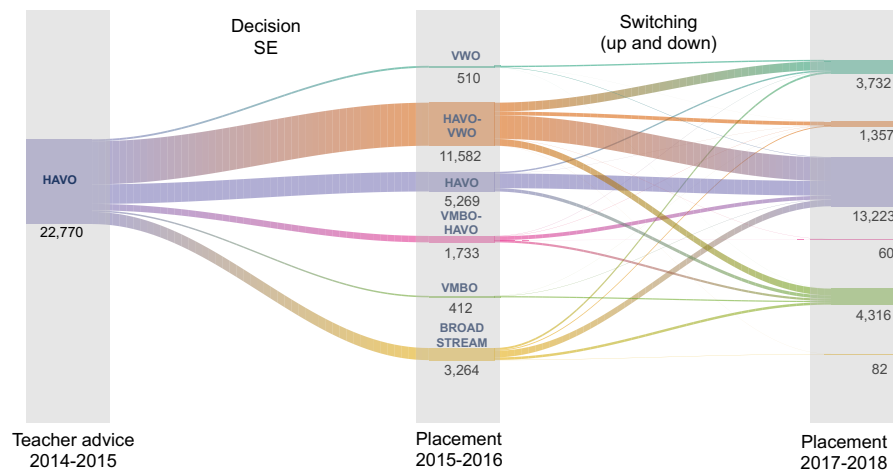
Figure 6.3 shows the level that pupils with a teacher’s advice for vmbo, havo, or vwo were allocated initially in SE in 2015-2016 as well as the level they were taking eventually in 2017-2018. Appendix 6.C has a more detailed version of Figure 6.3 (to be consulted at <https://osf.io/qrg4e/>), as well as versions for pupils with a composite vmbo-havo or havo-vwo advice. Here we discuss the difference between teacher’s advice and initial placement (Figure 6.3, left), pupils’ switching-behaviour (Figure 6.3, right) and whether this switching-behaviour has an explanation in the nature of the teacher’s or EPST advice.

Difference between teacher’s advice and initial placement

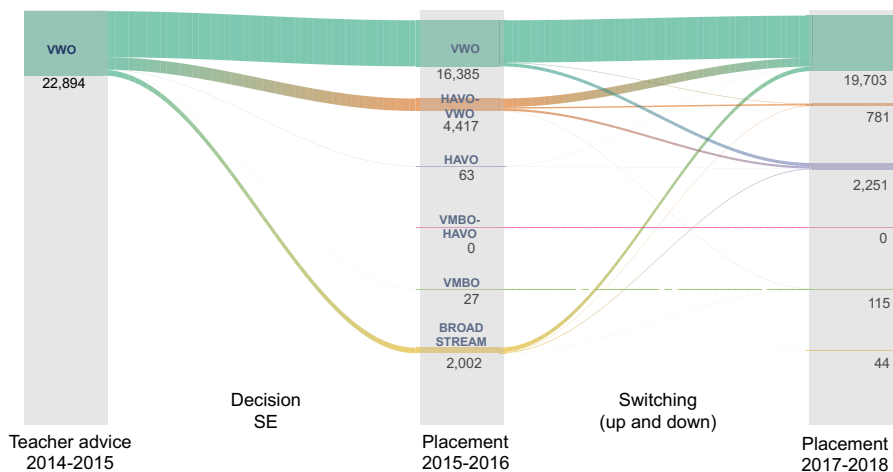
SE schools are not required to place pupils in exact agreement with the teacher’s advice. Parents and pupils, for example, play an important part in choosing an educational level. Some schools, moreover, offer a ‘broad stream’, (i.e. a mix of all SE levels in one class), where splitting up for the different levels then occurs after 1 or 2 years. Figure 6.3 shows that pupils with a vmbo or a vwo teacher’s advice relatively often also start at a vmbo or vwo level in 2015-2016, whereas only a quarter of the pupils with a havo teacher’s advice ended up in a first-year havo class. These pupils often start in a mixed havo-vwo first year and more often use the ‘broad stream’ SE mix.



(a)



(b)



(c)

Figure 6.3: Displayed are, for all pupils with a vmbo, havo, or vwo advice, the placing process in SE (the lines between teacher’s advice and initial placement), and how often pupils have switched to another level (the lines between initial placement in 2015-2016 and eventual placement in 2017-2018). Note that some lines are lacking because of the risk of identifiable data.

Difference between initial placement and year 3 placement

Pupils with a vmbo or vwo teacher's advice show a relatively high measure of stability (Figure 6.3): often they do indeed start and finish at the vmbo or vwo level. Pupils with a havo teacher's advice show greater mobility both upwards and downwards (i.e., at least half a level in both scenarios). Especially downward switching from havo to vmbo is notable. Upward or downward mobility occurs relatively more often if these pupils have started their school career in a 'broad stream' or combined (vmbo-havo, havo-vwo) class.

Switching-behaviour in relation to teacher's and EPST advice

Switching-behaviour can have a variety of reasons. Switching can be required if for instance the type of broader first year education is no longer on offer in the third year. For example, after one or two years of combined vmbo-havo classes the pupil has to switch to either vmbo or to havo exclusively. Another possibility is that the original teacher's or EPST advice was accurate, but that initial placement in SE did not (wholly) match. And then there can also be all kinds of circumstances, unforeseen by the teacher and the EPST, that cause a pupil's upward or downward mobility. Figure 6.4 comprises the switching-behaviour of the 119,751 pupils. It shows that 39,304 pupils (32.8%) made a switch between 2015-2016 and 2017-2018. Of these 39,304 pupils, 15,279 moved up (38.9% of all switchers) and 24,025 pupils moved down (61.1% of all switchers).

In more than half of all cases of switching, the switch agrees with both teacher's and EPST advice: 55.2% for switching upward and 55.5% for switching downward ('agreement' here refers to some measure of overlap with the advice). When teacher's and EPST do not overlap, it is the EPST advice that has greater predictive value than the teacher's advice for pupils' eventual upwards mobility: more than a quarter (25.8%) of the pupils make a switch upward to a level that matches the EPST advice exclusively. Less than 10% switch upward to a level that matches the teacher's advice exclusively. In the cases of downward mobility, it is the teacher's advice that has a (somewhat) better predictive value for the level pupils eventually take: 18.7% make a downward switch to a level that was exclusively advised by the teacher, compared to 11.5% foreseen exclusively by the EPST advice.

Adjustment in light of a higher EPST advice

Above, we mentioned that few schools adjust the teacher's advice to a higher level in light of higher EPST results. Seeing that more than a quarter of the pupils making an upward switch end up in a level that matches the EPST advice, the question is whether this restraint in upward adjustments is justified. In this section, we therefore posit the following hypothetical question: Would pupils have benefitted from a mandatory rather than an optional adjustment of the teacher's advice, if the test results indicated a higher SE level? We will answer this question by considering the placement percentages for 2017-2018, that is, the percentages of pupils with a certain teacher's and EPST advice that end up at vmbo, havo, or vwo.

For example, when considering the 37,332 pupils with an unequivocal vmbo advice (meaning that the teacher and the EPST both advice for vmbo), we see that only 2.1%

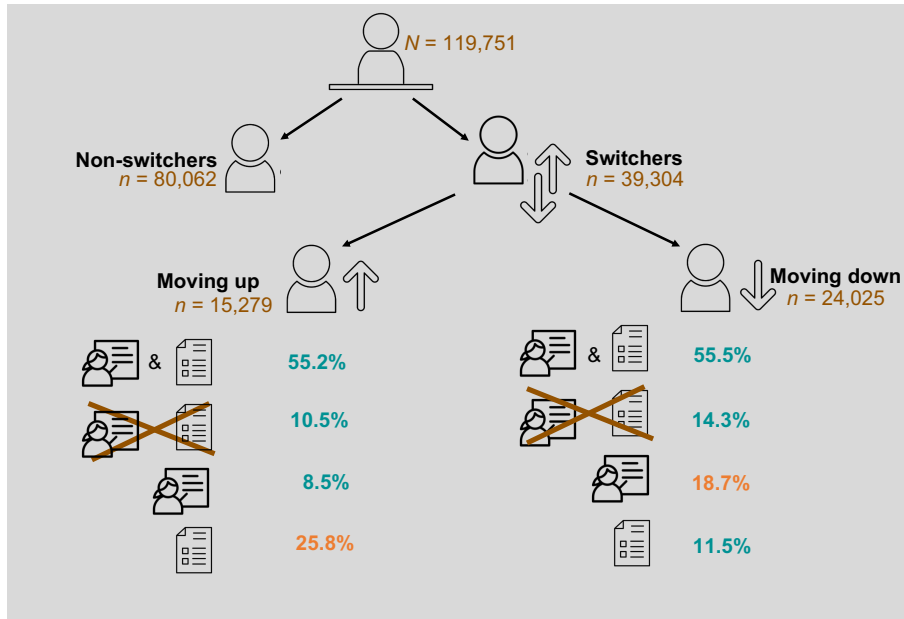


Figure 6.4: Schematic overview of pupils' switching-behaviour, also showing the percentage of switching-behaviour that can be explained with the original teacher's and EPST advices.

of the pupils with an unequivocal vmbo advice end up at the havo level in 2017-2018. When considering the 8,710 pupils with a vmbo teacher's advice and a vmbo-havo EPST advice, we see that in this group 10.1% of the pupils end up at the havo level in 2017-2018. The difference between the two percentages (2.1% unequivocal advice, 10.1% diverging advice) is interesting, because it tells us that the chance of eventual havo placement after three years is higher with a vmbo-havo EPST advice than with a vmbo EPST advice. This means that for some pupils with a teacher's vmbo advice and a vmbo-havo EPST advice, it would have been fitting to adjust the teacher's advice to a vmbo-havo advice. In the same way we have considered the other placement percentages with all the possible combinations of the teacher's advice and a higher EPST advice, comparing them to the placement percentages with unequivocal vmbo, havo, or vwo advice. All comparisons are described in detail in Appendix 6.D (see <https://osf.io/qrg4e/>).

In the following sections we will discuss the most important conclusions of these comparative analyses. We will also give a few examples. We split this up into two sections: the next section discusses the pupils where teacher's and EPST advice show overlap; the section thereafter discusses the pupils with no overlap of teacher's and EPST advice. The first group comprises almost three quarters of all the pupils with a higher EPST advice.

Teacher's and EPST advice show overlap

Overlap between teacher's and EPST advice occurs when either of the two advices constitutes a composite advice, for example a havo-vwo teacher's advice with a vwo EPST advice, or a havo teacher's advice with a havo-vwo EPST advice. In the case of a composite teacher's advice, relatively many pupils switch to the level that does not match

the single EPST advice. In the example of a havo-vwo teacher's advice with a vwo EPST advice, 30.0% of the pupils end up at the havo level in 2017-2018. This percentage is much lower for pupils with an unequivocal vwo advice (6.4%). In the case of a composite EPST advice, relatively many pupils switch to the level that does not match the single teacher's advice. In the example of a havo teacher's advice with a havo-vwo EPST advice, 19.2% of the pupils make the upward move to vwo. This is almost twice as high as the percentage of pupils with an unequivocal havo advice that end up at the vwo level (10.1%). On the basis of these two examples (and the other examples in Appendix 6.D) it seems fitting to take the composite teacher's or EPST advice as leading.

Teacher's and EPST advice do not overlap

A lack of overlap between teacher's and EPST advice occurs mainly when both the teacher's advice and the EPST advice are single, like a havo teacher's advice and a vwo EPST advice. In such cases the percentages of upward mobility are relatively high. In the example of a havo teacher's advice with a vwo EPST advice, 36.9% of these pupils end up at the vwo level. This is a much higher percentage than the percentage of vwo pupils with an unequivocal havo advice (10.1%).

The other combinations of single teacher's and EPST advice show a similar pattern. In other words, even though pupils are still most often placed in accordance with their teacher's advice in 2017-2018, a relatively high number of pupils perform wholly or partially according to their EPST advice. In the example of a havo teacher's advice and a vwo EPST advice, therefore, adjusting the teacher's advice to a havo-vwo advice might have been the right decision.

Conclusion

A controversial legislation change was implemented in the school year of 2014-2015 in the Netherlands, stipulating that the advice of the teacher instead of the advice resulting from the End of Primary School Test (EPST) was to be leading in the admission procedure to secondary education. This article uses CBS (Netherlands Statistics) data to review how pupils fared after making the transition to secondary education at the end of 2014-2015. In the following, we answer each of the four questions listed in the introduction in a summary way.

1. Can we speak of a (systematic) discrepancy between the teacher's advice and the EPST advice in 2014-2015? Despite the partial overlap between teacher's and EPST advice, we can indeed establish a systematic discrepancy for the 2014-2015 cohort. The EPST advice was often higher and it was also more often a composite advice (a combination of two adjacent education levels).
2. In retrospect, how do EPST and teacher's advice compare for predicting placement at a certain education level? The vast majority of pupils were taking an education level that matched both the initial teacher's advice and the EPST advice after three years. If this was not the case, we saw more often a match with the teacher's advice for pupils ending up at the vmbo level; whereas at the vwo level a match with the

EPST advice occurred more often. For *havo*, pupils were also placed more often in agreement with the teacher's advice, though the gap in percentages between teacher and EPST advice was smaller.

3. How often have pupils switched between the different SE levels and can these switches be explained by the teacher's advice and/or the test advice? Especially pupils with a *havo* teacher's advice turned out to switch relatively often from one education level to another between year 1 and 3 of secondary education. Some extent of mobility also occurred in the other two levels. In the case of pupils switching upwards, this was oftener to a level that matched the original EPST advice. Pupils switching to a lower level, on the other hand, more often made a switch in accordance with the original teacher's advice.
4. Would pupils have benefitted from a mandatory rather than an optional adjustment of the teacher's advice, if the test results indicated a higher SE level? On the whole it seems to be beneficial to combine the teacher's evaluation and the EPST results into a composite advice containing either both single advices or one that contains the composite advice of either teacher or EPST.

Discussion

The above findings must be placed somewhat in perspective. In the present article, for example, we consider the predictive value of EPST and teacher's advice after three years. Although this certainly gives an indication of the (added) value of both these advices, it does not completely answer the question of which is more accurate. Pupils' careers in SE are subject to all kinds of unpredictable factors that determine actual study success ('Onderadvisering in beeld', Inspectorate of Education, 2007), like educational circumstances (poor or favourable) or circumstances at the level of the adolescent.

An example of school-level factors is the tendency to reduce the 'broad stream' mix of all SE segments to a one-year period (the first year) and increasingly situate the different segments *vmbo*, *havo*, and *vwo* at separate locations (Van De Werfhorst, Elffers & Karsten, 2015). Another reason for the discrepancy between initial and eventual SE placement can sometimes be found, according to the Inspectorate of Education, in differences of pedagogical-didactic approaches between PE and SE (Inspectorate of Education, 2007; also see Korpershoek et al., 2016).

Furthermore, the teacher's advice and the EPST advice differ in the use of single as opposed to composite categories. Pupils less often receive a composite advice from teachers than from the EPST in 2014-2015. The limited use of composite advice can have different causes. With the new legislation, for example, teachers could only advise a maximum of two school types, where before there was no maximum. The idea behind this decision was that schools would thus be encouraged to give a well-considered and focused advice, but possibly this has also led to a larger number of single advices (See the House of Representatives Letter 'Tussenevaluatie wet eindtoetsing PO', 26 January 2016). In addition, in many regions covenants between PE and SE schools were made to the effect that teachers would solely or preferably write single advices.

An example of circumstances at the level of the adolescent is the case of the advice being too low. The pupil might conform to the lower educational level. A switch to a

higher, more suitable level will not occur; the too low advice then becomes a self-fulfilling prophecy.

Finally, it is also important to point out that we draw our findings from a single cohort. Our conclusions are valid for the pupils who received their teacher's and EPST advices in 2014-2015. We did not examine possible differences with earlier and later cohorts.

Practical Implications

In spite of the above-mentioned limitations, the present article offers a clear indication that the teacher's advice and the EPST advice each have certain benefits. The answer to the question of who knows better, the teacher or the End of Primary School Test – as in the title of this article – therefore is not that one is always superior to the other. It turns out that the teacher's advice more accurately predicts the eventual placement of especially vmbo pupils and to a lesser extent of havo pupils, whereas the EPST predicts better which pupils will end up at vwo level. Also, pupils who made an upward switch tend to end up at a level that agrees with the EPST advice, whereas for downward switchers the move is more in agreement with the teacher's advice.

In order to make optimal use of both advices, we therefore recommend combining them. Pupils then would not be judged by the outcomes of a single test, nor would they be too strongly dependent on the judgment of a teacher. A practical solution is to stimulate composite advices that comprise both teacher's and EPST advice. The proposal to include the EPST advice in the teacher's advice as well as aim more often for composite advices agrees with the sector view formulated by the Primary Education Council (PO-raad, 2018). It also agrees with the policy change in the 2018-2019 school year that the advice resulting from the EPST-score is always composite (with the exception of the single vwo advice)⁴⁰.

Note that our outcomes cannot be applied on a one-to-one basis in advising pupils for the transition to secondary education. In other words, though our findings show that generally the EPST advice is better at predicting which pupils end up at the vwo level, this should not be taken to mean that every vwo EPST advice should take precedence over the teacher's advice. Similarly, for the teacher's advice: not every vmbo advice by teachers is justified. Writing an advice continues to be a matter of careful consideration.

All in all, then, our message is not to opt for either the teacher's or the EPST advice, but to look for the optimal combination of both types of advice for each pupil. As for standardising such an optimal combination: here is a role for future research.

Chapter 7

HOW THE CHOICE OF DISTANCE MEASURE INFLUENCES THE DETECTION OF PRIOR-DATA CONFLICT



Kimberley Lek & Rens Van De Schoot

Published in Entropy, 2019, Volume 21, issue 5

Chapter 7

How the Choice of Distance Measure Influences the Detection of Prior-Data Conflict

Abstract

The present chapter contrasts two related criteria for the evaluation of prior-data conflict: the Data Agreement Criterion (DAC; Bousquet, 2008) and the Criterion of Nott et al. (2016). One aspect that these criteria have in common is that they depend on a distance measure, of which dozens are available, but so far, only the Kullback-Leibler has been used. We describe and compare both criteria to determine whether a different choice of distance measure might impact the results. By means of a simulation study, we investigate how the choice of a specific distance measure influences the detection of prior-data conflict. The DAC seems more susceptible to the choice of distance measure, while the criterion of Nott et al. seems to lead to reasonably comparable conclusions of prior-data conflict, regardless of the distance measure choice. We conclude with some practical suggestions for the user of the DAC and the criterion of Nott et al.

Keywords: prior-data conflict; distance measure; Kullback-Leibler; data agreement criterion

Introduction

Any Bayesian model consists of at least two ingredients: a prior and a sampling model for the observed data. The prior contains the a priori beliefs about the parameter(s) and can, for instance, be based on expert knowledge, if elicitation methods are used (O'Hagen et al., 2006; Kuhnert, Martin & Griffiths, 2010; O'Leary et al., 2009; Martin et al., 2011; Zondervan-Zwijnenburg, Van De Schoot-Hubeek, Lek, Hoijsink & Van De Schoot, 2017; Lek & Van De Schoot, 2018). Consecutive inferences based on the Bayesian model are built on the assumption that the (expert) prior is appropriate for the collected data. In case of a prior data conflict, this assumption is violated. In such a conflict, the prior primarily favors regions of the parameter space that are far from the data mass. Such a conflict can be caused by the unfortunate collection of a rare data set, a problem

with the prior, or both of these factors (Bouquet, 2008). To avoid erroneous inferences, checking for prior-data conflict should be ‘part of good statistical practice’ (Evans & Moshonov, 2006, p. 894). In the current chapter, we focus on two checks, developed for the detection of prior-data conflict. First, Bousquet (see Evans & Moshonov, 2006; see also Veen, Stoel, Schalken, Mulder, Van De Schoot, 2018) suggested that the expert prior be compared with a non-informative, reference prior in his “Data Agreement Criterion” (DAC). When the expert prior has a larger distance to a reference posterior—a posterior based on the data and the non-informative, reference prior—than the non-informative prior, it can be concluded that the expert and data are in conflict. Second, Nott et al. (see Nott, Xueou, Evans & Englert, 2016) suggested that the distance between the expert prior and resulting posterior should be measured directly. According to this method, the expert and data are in conflict when this distance is surprising in relation to the expert’s prior predictive distribution (see also Hoijtink & Van De Schoot, 2018). One thing that the DAC (Bouquet, 2008) and the criterion of Nott et al. (2016) have in common is that they depend on a distance measure. Currently, both the DAC and the Nott et al. criteria rely on the Kullback-Leibler divergence (heretofore abbreviated as KL). One of the advantages of the KL is that it has an intuitive interpretation—the informative regret due to the prior choice—and some favorable analytical properties, such as its invariance to reparameterization (Kullback & Leibler, 1951). However, the KL is not the only available option for the measurement of distance between statistical distributions. To illustrate: the “Encyclopedia of distances” (Deza, 2009) consists of 583 pages filled with distance measures, and their list is not even exhaustive. How the DAC and the Nott et al. criteria behave for different distance measure choices has yet to be investigated. Therefore, the goal of the current chapter is to investigate how the implementation of different distance measures in the DAC and Nott et al. criteria influence the detection of prior-data conflict. In the remainder, we first discuss both prior-data conflict criteria in detail. Thereafter, we discuss the design and results of a simulation study into the effect of the choice of distances on the conclusion of prior-data conflict, using a variety of distance measures (also see the Appendix for an overview; <https://osf.io/qrg4e/>). We end with a general discussion and practical recommendations for users, who would like to test for a prior-data conflict.

Prior-Data Conflict Criteria

DAC

Computation of Prior-Data Conflict

The Data Agreement Criterion (DAC; Bouquet, 2008) is based on the ratio of two distances, denoted by Δ . The first distance is between the expert prior $\pi_i^E(\theta)$ and the posterior $\pi^B(\theta \mid \mathbf{y}_n)$, where θ denotes the parameter of interest, and upper script B is used for a ‘benchmark’, upper script E for an ‘expert’, and subscript $i, 1, \dots, I$ to distinguish between experts. As reflected in the expression $\pi^B(\theta \mid \mathbf{y}_n)$, the posterior is based on dataset \mathbf{y}_n and prior $\pi^B(\theta)$. The idea is to choose $\pi^B(\theta)$ such that the posterior is dominated by the data \mathbf{y}_n . Following Bernardo (1979) and Berger, Bernardo & Sum (2009), in order to affect the posterior as little as possible, a reference prior should be

chosen for $\pi^B(\theta)$. The second distance is between the posterior $\pi^B(\theta \mid \mathbf{y}_n)$ and the non-informative prior $\pi^B(\theta)$. The ratio between the two distances results in the DAC:

$$\text{DAC}_i = \frac{\Delta\left(\pi^B(\theta \mid \mathbf{y}_n) \parallel \pi_i^E(\theta)\right)}{\Delta\left(\pi^B(\theta \mid \mathbf{y}_n) \parallel \pi^B(\theta)\right)} \quad (7.1)$$

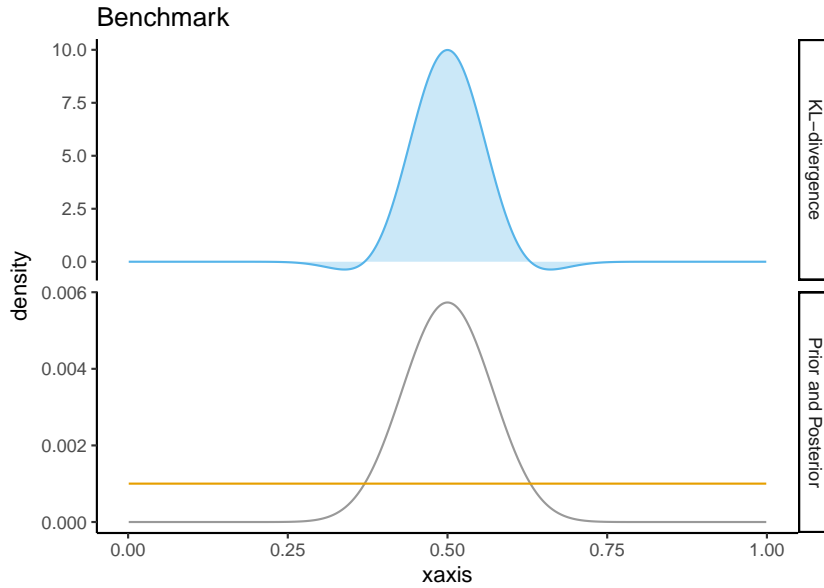
In the work of Bouquet (2008) and Veen et al. (2019), the KL-divergence is used for Δ , denoted here by Δ_{KL} . The KL-divergence expresses the loss of information that occurs when we rely on the expert prior $\pi_i^E(\theta)$, instead of on the posterior $\pi^B(\theta \mid \mathbf{y}_n)$:

$$\Delta_{KL}\left(\pi^B(\theta \mid \mathbf{y}_n) \parallel \pi_i^E(\theta)\right) = \int_{\Theta} \pi^B(\theta \mid \mathbf{y}_n) \log \frac{\pi^B(\theta \mid \mathbf{y}_n)}{\pi_i^E(\theta)} d\theta. \quad (7.2)$$

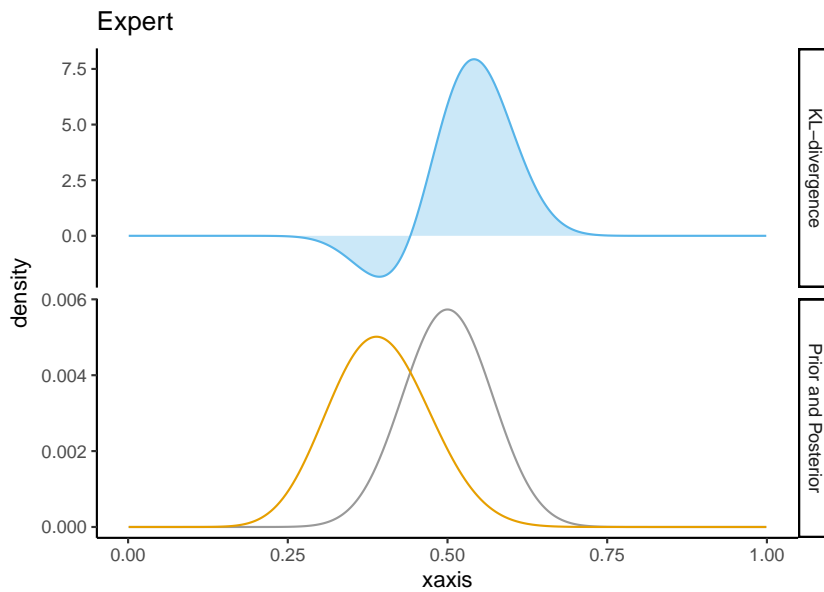
Here, Θ is the set of all possible values for the parameter θ . Figure 7.1a illustrates the KL-divergence between the prior $\pi^B(\theta)$ and posterior $\pi^B(\theta \mid \mathbf{y}_n)$. The lower part of this figure shows the prior $\pi^B(\theta)$ (orange color) and the $\pi^B(\theta \mid \mathbf{y}_n)$ (in grey), and the upper part of this figure shows the corresponding KL-divergence, which is equal to the highlighted area under the curve. Figure 7.1b illustrates the KL-divergence between the expert prior $\pi_i^E(\theta)$ and the posterior $\pi^B(\theta \mid \mathbf{y}_n)$ (see Veen et al., 2018). To compute the DAC_i the KL-divergence in the upper part of Figure 7.1b is divided by the KL-divergence in the upper part of Figure 7.1a. In this example, the KL-divergence for the prior $\pi^B(\theta)$ is 1.26, and the KL-divergence for the expert prior $\pi_i^E(\theta)$ equals 0.89. Since this value is lower than 1.26 (the ‘‘benchmark’’ KL-divergence), we would, in this example, conclude that there is no prior-data conflict, according to the DAC criterion. Note that the DAC (Equation (7.1)) does not necessarily have to be based on the KL-divergence (Equation (7.2)); any other distance or divergence measure can be substituted for Δ .

Definition of Prior-Data Conflict

As is shown by Equation (7.1), the distance between the prior $\pi^B(\theta)$ and posterior $\pi^B(\theta \mid \mathbf{y}_n)$ serves as a benchmark. When the distance between the expert prior $\pi_i^E(\theta)$ and posterior $\pi^B(\theta \mid \mathbf{y}_n)$ exceeds the benchmark (i.e., when $\text{DAC}_i > 1$), it is concluded that there is a prior-data conflict. As stated by Bousquet (2008), $\pi^B(\theta)$ can be seen as a fictitious, oblivious expert, perfectly in agreement with \mathbf{y}_n . When this ‘ideal’, fictitious expert $\pi^B(\theta)$ is more in agreement with $\pi^B(\theta \mid \mathbf{y}_n)$ than the actual expert, it is concluded that there is a prior-data conflict. According to Bousquet (2008), this happens in the following two scenarios: (1) when the expert ($\pi_i^E(\theta)$) favors the regions of θ that are far from the data mass (conflict in location), or (2) when the prior information on θ is far more precise than the information from the data \mathbf{y}_n (conflict in information uncertainty). The latter scenario mainly arises with small data sets. Note that a prior-data conflict, as defined by Bousquet (2008), does not necessarily point to a problem with the expert prior. The DAC is developed as a simultaneous check of the sampling model, and the expert data and may both have symmetric roles in the conflict.



(a)



(b)

Figure 7.1: Illustration of the Data Agreement Criterion (DAC), with (a) the Kullback-Leibler (KL)-divergence between the posterior $\pi^B(\theta | \mathbf{y}_n)$ and non-informative prior $\pi^B(\theta)$, and (b) the KL-divergence between the posterior $\pi^B(\theta | \mathbf{y}_n)$ and an expert prior $\pi_i^E(\theta)$. Note the difference in the y-axis scale between the upper parts and lower parts of Figure (a) and (b).

Pros and cons

The DAC is appealing to the applied user because of its clear, binary decision. Other than with p -values (see Nott et al., 2016; next section), the applied user does not have to choose a threshold him- or herself. A disadvantage, however, is the necessity to specify a suitable non-informative, benchmark prior $\pi^B(\theta)$. Choosing such a prior is a delicate matter, since there are many alternative candidates to be considered, which all directly influence the conclusion and thus the definition of the prior-data conflict. This is especially problematic when an improper prior is chosen. In that case, the notion of distance gets lost in the determination of $\Delta\left(\pi^B(\theta | \mathbf{y}_n) || \pi^B(\theta)\right)$. It might also feel counterintuitive that, in order to compare an expert prior $\pi_i^E(\theta)$ with \mathbf{y}_n , the specification of another prior $\pi^B(\theta)$ is required.

Criterion of Nott et al.

Computation of Prior-Data Conflict

Inspired by the work of Box (see Box, 1980) and Evans and Moshonov (Evans & Moshonov, 2006; Evans & Jang; 2011; Evans & Moshonov, 2007), the criterion developed by Nott et al. (2016) is based on the premise of Bayesian prior predictive model checking. The prior predictive distribution informs us as to which data are considered plausible a priori by the respective expert. When the collected data \mathbf{y}_n are surprising under the prior predictive (i.e., located in the tails of the prior predictive distribution), this signals that something is wrong. More formally stated, the idea is that there is a discrepancy function $D(\mathbf{y}_n)$ of the data \mathbf{y}_n (see Cox & Hinkley, 2019) and that, for the prior predictive distribution, a Bayesian p -value is computed as (see p. 3, Hoijtink & Van De Schoot, 2018):

$$p = P\left(D(Y) \geq D(\mathbf{y}_n)\right), \quad (7.3)$$

where Y is a draw from the prior predictive distribution. A small p -value results when $D(\mathbf{y}_n)$ is large (i.e., surprising), compared to $D(Y)$. In principle, any discrepancy function can be chosen that suits a specific model-checking goal. To check for prior-data conflict, Evans and Moshonov (2006) consider $1/m_T\left(T(\mathbf{y}_n)\right)$ as their discrepancy function, where m_T is the prior predictive density of a minimal sufficient statistic T . Despite their promising results (see Evans & Jang, 2011; Evans & Moshonov, 2007), a drawback of this suggestion is that the check is not invariant in relation to the choice of the minimal sufficient statistic, when Equation (7.3) is continuous (see Jang, 2010). Nott et al. (2016) solve this problem elegantly by considering another discrepancy function: the distance between the prior $\pi_i^E(\theta)$ and posterior $\pi_i^E(\theta | \mathbf{y}_n)$; $\Delta\left(\pi_i^E(\theta | \mathbf{y}_n) || \pi_i^E(\theta)\right)$. The advantage of $\Delta\left(\pi_i^E(\theta | \mathbf{y}_n) || \pi_i^E(\theta)\right)$ is that it depends on the data \mathbf{y}_n only through the posterior distribution $\pi_i^E(\theta | \mathbf{y}_n)$. Hence, the statistic $\Delta\left(\pi_i^E(\theta | \mathbf{y}_n) || \pi_i^E(\theta)\right)$ is a function of any sufficient statistic and thus invariant in relation to the particular choice of T . To express the distance between $\pi_i^E(\theta)$ and $\pi_i^E(\theta | \mathbf{y}_n)$, Nott et al. (2016) consider a class of

divergences, of which the KL-divergence is a special case. This class contains the Renyi divergences in the order of α , here denoted by $\Delta_{R\alpha}$. $\Delta_{R\alpha}$ can be explained as a measure of how much beliefs change from prior to posterior, comparable to in relation to the notion of relative belief (see Evans, 2015). The earlier mentioned KL-divergence is a special case of this class, with $\alpha \rightarrow 1$. For the Nott et al. (2016) criterion, Equation (7.3) can thus be rewritten as:

$$p_i = P\left(\Delta_{R\alpha}\left(\pi_i^E(\theta|Y) \parallel \pi_i^E(\theta)\right) \geq \Delta_{R\alpha}\left(\pi_i^E(\theta | \mathbf{y}_n) \parallel \pi_i^E(\theta)\right)\right), \quad (7.4)$$

where

$$\Delta_{R\alpha}\left(\pi_i^E(\theta | \mathbf{y}_n) \parallel \pi_i^E(\theta)\right) = \frac{1}{\alpha - 1} \log \int_{\Theta} \pi_i^E(\theta | \mathbf{y}_n) \left[\frac{\pi_i^E(\theta | \mathbf{y}_n)}{\pi_i^E(\theta)}\right]^{\alpha-1} d\theta, \quad (7.5a)$$

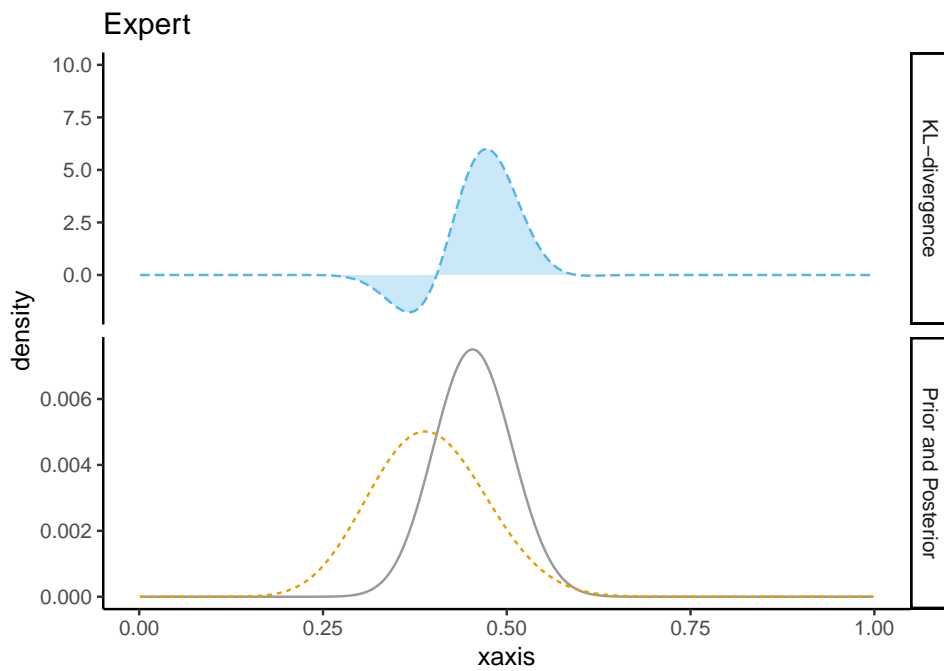
and

$$\Delta_{R\alpha}\left(\pi_i^E(\theta | Y) \parallel \pi_i^E(\theta)\right) = \frac{1}{\alpha - 1} \log \int_{\Theta} \pi_i^E(\theta | Y) \left[\frac{\pi_i^E(\theta | Y)}{\pi_i^E(\theta)}\right]^{\alpha-1} d\theta. \quad (7.5b)$$

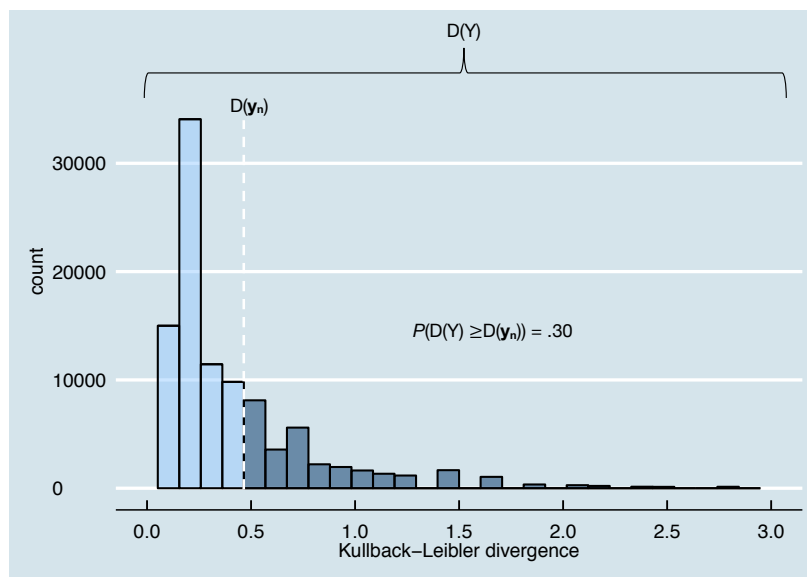
Figure 7.2 illustrates the criterion of Nott et al. for the same $\pi_i^E(\theta)$ and \mathbf{y}_n , as used in Figure 7.1. In Figure 7.2a, the KL-divergence between the prior $\pi_i^E(\theta)$ (lower panel; orange color) and posterior $\pi_i^E(\theta|\mathbf{y}_n)$ (lower panel; in grey) is illustrated. The area under the curve is equal to 0.43. In Figure 7.2b, this value (0.43) is compared to 10^5 KL-divergences, obtained by repeatedly drawing Y from the prior predictive distribution of the expert i . In this example, 30% of the KL-divergences in Figure 7.2b exceed 0.43 (see the dark blue bins in Figure 7.2b). We would therefore conclude that there is no prior-data conflict in this example. Note that, in prior-predictive checking, the user has considerable significant freedom in choosing a suitable discrepancy function $D(\mathbf{y}_n)$. Therefore, the distance measure $\Delta_{R\alpha}$ in the criterion of Nott et al.(2016) can be replaced by any other distance or divergence measure.

Definition of Prior-Data Conflict

In the work of Nott et al. (2016), it is found that prior-data conflict exists when the observed distance between the prior $\pi_i^E(\theta)$ and posterior $\pi_i^E(\theta | \mathbf{y}_n)$ is surprising in relation to the corresponding prior predictive distribution. In other words, if the data \mathbf{y}_n would have been in line with the prior $\pi_i^E(\theta)$, we would not expect to observe such a large distance $\Delta_{R\alpha}\left(\pi_i^E(\theta | \mathbf{y}_n) \parallel \pi_i^E(\theta)\right)$. Hence, the prior $\pi_i^E(\theta)$ and the data \mathbf{y}_n appear to be in conflict. Note that this definition of prior-data conflict is vastly different from the earlier definition of the DAC. First, although both rely on the KL-divergence, in the criterion of Nott et al., the prior $\pi_i^E(\theta)$ is directly compared to its corresponding posterior $\pi_i^E(\theta | \mathbf{y}_n)$, instead of $\pi^B(\theta | \mathbf{y}_n)$, as in the DAC. The criterion of Nott et al. (2016) and the DAC therefore ask different questions: how much information do we lose when relying on $\pi_i^E(\theta)$, instead of $\pi^B(\theta | \mathbf{y}_n)$ (DAC)? To what extent does our information change, when moving from $\pi_i^E(\theta)$ to $\pi_i^E(\theta | \mathbf{y}_n)$ (Nott et al.)? Second, the



(a)



(b)

Figure 7.2: Illustration of the criterion of Nott et al. (2016), with (a) the KL-divergence between the posterior $\pi_i^E(\theta | \mathbf{y}_n)$ and expert prior $\pi_i^E(\theta)$, i.e., $D_i(\mathbf{y}_n)$ (see Equation 4a), and (b) the distribution of KL-divergences between the posterior $\pi_i^E(\theta | Y)$ and expert prior $\pi_i^E(\theta)$, for 105 draws y from the prior predictive.

observed values $\Delta_{KL}\left(\pi^B(\theta | \mathbf{y}_n) \parallel \pi_i^E(\theta)\right)$ and $\Delta_{R\alpha}\left(\pi_i^E(\theta | \mathbf{y}_n) \parallel \pi_i^E(\theta)\right)$ are calibrated differently. In the DAC, prior-data conflict exists when more information is lost than with the benchmark prior $\pi^B(\theta)$. In the criterion of Nott et al., the p -value in Equation (7.4) calibrates the change in information by expressing how surprising this change is. Third, the uncertainty of the expert (i.e., the variance of $\pi_i^E(\theta)$) is treated differently in the criterion of Nott et al. than in the DAC. In the DAC, prior-data conflict can occur when there is no mismatch in location (i.e., both $\pi_i^E(\theta)$ and $\pi^B(\theta | \mathbf{y}_n)$ place their mass primarily in the same region of θ) but there is a mismatch in information uncertainty (i.e., the variance of $\pi_i^E(\theta)$ is smaller than the variance of $\pi^B(\theta | \mathbf{y}_n)$). In the criterion of Nott et al., the uncertainty of the expert simply reflects the variety of data that are deemed plausible by this expert, as captured in the prior predictive distribution. Therefore, when the location of the prior and posterior match, we do not find a prior-data conflict with the criterion of Nott et al., even if that prior variance is relatively small. Finally, the DAC and the criterion of Nott et al. respond differently, when little data is available. With a small sample size, the criterion of Nott et al. may fail to detect an existing prior-data conflict. A bad prior may pass the test, as the relatively large sampling variability is taken into account in the comparison with the prior predictive. In the DAC, the sampling variability is not taken into consideration; $\pi^B(\theta|\mathbf{y}_n)$ is treated as a temporary ‘truth’. Therefore, we should be careful not to falsely interpret a prior-data conflict as a problem with the prior in a situation where there is little data. Such a conflict may very well be the result of the unavailability of representative data and a mismatch in the amount of information captured in $\pi_i^E(\theta)$ and $\pi^B(\theta | \mathbf{y}_n)$.

Pros and Cons

One major advantage is that - aside from Bousquet (2008) - the criterion of Nott et al. (2016) does not depend on the definition of a non-informative prior $\pi^B(\theta)$. A possible disadvantage, however, is that the p -value in Equation (7.4) is often misinterpreted in practice (see Bernardo & Smith, 2000). The p -value in Equation (7.4) should be used as a measure of surprise rather than support (Jang, 2010), and not be mistaken for the probability of a (non-)conflict between $\pi_i^E(\theta)$ and $\pi_i^E(\theta | \mathbf{y}_n)$. Given that the definition of prior-data conflict is quite different for the DAC and the criterion of Nott et al., choosing one or the other should be based on a clear and well-informed opinion of what prior-data conflict entails.

Simulation Study

Goal of the Simulation

In principle, the KL-divergence in Equations (7.1) and (7.4) can be replaced by any other distance measure. The goal of our simulation is therefore to investigate the role of the distance measure in the detection of prior-data conflict. The main question is: How much does the choice of distance measure eventually impact the conclusions of prior-data conflict by the DAC and the criterion of Nott et al.? In other words: How robust are the DAC and the criterion of Nott et al. in relation to the choice of distance measure?

Simulation Design

Scenario

In our simulation, we focus on one example scenario, in which θ is a one-dimensional parameter. Specifically, $\mathbf{y}_n \sim \text{Binomial}(n, \theta)$ and $\pi_i^E(\theta)$ is a conjugate Beta distribution $\text{beta}(\alpha, \beta)$ on $[0,1]$. In this scenario, $\pi^B(\theta)$ — the benchmark prior in the DAC — is set to $\text{beta}(\alpha = 1, \beta = 1)$, and the prior predictive distribution for the criterion of Nott et al. equals a Beta Binomial, with parameters α, β and n . θ set to .5 in the population, whereas α, β and n are varied in the simulation. Note that α and β together express the location of the expert prior (i.e., the region with most of the prior’s mass), $\frac{\alpha}{\alpha + \beta}$, and the (un)certainty of the expert. $\alpha + \beta \cdot \frac{\alpha}{\alpha + \beta}$ are set to .05, .07, .09, \dots .95, respectively, and $\alpha + \beta$ is varied from 10, 12, 14, \dots 200. The sample size of the data, n , is varied between 50, 100 and 200. As explained previously, twelve distance measures are compared. Figure 7.3 summarizes the simulation design.

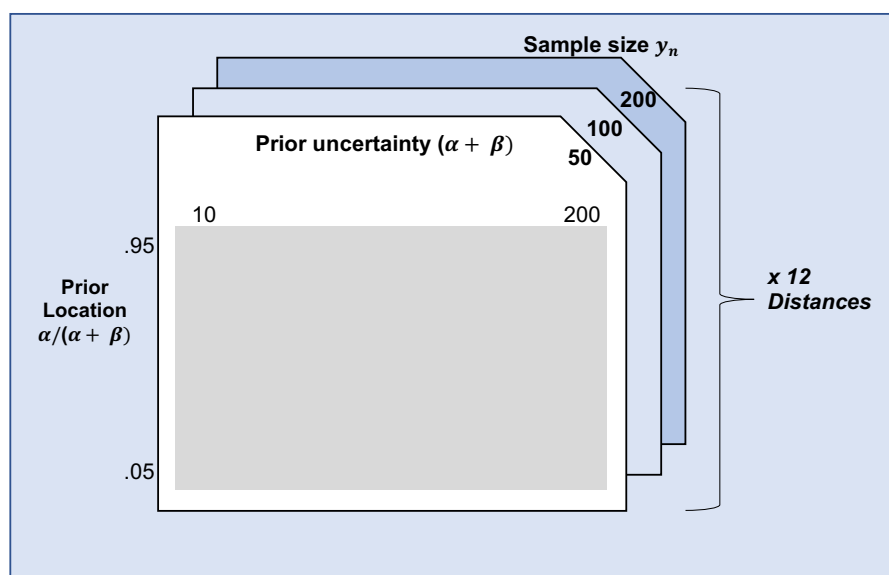


Figure 7.3: Illustration of the simulation design. The prior location $\frac{\alpha}{\alpha + \beta}$, prior uncertainty $\alpha + \beta$, and sample size of \mathbf{y}_n were varied, as well as the distance measure used.

Steps

We started the simulation by taking 1,000 samples from the respective population, with $\theta = .5$ and sample sizes equal to 50. We repeated this sampling process for the other sample sizes ($n = 100$ and 200). Subsequently, multiple expert priors $\pi_i^E(\theta)$ were constructed, based on all possible combinations of the location (i.e., $\frac{\alpha}{\alpha + \beta} =$

0.05, \dots 0.95) and expert uncertainty (i.e., $\alpha + \beta = 10, 12, \dots, 200$). For all these expert priors, we then determined $\pi_i^E(\theta \mid \mathbf{y}_n)$ (for the criterion of Nott et al.) and $\pi^B(\theta \mid \mathbf{y}_n)$ (for the DAC) and calculated their respective distances to $\pi_i^E(\theta)$ using the variety of distance measures, named in the next section. The calculation of the DAC and p -value of the criterion of Nott et al. (2016) followed, using Equations (7.1) and (7.4), respectively. Finally, all DAC values above 1 were flagged as prior-data conflicts. For the criterion of Nott et al. (2016), we considered $p \leq .05$ as indicative of a prior-data conflict. All analyses were performed using RStudio (RStudio Core Team, 2015) and the R packages, “Philentropy” (Drost, 2019), “distr” (Ruckdeschel, Kohl, Stabla & Camphausen, 2006) and “distrEx” (Ruckdeschel et al., 2006).

Distance measures

The KL-divergence, used in both the DAC and the criterion of Nott et al., is an instance of an f -divergence, also known as an Ali-Silvey divergence (Ali & Silvey, 1966). The more general Rényi divergences, considered in Nott et al., are related to this class. One of the distance measures we investigate in our simulation (the total variation distance) belongs to the class of f -divergences (see Liese & Vajda, 2006). We also consider distance measures that are not part of this class: Hellinger, Kolmogorov, Euclidean, Manhattan, Sorensen, Intersection, Harmonic mean, Bhattacharyya, Divergence, Jeffreys and Jensen-Shannon (leading to twelve distance measures in total). See Cha (2007) for more information on these distance measures.

Results

As stated previously, the main goal of the simulation was to investigate the robustness of the DAC and the criterion of Nott et al. in relation to the choice of the twelve distance measures under study. Below, we split the results of the simulation into two parts. In the first part, we look at how the conclusion of prior-data conflict varies with different choices of distance measures, when we vary the expert priors for a specific sample \mathbf{y}_n (see Figure 7.3). Ideally, whether or not the varying expert priors are found to be in conflict with the specific sample should not depend on the choice of distance measure. In this first part, we act as if this fixed, single sample is ‘the truth’. Of course, the sample is drawn from a population. Therefore, in the second part, we look at varying samples \mathbf{y}_n and specific expert priors. This way, we can investigate how sampling variability influences the behavior of the DAC and the criterion of Nott et al., when different distance measures are used.

Robustness in Relation to the Choice of Distance Measure: Specific Sample and Varying Expert Priors

In this section, we investigate varying expert priors to determine whether they are in conflict with one specific sample, $\mathbf{y}_n \sim \text{Binomial}(n = 100, \theta = 0.5)$. To limit the number of plots here, a part of the plots is shown in the Appendix. Specifically, the plots in the Appendix visualize the DAC and p -values for varying expert priors, for each of the twelve

distance measures separately. The x-axes show the expert prior uncertainty $(\alpha + \beta)$ and the y-axis prior location $\frac{\alpha}{(\alpha + \beta)}$, such that, for every combination of uncertainty and location, the DAC and p -value can be read from the plots. The contour lines show which expert priors lead to comparable DAC and p -values, such as $\text{DAC} = 0.2, 0.3, \dots, 1.1$ and $p = 0.05, 0.1, \dots, 0.9$. The most interesting contour lines are $\text{DAC} = 1$ and $p = 0.05$, since these lines distinguish the expert priors, for which a prior-data conflict and no prior-data conflict is concluded. As seen in the DAC plots, the Jeffreys divergence leads to the most lenient decision of a prior-data conflict. Indeed, when Jeffreys divergence is used, relatively many expert priors lead to a DAC value below 1 (i.e., many combinations of expert uncertainty and location fall in the area of DAC values < 1). Kolmogorov, on the other hand, leads to the most stringent decision of prior-data conflict. All divergence-based measures show problematic behavior. Using ‘divergence’, DAC values below 1 are solely found for a small group of expert priors. In the criterion of Nott et al., divergence shows divergent behavior, especially for expert priors with a low $\alpha + \beta$. To a lesser extent, this problematic behavior at low values of $\alpha + \beta$ is also found when using the Euclidean distance in the criterion of Nott et al. Apart from the divergence and Euclidean distance, the criterion of Nott et al. generally shows behavior more comparable to the p -value for different distance measures than the DAC. Figure 7.4a combines the twelve DAC plots from the Appendix into one plot, in which only the contour lines, ‘DAC = 1’, are displayed. By doing so, these DAC = 1 boundaries can more easily be compared over the twelve distance measures. Figure 7.4b does the same for the twelve p -value plots from the Appendix, for the contour lines $p = 0.05$. The contour lines (DAC = 1 and $p = 0.05$), of the distance measure, which leads to the most stringent conclusion of prior-data conflict, mark the area (i.e., the combinations of expert prior uncertainty and location), for which all distance measures agree that there is no prior-data conflict. In Figure 7.4a and 7.4b, this area is colored light blue. As stated before, in Figure 7.4a, this area is determined by Kolmogorov. The contour lines (DAC = 1 and $p = 0.05$), belonging to the distance measure which leads to the most lenient conclusion of prior-data conflict, mark the area for which all distance measures agree that there is a prior-data conflict. As stated before, in Figure 7.4a, this area is determined by Jeffreys divergence. In Figure 7.4a and 7.4b, this area is colored dark blue. The area in between the contour lines of the most lenient and most stringent distance measures illustrates the area of no consensus, i.e., the expert priors, for which some distance measures would lead to a conclusion of prior-data conflict and others not. The larger this area is, the more influence the choice of distance measure has on the conclusion of prior-data conflict. The white lines in this area illustrate the exact DAC = 1 and $p = 0.05$ contour lines of the distance measures. Comparing Figure 7.4a and 7.4b clearly shows that a different choice of distance measure more profoundly influences the conclusion of the DAC than the criterion of Nott et al. When used in the criterion of Nott et al., the distance measures primarily disagree with each other with for the expert priors that have $\frac{\alpha}{(\alpha + \beta)}$ far from 0.5 (0.1, 0.2, 0.8 and 0.9) and a low $(\alpha + \beta)$ (smaller than 50). The outer line separating the area with no consensus from the area with consensus on prior-data conflict is formed by the Divergence distance. Of the twelve distance measures under study here, this distance measure thus leads to the most lenient decision of prior-data conflict, at least in this specific scenario. For only a relatively small area, the distance measures agree that there is no prior-data conflict, when used in the DAC (see Figure 7.4a). Other than in the criterion of Nott et al., differences in conclusions of (no) prior-data conflict are not restricted to certain

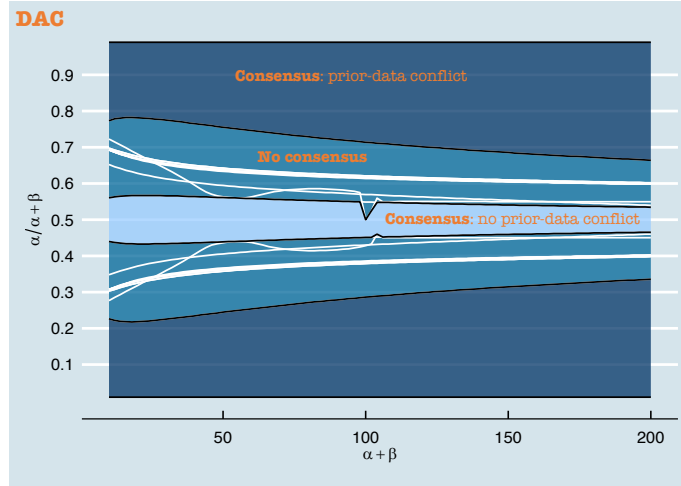
combinations of prior location $\frac{\alpha}{(\alpha + \beta)}$ and uncertainty $(\alpha + \beta)$.

Robustness in Relation to the Choice of Distance Measure: Specific Expert Prior and Varying Samples

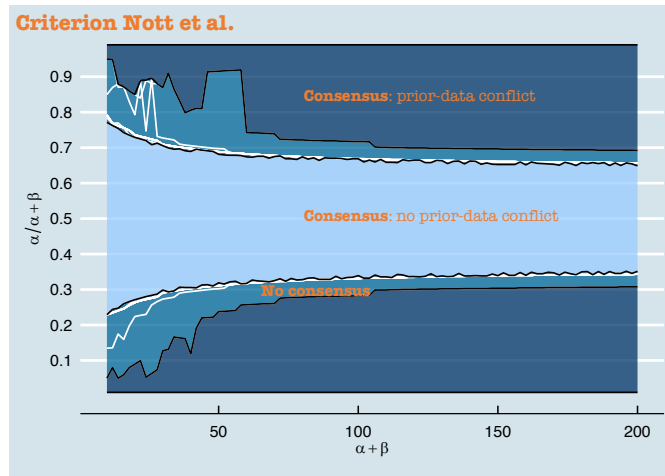
In this section, we investigate how many times the DAC and p -values lead to a conclusion of prior-data conflict, when repeated samples are drawn, and the expert prior is fixed. Figures 7.5a–7.5c show an overview of the results. The stacked histograms show the number of times the DAC and the criterion of Nott et al. do (red color) or do not (green color) reach a conclusion of prior-data conflict. Every stacked histogram is based on a Beta prior, with a certain combination of $\frac{\alpha}{(\alpha + \beta)}$ (row) and $\alpha + \beta$ (column) and 1,000 repeated samples $\mathbf{y}_n \sim \text{Binomial}(n, \theta)$ of with $\theta = 0.5$ and $n = 50$ (Figure 7.5a), $n = 100$ (Figure 7.5b) or $n = 200$ (Figure 7.5c). Every bin of the stacked histogram corresponds to one of the 12 distance measures. The first bin, for instance, corresponds with the total variance distance, and the second one corresponds with the Hellinger distance. When a bin within a stacked histogram is completely red, this means that all 1,000 repeated samples led to a conclusion of prior-data conflict, when this specific distance measure was used. The dashed horizontal line and the corresponding percentage printed at the bottom of each histogram show the average number of times the distance measures lead to a conclusion of no prior-data conflict. To ease comparison, the left side of Figure 7.5a–7.5c shows the results for the DAC, and the right side of Figure 7.5a–7.5c shows the results for the criterion of Nott et al. The most important conclusion that can be drawn from Figure 7.5a–7.5c is that the criterion of Nott et al. is less sensitive to the choice of the twelve distance measures than the DAC. Comparing the right side (criterion Nott et al.) with the left side (DAC) of Figure 7.5a–7.5c, the coloring of the bins is more comparable over the twelve distance measures with when the Nott et al. criterion rather than the DAC is used. Only one of the twelve distance measures shows problematic behavior: “Divergence” (bin nr. 10). When “Divergence” is used as a distance measure in the criterion of Nott et al., the conclusion of prior-data conflict becomes more lenient. When $\frac{\alpha}{(\alpha + \beta)} = .2$, $\alpha + \beta = 50$ and $n = 50$, for instance, almost half of the 1,000 repeated samples would result in a conclusion of “no prior-data conflict”. Using any of the other distance measures would, however, lead to a conclusion of prior-data conflict in almost all samples.

Conclusion

The goal of the current chapter was to investigate how the implementation of different distance measures in the DAC and Nott et al. criteria influence the detection of prior-data conflict. Overall, the criterion of Nott et al. seems less sensitive to the choice of distance measure than the DAC. In the criterion of Nott et al., the user seems to have more freedom in the choice of distance measure. With the exception of one distance measure, all distance measures led to a comparable conclusion of prior-data conflict (see again Figure 7.5). This is advantageous, as even the simplest (and fastest computable) distance measures (for instance, the Manhattan distance) can be chosen, without changing the interpretation of

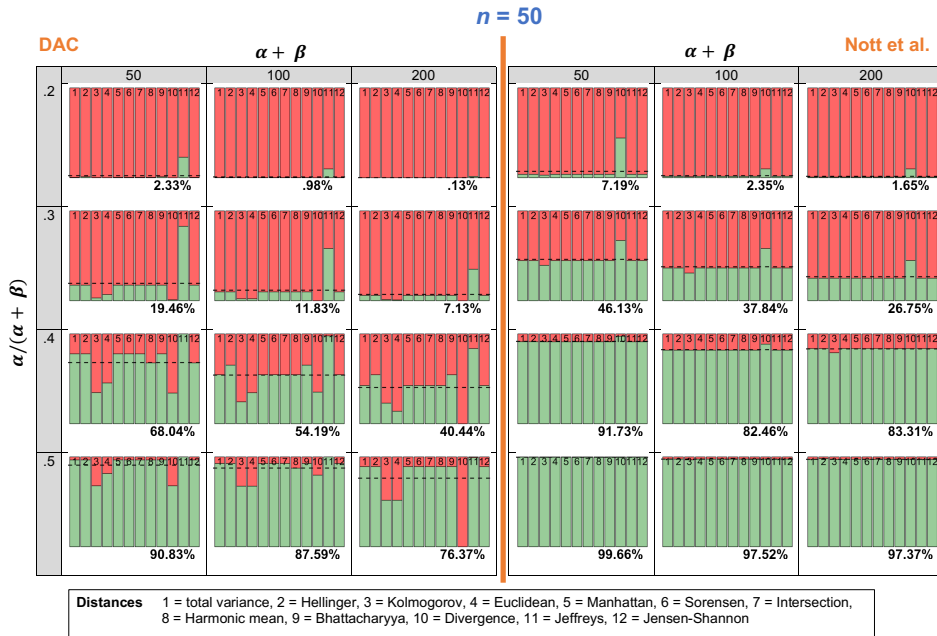


(a)

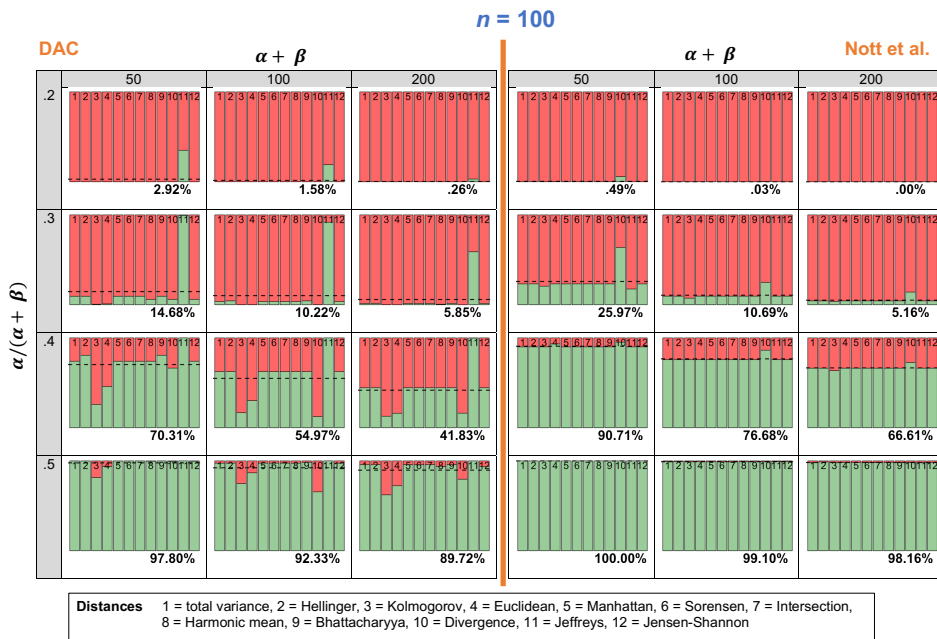


(b)

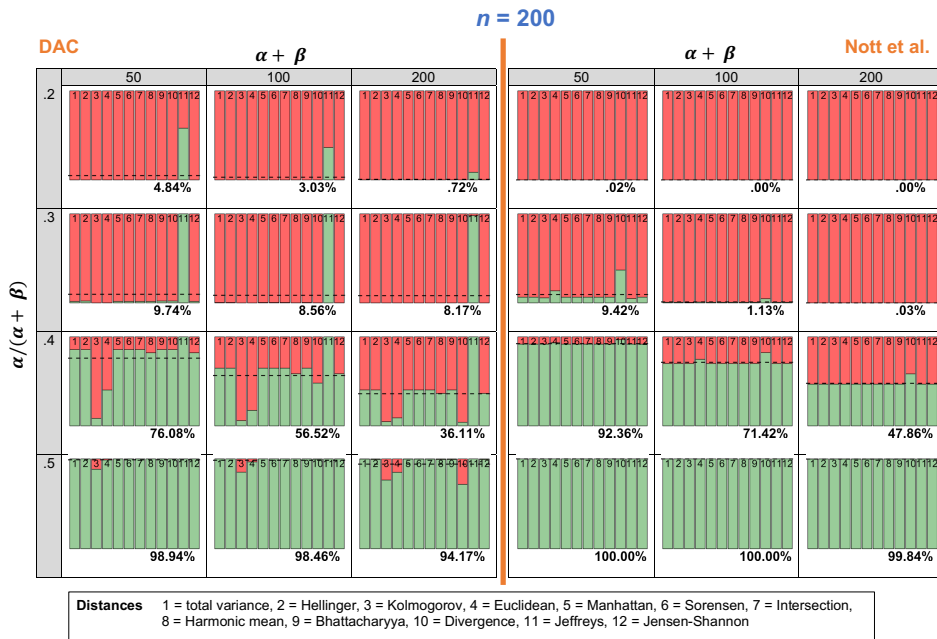
Figure 7.4: The white lines in this figure show the $DAC = 1$ (a) and $p < 0.05$ (criterion of Nott et al.) (b) boundaries for the twelve distance measures, when the sample is fixed, and $\alpha + \beta$ (x-axis) and $\frac{\alpha}{\alpha + \beta}$ (y-axis) are varied. The darkest blue area illustrates the expert priors (i.e., the combination of $\alpha + \beta$ and $\frac{\alpha}{\alpha + \beta}$) for which all distance measures would lead to a conclusion of prior-data conflict. The lightest blue area shows the expert priors that consistently lead to a conclusion of no prior-data conflict. The area in between these two areas is inconclusive and depends on the choice of distance measure.



(a)



(b)



(c)

Figure 7.5: This Figure shows how many times, out of 1,000 repeated samples $\mathbf{y}_n \sim \text{Binomial}(n, \theta)$ of $\theta = 0.5$ and $n = 50$ (a), $n = 100$ (b) or $n = 200$ (c), the twelve distance measures led to a conclusion of prior-data conflict. Every stacked histogram is based on an expert prior, with a certain combination of $\frac{\alpha}{(\alpha + \beta)} = .2, .3, .4$ or $.5$ (rows) and $\alpha + \beta = 50, 100$ or 200 (columns). The red bins correspond to a conclusion of prior-data conflict, and the green bins correspond to a conclusion of no prior-data conflict.

prior-data conflict. In our specific scenario, we did see that some distance measures led to a more lenient treatment of experts, who were very insecure and incorrect. Concerning the DAC, substituting the KL-divergence (see Equation (7.2)) for another distance measure is only encouraged, when the KL results in issues. One reason to not use the KL-divergence in the DAC, for instance, is that it can lead to computational problems (see Cha, 2007). In case of such computational problems, the best strategy is to search for a distance measure without these computational issues and with comparable behavior, so as to not jeopardize the interpretation of the DAC. Plotting the size of the DAC (or the criterion of Nott et al.), compared to the characteristics of the (possible) expert priors, as we did in Figure 7.4 and in the Appendix (see <https://osf.io/qrg4e/>), can help to spot potential problems with the distance measure. Such a plot can also help to check whether the distance measure leads to a relatively strict or relatively lenient conclusion of prior-data conflict. The results of our study are limited to a specific example scenario and a limited set of twelve distance measures. We chose this limited scope, as it was impossible to investigate the hundreds of possible distance measures and anticipate all possible scenarios. Our study can, therefore, best be read as an example of how the choice of distance measure may influence the detection of prior-data conflict with the DAC and the criterion of Nott et al. Whether or not the detection of prior-data conflict depends on the choice of distance measure in another scenario should be checked beforehand by the investigator.

Discussion

In my PhD thesis, I have had the opportunity to investigate the merits and disadvantages of teacher judgments and test results in the specific example of the transition from Dutch primary to secondary education and beyond. Below, I discuss some ideas I have about this transition, based on what I have learned during my PhD. Note that although this discussion is particularly aimed at the Dutch situation, this does not mean that these discussion points are only relevant in the Dutch context. In several other European countries (for instance Flanders [the northern part of Belgium], Germany, Luxembourg and Switzerland), entry into secondary schools is based on a selection process in which teacher judgment plays a major role (Pit-ten Cate, Krolak-Schwerdt & Glock, 2016; Van de Werfhorst & Mijs, 2010). And even in school systems that are less academically selective (for instance in Scandinavian countries, the UK and the United States; see OECD, 2016), teacher judgments impact the ongoing instructional decision-making within the classroom, ultimately influencing the academic pathway of students (Meissel, Meyer, Yao & Rubie-Davies, 2017; Sudkamp, Kaiser & Moller, 2012). Thinking about the role of teacher judgments and the possible supplementary role of (standardized) test results is therefore relevant in any educational system.

False dichotomization

The discussion surrounding the transition from primary to secondary education in the Netherlands is largely limited to two options. Option one dictates that tracking in secondary education should be based on a holistic, unaided judgment of the sixth grade teacher⁴¹. The second option is to base tracking on the result of a single, summative End-of-Primary-school test. Proponents of the latter option usually point to the potential bias of the ‘subjective’ teacher judgment (see, for instance, Timmermans et al., 2018) to justify the usage of the ‘objective’ End-of-Primary-school test (Niessen & Meijer, 2016; Dronkers, March 2013). Opponents of the latter option often mention the limited scope and time frame of the End-of-Primary-school test to express their preference for the teacher judgment (Kammerman & Vasterman, 2015). In my opinion, the dichotomization - teacher versus test - is overly simplistic. The danger of such an oversimplification is that we end up in a right and wrong game and that we fail to consider other possibilities in the transition from primary to secondary education.

Why limit ourselves to a *single* test result?

There is no reason why, for instance, we should limit ourselves to a single End-of-Primary-school test result. In the Netherlands, as in many other countries, there are elaborate monitoring systems in place with multiple rich, standardized tests for every grade (see Faber, Van Geel & Visscher, 2013; Van Geel & Visscher, 2013). Although these standardized tests are formative rather than summative⁴² and not developed for track recommendation purposes, it is worthwhile to investigate their potential (Geven et al., 2018). According to Van Aarsen (2013), the scores on tests from the Cito monitoring system⁴³ (especially grade 4 - grade 6) are more predictive of actual track placement four years later than the (Cito) End-of-Primary-school test result. Additionally, previous math test scores from the Cito monitoring system across grade four to six correlate highly with the math component of the (Cito) End-of-Primary-school test result (Hirsch, 2017). Assuming at least a temporarily stable student ability, we could therefore attempt to accumulate the results of all or some of these tests and combine these with the End-of-Primary-school test to gain a much richer understanding of the abilities of a pupil. Doing so would have advantages over the single End-of-Primary-school test. Since the test scores are correlated, the reliability of their composite for instance exceeds that of any of the single tests (He, 2009). It is even possible to combine (i.e., weight) the tests such that the reliability of the composite is maximized (He, 2009)⁴⁴. The same holds for the external validity: it is possible to maximize the relationship between the composite and an external criterion (e.g., the actual track placement a couple of years later). Furthermore, when multiple tests are combined, we do not have to fear for the phenomenon of ‘regression towards the mean’ in the track allocation process (Kelley, 1947). Because of random measurement error, the result of a *single* test can be rather extreme (low or high) compared to a pupil’s ‘true’ score. Upon retesting, the test score is likely to be closer to the ‘true’ score of the pupil and over multiple test takes, the random measurement errors tend to cancel out. Often mentioned critiques of the End-of-Primary-school test by opponents of this test could also be solved by incorporating test results from the monitoring system. The limited scope of the test would be extended, for instance, and the resulting score would reflect the pupil’s performance over a longer period of time. Shifting the emphasis away from the End-of-Primary-school test may also have the advantage that pupils feel less pressured (Jansen, 2015) and parents are less tempted to spend inordinate amounts of money on trainings for this test (Hoogstad, 2012). For the idea of combining the EPST with other tests to work, the purpose of the monitoring system should primarily remain to inform the teacher about the learning opportunities of pupils (i.e., formative testing), with information for the track recommendation in sixth grade as a ‘by-product’. Otherwise, extending the EPST with tests from the monitoring system might only result in an earlier selection and therewith a narrowing of the curriculum in the direction of the content and skills that are tested (i.e., teaching to the test; see Brennan, 2006). It is important to realize that the balance between formative and summative testing is delicate, as pupils should have the opportunity to make mistakes and learn from them, without the fear that these mistakes have a long lasting effect on their eventual track recommendation. Additionally, when low-stakes tests get a ‘high-stakes role’, we should be aware of the fact that low- and high stakes test conditions differ, influencing pupil’s obtained test scores (see Brennan, 2006).

Why should the teacher recommendation be reached *holistically*?

There is also no reason why the teacher's judgment should be reached holistically. All the steps involved in the judgment process (i.e., selection of useful data sources about a pupil, (re)collection of this data, weighting of the data to come to a decision and reweighting the data when the End-of-Primary-school test result is higher than expected) could be completed separately in a systematic, transparent matter. Based on recent literature, we know that possible biases associated with intuitive judgement could be corrected when such a rational approach is taken (Schildkamp, Poortman & Handelzalts, 2016; Vanlommel & Schildkamp, 2019). Additionally, the reliability of the teacher recommendation could be improved. Specifically, disaggregating the otherwise overall judgment in smaller judgments reduces random measurement error, as human's cognitive abilities are limited (see the classical work of Kahneman & Tversky, 1982). According to the Dutch Inspectorate of Education, the weighting of test data by teachers could be based on, for instance, the trust teachers have in their own judgment, the value teachers attach to test results, the idea teachers have of the requirements in secondary education, the 'risks' that teachers are willing to take with their students, et cetera (2018). Performing the judgment process in a step-by-step fashion helps teachers to at least become aware of these motives, and implicit assumptions can be tested. From a research perspective, breaking down the teacher judgment process in smaller, transparent steps also has the advantage that we get to unravel the black box that teacher judgment currently is. When we are able to follow all judgments steps, we obtain a better understanding of how teachers 'make sense of data' (see Vanlommel & Schildkamp, 2019) and how they reach their final judgment.

Why should the teacher judgment be *unaided*?

Another argument in favor of a more systematic approach to teacher recommendation, is that it becomes easier to support the teacher where this support is needed or wanted. In the field of clinical judgment, research has focused on the question how statistics can help the expert to reach optimal decisions. They found that formal statistical rules could be used by the expert (e.g., the clinician or in our case the teacher) in every step of the decision process (i.e., selection of data, (re)collection of this data and (re)weighting of the data to come to a decision; Dawes, 2002). When information is available about a pupil's current track placement and a variety of - by the teacher deemed - relevant indicators (e.g., scores on previous tests, observations in the classroom et cetera), statistics can for instance assist in selecting those indicators that are generally predictive of actual track placement. Statistics can also help the teacher in weighting these indicators based on their predictive power. Even if the expert (i.e., teacher) prefers to stay in control of the selection and relative weighting of the data, Dana and Thomas (2006) show that statistical rules can help to reach a more accurate and more consistent ultimate decision. To be clear: in this proposal, the teacher is not *replaced* by mere statistics, as statistics are generally not concerned with unique individuals but with averages and aggregates (Hart, 2003). Rather, the teacher retains his idiographic focus, while the statistical toolbox is used to ensure procedural reliability and a consistent adherence to one's own judgment.

policy.

Another way to aid the teacher recommendation is by developing protocols. Currently, many local protocols are being created and updated to streamline the way teacher recommendations are formulated, communicated and evaluated (see Smeets, Van Kuijk and Driessen, 2014). Despite the clear potential of these protocols, there are some caveats. Protocols differ in the (non-)cognitive factors they deem relevant and whether or not formal decision rules are part of this protocol (Dutch Inspectorate of Education, 2014). Moreover, when formal decision rules are established, they vary over protocols. In the ‘plaatsingswijzer’⁴⁵, for instance, a selection of scores from the monitoring system in grade 4 - grade 6 is translated into a track recommendation⁴⁶. Subsequently, it is checked whether each of these tests points in the same (track) direction. In the ‘plaatsingswijzer’, every test from the monitoring system thus plays an equal role in the determination of the track placement and consistency over tests is deemed important. To contrast: in Almere’s protocol (see Smeets, Van Kuijk and Driessen, 2014), tests from the monitoring system are weighted by a formula (updated on the basis of information collected in Almere). Consistency of the test results plays a less profound role in their protocol. The danger of the protocol diversity is that similar pupils may receive different track recommendations in different regions of the Netherlands. What is currently lacking is a comparison of protocols to understand how exactly these protocols impact track recommendation practices. Furthermore, although there are national standards (see Amsing, Bosch & Rouweler, 2009) and some popular general guidelines (see, for instance, De Wijs & Hollenberg, 2014), it is unclear whether and how these play a role in the transition from primary to secondary education in different regions of the Netherlands. In my own comparison of protocols, I often found it hard to find any rationale for the selection of factors and establishment of decision rules. It was also difficult to find any information on how the decision rules are updated over time. Transparency is needed to ensure that protocols are up-to-date and based on relevant and recent data.

Why not *combine* teacher and test?

One obvious yet often overlooked option is to statistically *combine* teacher judgment and End-of-Primary-school test result. Teacher judgment and test result could be weighted, for instance, in such a way that the resulting combination of the two is ‘better’. ‘Better’ could for instance mean ‘more predictive of the actual track placement of the student in the third year of secondary education’. But we could also optimize the weights taking into account the utility of the placement decision. There are, for instance, indications that a track recommendation that is too optimistic is less detrimental than a track recommendation that is too conservative. Under-recommended pupils are largely unable to switch up, possibly due to self-fulfilling prophecy effects (see Jussim & Harber, 2005). Over-recommended pupils, on the other hand, are generally capable of maintaining their optimistic track placement (De Boer, Bosker & Van Der Werf, 2007; Driessen, 2011; Geven et al., 2018; Timmermans et al., 2013). Taking into account that classmates are generally not ‘hurt’ by the presence of somewhat lower ability pupils (except for the pupils with the highest abilities; see Feron et al., 2015; Diris, 2012; Van Elk et al., 2009), we could weight teacher judgment and test result such that somewhat higher track indications are favored.

It is also possible to weight the teacher judgment and test result based on teacher/measurement uncertainty. Specifically, we could ask the teacher to not only express his or her track recommendation, but also how (un)certain he or she is about this recommendation. If we use this information to construct a Bayesian prior (as we did in chapter 5), teacher judgment and test result can be combined in a so-called Bayesian posterior (for a practical example, see Wagenmakers, Morey & Lee, 2016). This Bayesian posterior weights the teacher judgment (captured in the prior) and the test result based on their (un)certainty and expresses the (un)certainty in this combined advice. One advantage of this Bayesian weighting approach is that it provides secondary schools with a track recommendation *and* an indication of whether this pupil might benefit from a track level lower or higher. Since background characteristics of students tend to play a larger role in the teacher recommendation when the teacher is insecure (Inspectorate of Education, 2018), uncertainty in the teacher recommendation is valuable information to have. Other advantages are that multiple (teacher) judgments could be combined easily and that the compatibility of teacher judgment and test result(s) can be statistically checked by looking for prior-data conflict (see chapter 7). For the teacher, the advantage is that when (s)he expresses uncertainty regularly, this helps in developing calibrated judgments, meaning that the teacher shows more certainty with correct judgments than with incorrect ones (Gabriele & Park, 2016).

Feedbackcycle

All the suggestions above (i.e., combining multiple tests, systematically breaking down the judgment process in smaller steps, using statistical rules and protocols, weighting teacher judgment and test result) only work if data on pupil's current track placement is collected and analyzed on a regular basis. The 'optimal' weights for teacher judgment and End-of-Primary-school test result, for instance, probably change over time. And to fine-tune their judgment process, teachers need data on how their teacher recommendations differ from pupils actual track placement in secondary education. Currently, secondary schools are expected to share data on how pupils progress with the former primary school. Secondary schools differ, however, in how much information they share and not all schools for primary education use this data to evaluate the track allocation process (Smeets, Van Kuijk & Driessen, 2014). What is needed is a systematic approach on a national-, school- and teacher level. A valuable source of information on a national level is the annual 'Staat van het Onderwijs'-report created by the Dutch Inspectorate of Education (see <https://www.onderwijsinspectie.nl/onderwerpen/staat-van-het-onderwijs>). Together with the POraad, we are also exploring the idea of repeating our study (see chapter 6) on a yearly basis. Promising sources of information on the school- and teacher level are the elaborate reports sent by the National Cohortstudy Educational Careers ('Nationaal Cohortonderzoek Onderwijs', NCO; initiated by The Netherlands Initiative for Education Research NRO) and the available information on vensters.nl. Both sources are especially helpful in detecting general bias: whether teacher recommendations are systematically too high or too low compared to 'comparable schools'. To investigate *specific* bias (i.e., whether teacher recommendations are systematically too high or too low for specific subgroups; Timmermans et al., 2015), more information is needed. In the [vensters](http://vensters.nl) report⁴⁷, for instance, a detailed overview is presented with the percentage

students who switched up and down, conditional on their teacher recommendation and End-of-Primary-school test result. To investigate specific bias, it would be informative to know how these percentages differ for students with different socio-economic background or ethnicities, for instance. On top of a systematic approach of data collection and sharing, teachers and schoolleaders also need to be supported in how to optimally use the information to correctly adjust the track allocation process. The work of Schildkamp (see, for instance, Schildkamp et al., 2018; Vanlommel & Schildkamp, 2018) shows how teachers and schoolleaders currently interpret data and how their data usage can be successfully improved.

Smart Testing

Note that just as (test) data can be used to improve teacher judgments, teacher judgments in their turn may be useful to improve (test) data collection. Teacher judgments could, for instance, play a valuable role in tests that are constructed based on multistage testing (see Hendrickson, 2007). In multistage testing, preconstructed sets of items are administered adaptively, starting with a routing test containing items with a range of difficulty values (i.e., relatively easy and difficult items). Based on their performance on this routing test, pupils are subsequently administered a second set of items with a more narrowly focused difficulty, matching their initial estimated ability (this process is repeated when there are more than two stages). Thus, multistage tests are aimed to optimize the match between the difficulty of the presented items and the ability of students based on the performance on earlier administered parts of the same test. This has the advantage that less items are needed in order to estimate the ability of students and especially in the last part of the test, less guessing and slipping is likely to be observed due to the tailoring of items to persons. In sum, well-designed multistage tests manage to measure a broader range of abilities in a efficient and tailored way compared to linear tests. To shorten multistage tests and to minimize pupil's fatigue and frustration, teacher judgment could possibly replace the initial routing test. As shown by Kim, Moses and Yoo (2015; also see Kim & Moses, 2014) misrouting generally has a minimal impact on the final pupil's proficiency estimate. Therefore, even if the teacher judgment is wrong occasionally, they still have the potential to shorten the multistage test without biasing the final test result.

Teacher judgments could also play an important role in testing as part of the schools' monitoring system. Currently, every pupil completes the same amount of tests within this system. The teachers' knowledge gain as a result of these tests is, however, not the same for every pupil. Some pupils, for instance, show a consistent pattern from test to test and the teacher has a clear picture of the abilities and challenges of these pupils. Completing another test does therefore not provide the teacher with much (new) information. Other pupils fluctuate much more in their performances from test to test, leaving the teacher with more insecurity about their learning achievements and possibilities. For these pupils, every test provides new and useful information, helping the teacher to adapt instruction accordingly. If it was possible to incorporate teacher knowledge and teacher insecurity in the school's monitoring system, we could test pupils based on which information is lacking. Providing a minimal set of tests for every pupil, the teacher could select specific extra tests for pupils (s)he is insecure about. By incorporating teacher judgments in the monitoring system in this way, testing 'for the sake of testing' is replaced by 'testing as

an instrument to gain information about pupils that is currently lacking'. Ideally, for teachers this helps to experience the value of test results as tests become more integrated within the daily educational practice.

How to proceed?

As shown above, there are many options in between a completely unaided, holistic teacher judgment and a single, End-of-Primary-school test result in the transition from primary to secondary education. It is useful to explore how we can make optimal use of the advantages of teacher judgments and test results, without necessarily opting for just one of the two. Indeed, investigating how teacher judgments and test results can strengthen one another is a useful endeavor in an era characterised by an intensified focus on standardized testing together with an increasing testing resistance in the educational field. The suggestions above (i.e., combining multiple tests, systematically breaking down the judgment process in smaller steps, using statistical rules and protocols, weighting teacher judgment and test result, et cetera) show that I plea for more standardization and comparability over schools in their transition procedures, substantiated by systematic data collection. Such a consistency is necessary to ensure a 'fair' transition process, in which pupils' track recommendations are minimally influenced by pupil's demographic characteristics, school characteristics and school region. What I do *not* imply, however, is that such a standardization should be enforced upon schools in a top-down manner. Such a top-down enforcement would run counter to the notion that the expertise, ideas and concerns of teachers and other (educational) experts should be taken seriously. The transition from primary to secondary education is so complex and so delicate that it should not be based on the beliefs and expertise of a *single* group (e.g., psychometricians, educational scientists, educational policymakers, schoolleaders, teachers). The challenge is to bundle forces bottom-up, to support fruitful initiatives (such as with the different protocols that were mentioned) and to investigate how these initiatives could be implemented on a larger, national scale, once their merits and potential pitfalls have been investigated. As I have mentioned before, collecting data on a regular basis helps to substantiate ideas and initiatives. When this data is evaluated and discussed by all researchers, policy makers and educational practitioners involved in the transition from primary to secondary education, we could ideally end up in a feedbackcycle in which all these 'stakeholders' learn and improve the transition from primary to secondary education *together*. To some critical readers, the idea of bundling expertise and sharing ideas in this way might sound unrealistic. There are, however, examples that show the potential of such a collective learning approach. One example is the Educational Agenda Limburg (EAL); a large-scale cooperation between a number of universities, hbo (higher professional education) - and mbo (senior secondary vocational education) -schools and institutions for primary and secondary education, focused on improving transitions in education such as the transition from primary to secondary education (see <https://www.educatieveagendalimburg.nl>). The 'research gang' (Dutch: 'Onderzoeksbende', see <https://onderzoeksbende.nl/>), consisting of representatives of the Dutch boards for primary and secondary education, mbo's, hbo's and universities, is another example of an initiative aimed at connecting teachers, schoolleaders, policy makers and researchers. Recently (April, 2019) they published a report on how this connection could be established and improved, based on many

interviews they had with educational experts and other stakeholders. Another example: NRO (the Netherlands Initiative for Education Research, founded in 2012) facilitates the connection between educational practice and research by funding research that is conducted in consortia consisting of (primary) schools and universities and by the founding of so-called ‘workplaces for educational research’, in which researchers and educational practitioners work together to develop their school(s) based on research. Finally, since 2008 universities started to offer an academic teacher education program (ALPO). On top of didactic and pedagogical skills, these academic teachers are equipped with an understanding of (educational) research and the ability to conduct such research themselves. These academic teachers can therefore play a vital role in bridging the gap between educational practice and (educational) research (see PO-raad, 2018). Although there are still many challenges to tackle⁴⁸, I believe that if we are able to establish a sustainable cooperation between educational practice and educational research, the transition from primary to secondary education can be optimized for every pupil, making use of the advantages of test(s) and teacher expertise.

Footnotes

¹The Organisation for Economic Co-operation and Development (OECD) has 36 member countries from North and South America to Europe and the Asia-Pacific. see www.oecd.org.

²Methodological concerns are dealt with and a more recent SAT-dataset is used. See Santelices and Wilson (2010) for a detailed overview of improvements.

³For instance when rank ordering schools based on test results that are not validated for this purpose.

⁴The College for Tests and Exams, an independent administrative body of the Dutch government, offers the End of Primary School Test which is developed by the CITO. In the 2014-2015 school year Primary schools could also opt for the alternatives ROUTE 8 (developed by A-VISION) and the IEP end test (Bureau ICE). Since then, the Dia end test and the AMN end test have also become available.

⁵Note that in the context of educational testing, especially within the Item Response Theory literature, the term ‘differential item functioning’ is often used to refer to measurement non-invariance in specific items (see Millsap, 2011; Mellenbergh, 2011). In chapter 3, I stick to the more general term ‘measurement invariance’.

⁶The term ‘teacher judgment’ is sometimes misleading, as it might give the impression that teachers reach these judgments in isolation. In reality, this is rarely the case, as teachers operate in teams and often a single class is taught by multiple teachers part-time.

⁷Most research on the accuracy of teacher judgments simply assesses the correlation between the teacher judgment and student’s actual performance on a related test. See Südkamp, Kaiser & Möller, 2012.

⁸Note that the expert elicitation can easily be extended to multiple colleague teachers. Either the expert elicitation is completed for all teachers at once or the expert elicitation is performed separately combining the teachers’ elicitation results statistically.

⁹Or, more generally, a test participant.

¹⁰More specifically, all intelligence tests based on the Classical Test Theory (as explained in the section ‘How to interpret WISC-III^{NL} confidence intervals?’) construct their confidence intervals in a similar fashion.

¹¹Note that the standard deviations of the raw subtest score distributions as displayed in Figure 1.1 are not represented in one of the WISC-III^{NL} tables. These values can be easily computed as follows: (Standard error of measurement)/ $\sqrt{(1 - \alpha)}$, where α represents the reliability coefficient for the specific subtest and age group. For example, with age group 6.5 and subtest ‘similarities’ this is: 1.42 (table 3.7), divided by $\sqrt{1 - 0.78}$ (table 3.5) = 3.03.

¹²Figure 1.5) shows how the WISC-III^{NL} first executes this conversion from total score to IQ score for each age group separately. Unsurprisingly, effects for age group did not emerge (the total scores are already based on standard scores).

¹³Figure 1.5 shows how the WISC-III first executes this conversion from total score to IQ score for each age group separately. Unsurprisingly, effects for age group did not emerge (the total scores are already based on standard scores).

¹⁴The reliability coefficient is estimated directly based on the norm population. The norm population therefore has a direct influence on the width of the confidence interval (options 1 and 2). If the norm population actually is not representative for children taking the WISC-III^{NL} intelligence test, this has far-reaching consequences for the interpretation of the confidence interval.

¹⁵For Julia (IQ = 97) the option 1 and 2 confidence intervals are exactly the same after rounding: 90 to 104.

¹⁶Note that over time, researchers have extended and revised the CTT-model for different purposes (see Hambleton & Jones, 1993). Here, we only discuss the most basic and common CTT-model.

¹⁷Note that there are also CTT-models in which different distributions are considered such as the binomial and beta-binomial (e.g., Van der Linden, 1979).

¹⁸The rationale is as follows. Since parallel measurements have experimentally independent errors, the joint distribution function over all (parallel) parts factor into the score distributions for each of the parallel parts separately (Lord & Novick, 1968, definition 2.10.1). Sampling scores one by one from their K respective distributions is thus stochastically the same as sampling from the joint distribution at once. For normal distributions such as this joint distribution, it is known by Cochran's theorem that the variance $\sigma^2(E_{*i})$ follows a scaled chi-squared distribution.

¹⁹Grouping by total score instead of 'true' score introduces some bias, depending on the reliability of the test. See Woodruff, 1990

²⁰Some bias will be introduced in the conditional SEM resulting from basing the groups j on observed rather than true test scores (see Woodruff, 1990).

²¹We simulated the "true" scores based on a Gaussian distribution. This choice is rather arbitrary; Classical Test Theory does not assume τ_i to be normally distributed in the norm population. Note that the resulting estimates of $\sigma^2(E_{*i})$ are not influenced by this choice.

²²The plots in appendix 2.A (see <https://osf.io/qrg4e/>) are all based on a 12-item test, with an overall reliability of 0.8 and $K = 2$ and the relationship depicted in Figure 2.5a. Plots for any other combination of number of items, K , overall reliability and relationship between "true" score and error variance can be requested from the first author

²³ $\epsilon = fc$ where c is 0.05 by default and f is a multiplicity factor that takes into account the number of parameters in the model. **Bconvergence** = .01; replaces c by 0.01, hence yielding a more stringent convergence criterion.

²⁴This chapter has introduced the concept of approximate measurement invariance and illustrated the use of its most basic variant. More complex variants, such as multilevel/hierarchical models and other types of Bayesian priors on differences, have fallen out of the scope of this chapter. For applications of multilevel/hierarchical models to measurement invariance, see Cheung and Au 2005; Davidov et al. 2012, in press; Jak et al. 2014a,b; Jak et al. 2013; Meuleman 2016. Additionally, it is not yet clear how exactly to compare models with different priors in the context of approximate measurement invariance. Some preliminary results show that the PPP and DIC are not so well suited and alternatives have been proposed (Hojtink & Van De Schoot, 2017). Furthermore, we have avoided issues external to measurement equivalence, such as overall model fit and concept equivalence (see, e.g., Meitinger 2014).

²⁵This study indicated that teachers developed higher expectations of pupils who were randomly labeled 'late bloomers'. As a consequence of these higher expectations, these 'late bloomers' enjoyed more positive attention, leading them to achieve higher results in return. The results of this study showed how influential teacher expectations could be for the school career of the pupil. Replication studies, however, largely failed to find the same effect.

²⁶The terms 'teacher expectations' and 'teacher judgments' are often used interchangeably in the literature. Generally, 'teacher judgments' refer to the beliefs a teacher has about current pupil comprehension and ability. 'Teacher expectations' are also focused on future academic achievements (Good & Brophy, 1997). Teachers' track recommendations can be based on their beliefs of pupil's current and future achievements.

²⁷We focus on these studies as the Netherlands, Luxembourg and Flanders have a comparable tracked school system. There are, however, some differences between the Netherlands, Luxembourg and Flanders that the reader should take into account. In the Netherlands, parents - for instance - do not have an

active role in the track advice. They can attempt to convince the teacher to adjust the advice, but the teacher does not have to take the concerns of the parents into account. In contrast, in Luxembourg the track advice results from an extensive exchange between teacher and parents. When teacher and parents disagree, an independent orientation committee determines the track placement, carefully weighting the concerns of the teacher and the parents. Compared to the Netherlands and Luxembourg, in Flanders the transition from primary to secondary education is more loosely organized; there are no formal regulations, no standardized tests and the advice of the teacher is not binding.

²⁸In the Netherlands, sixth grade pupils with a migration background are mostly the descendants of Turkish and Moroccan guestworkers arriving in the 1960s. Other immigrants come from former colonies (e.g., the Dutch Antilles and Suriname), stem from guest workers from other Western countries or are refugees/asylum seekers from the Middle east (Scheerens & Van Der Werf, 2018; Driessen, Slegers & Smit, 2008).

²⁹Information about accessing these files can be obtained by consulting the CBS website: <https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research>

³⁰most classes with fewer than 10 pupils only shared the teacher recommendation for one pupil. With only one or a few pupils per class, it is difficult to distinguish between pupil-level and class-level effects. Additionally, aggregate variables at the class level are unstable when based on few pupils.

³¹in the CITOtab file, a distinction is made between all subtracks of vmbo: vmbo-basisberoeps, vmbo-basis/kaderberoeps, vmbo-kaderberoeps, vmbo-kaderberoeps/vmbo-gemengd/theoretisch, vmbo-gemengd/theoretisch.

³²Note that the coefficient for the miscellaneous category is close to zero. This does not necessarily mean, however, that having a migration background other than Moroccan or Turkish does not influence the teacher recommendation. As this heterogeneous category contains immigrants from all over the world (except Morocco and Turkey) who migrated for a variety of reasons, it might well be that positive and negative effects within this miscellaneous group cancel each other out.

³³i.e., parent's annual average income > 50,000, WOZ-value equal to the sample's average, gender = boy, ethnicity = native Dutch, both EPST-scores equal to the sample's average, in a class with no parents with an annual average income below 50,000, with an average WOZ-value equal to the sample's average, with no non-native Dutch pupils and with an average class score on the EPST language and mathematics parts equal to the sample's average

³⁴Note that in many of these density plots, a small peak is visible around scores close to 550. This phenomenon is due to the way the EPST-score is constructed. The number of correct items on the test are linearly transformed in order to scale the observed scores to the reporting scale of 501 - 550 and to equate scores to results obtained in previous administrations. EPST-scores outside the reporting scale are truncated to the minimum and maximum possible reporting scores. (<https://www.cvte.nl/documenten/publicaties/2015/05/12/verantwoording-centrale-eindtoets-po>). Therefore the number of correct items resulting in an EPST-score of 501 and 550 are larger compared to other EPST-scores.

³⁵'Expert' here simply refers to the "person whose knowledge is to be elicited" (O'Hagan et al., 2006, p. 9).

³⁶Specifically, we created a list of random student codes. Every teacher got as many of these codes as they had pupils in their class, starting with the first student code of the list, then the second, et cetera. The student codes of different teachers thus overlapped, avoiding a link between certain student codes and specific teachers.

³⁷The College for Tests and Exams, an independent administrative body of the Dutch government, offers the End of Primary School Test which is developed by the CITO. In the 2014-2015 school year the schools could also opt for the alternatives ROUTE 8 (developed by A-VISION) and the IEP end test (Bureau ICE). Since then, the Dia end test and the AMN end test have also become available. The data used for this article are exclusively based on pupils who have completed the Cito End of Primary School Test.

³⁸i.e. the situation that pupils are allocated an education level that does not do justice to their capacities

³⁹These data were provided under strict conditions aimed at protection of personal details of pupils.

⁴⁰See <https://www.rijksoverheid.nl/onderwerpen/toelating-middelbare-school/vraag-en-antwoord/adviezen-in-groep-8>

⁴¹Note that sixth grade is the last grade of primary education in the Netherlands

⁴²Summative tests are meant to provide information on learning achievement in a certain period of time, whereas formative tests are meant to highlight how learning can be improved (Dolin et al., 2018)

⁴³an often used monitoring system in the Netherlands

⁴⁴Note that in He's 2009 paper, the tests were administered at approximately the same time. The formulas in He's paper need to be updated for the proposal to combine EPST-results with tests from the monitoring system, to take the time intervals into account

⁴⁵Originally developed in the Northern regions of the Netherlands, Friesland and Groningen. See plaatsingswijzer.nl

⁴⁶Every track is linked to an interval of possible scores called 'bandbreedtes'

⁴⁷For an example report, see <https://static1.squarespace.com/static/57a9d4ba725e256e5549a4f4/t/58eccb83197aea316e9ccfdc/1491913604265/Demorapport+Eindtoets.pdf>

⁴⁸The Dutch report 'Lerend onderwijs voor een lerend Nederland', written by the 'Onderzoeksbende' (April, 2019) lists a number of challenges we are currently facing. One example is that the culture of universities needs to change from one in which researchers are primarily rewarded for their publications in high impact international journals to one in which researchers are (also) rewarded for their contribution to the Dutch society.

Nederlandse samenvatting

Waarschijnlijk de bekendste en meest controversiële toets in het Nederlandse onderwijssysteem is de eindtoets in groep 8, waarvan Cito de belangrijkste aanbieder is. Deze toets is ooit (in zijn oorspronkelijke vorm) ontwikkeld door toetspionier Professor Adriaan de Groot in de jaren 60. De Groot was voor deze ontwikkeling geïnspireerd geraakt door met name de Amerikaanse ontwikkeling van Testtheorie, waar de nadruk lag op objectiviteit en de toepassing van wetenschappelijke inzichten (zie Gallagher, 2003). Volgens De Groot was het mogelijk met de eindtoets een objectief oordeel te vellen over leerlingen. Dit oordeel zou in scherp contrast staan met subjectieve oordelen van de leerkracht, die leerlingen mogelijk verkeerd zou beoordelen op basis van bijvoorbeeld sociaal-economische status en etniciteit (Mertens, 2016). De leerkracht die al die tijd als de onbetwiste en gerespecteerde expert werd gezien, moest steeds meer aan gezag inboeten.

Net als andere invloedrijke toetsen is de eindtoets veelvuldig bekritiseerd, met name vanuit het onderwijsveld zelf. Zo zou de eindtoets volgens sommigen (zie Karels, 2019) het gat tussen kinderen van hoger- en lager opgeleide ouders en kinderen met verschillende etnische achtergronden juist vergroten in plaats van verkleinen. Vermogende ouders kunnen immers toetstraining en oefenmaterialen bekostigen. Bovendien zou de eindtoets slechts een ‘momentopname’ laten zien, terwijl de leerkracht zicht heeft op de gehele basisschoolontwikkeling. En waar een toets zaken als discipline en motivatie niet kan meten, is de leerkracht hier wel van op de hoogte.

Het is deze kritiek die in schooljaar 2014/2015 tot een belangrijke wetswijziging leidt: niet langer zou de eindtoets in de overgang van primair naar voortgezet onderwijs een leidende rol innemen. Het vertrouwen in het oordeelsvermogen van de docent werd hersteld en de toets werd gedegradeerd tot een optionele ‘second opinion’. In het verhitte debat wat volgt op de wetswijziging, komen een aantal vragen naar voren: ‘is één van de twee (de toets of de docent) nu echt geschikter om de plaatsing van leerlingen te bepalen in het voortgezet onderwijs?’ en ‘in welke mate zijn toetsresultaten en docentoordelen nu beïnvloedt door factoren als sociaal-economische status en etniciteit, en hoe kunnen deze invloeden beperkt worden?’ Met dit proefschrift heb ik de kans gehad deze en gerelateerde vragen te onderzoeken in de context van de eindtoets en daarbuiten. Mijn proefschrift is verdeeld in drie delen. In deel 1 en 2 onderzoek ik enkele ‘gevaren’ en voordelen van respectievelijk toetsresultaten en docentoordelen. In het derde deel focus ik op de vergelijking en potentiële combinatie van toetsresultaten en docentoordelen.

Deel I: De toets

Om (meer)waarde te hebben, dient een toets allereerst goed gebruikt te worden. Dit klinkt misschien triviaal, maar in de praktijk blijkt dat een dergelijke ‘toetsgeletterdheid’ (Schenke & Meijer, 2018; Schildkamp & Poortman, 2015) onder toetsgebruikers vaak nog ontbreekt. Een van de redenen waarom het weleens mis gaat in de interpretatie en het gebruik van toetsresultaten is het niet (in voldoende mate) meewegen van meetonzekerheid, bijvoorbeeld uitgedrukt in zogenoemde ‘betrouwbaarheidsintervallen’. In de context van de IQ-test is dit een bekend probleem. Voor het **eerste hoofdstuk** van mijn dissertatie heb ik nauw samengewerkt met psychologen die IQ-testen afnemen in hun dagelijkse praktijk. Het doel van dit hoofdstuk was het op een begrijpelijke, niet-technische manier uitleggen van de verschillende betrouwbaarheidsintervallen in IQ-testen, hoe deze geïnterpreteerd dienen te worden en welke - vaak onbekende of genegeerde - assumpties hieraan ten grondslag liggen.

Één van de assumpties die door de psychologen als onrealistisch werd beschouwd, was het idee dat alle kinderen (of andere geëxamineerden) met evenveel meetonzekerheid zouden zijn gemeten (Sijtsma, 2009; Molenaar, 2004). Deze assumptie wordt vaak gemaakt in de zogenaamde Klassieke Testtheorie; de theorie waarop de meeste Nederlandse IQ-testen zijn gebaseerd. In het **tweede hoofdstuk** bespreek ik twee alternatieven waarin deze assumptie wordt versoepeld. Door middel van een simulatiestudie onderzoek ik welk van deze alternatieven geschikt is in welke testsituatie.

Naast dat toetsen goed geïnterpreteerd dienen te worden, moet er ook zorg voor gedragen worden dat deze bepaalde groepen (bijvoorbeeld met een bepaalde etnische achtergrond of sociaal-economische klasse) niet bevoor- of benadeelt. Dit wordt meetinvariantie genoemd (zie bijvoorbeeld Millsap, 2011). In het **derde hoofdstuk** richt ik mij op een nieuwe ontwikkeling binnen het meetinvariantie veld, die ervoor zorgt dat het checken van meetinvariantie gemakkelijker gaat en minder strikt is (zie Muthén & Asparouhov, 2013; Van De Schoot et al., 2013).

Deel II: De leerkracht

Net als bij toetsen zou ook het oordeel van de leerkracht ‘invariant’ moeten zijn ten opzichte van de sociaal-economische en etnische achtergrond van leerlingen. In mijn **vierde proefschrift hoofdstuk** onderzoek ik daarom - specifiek voor de overgang van primair naar voortgezet onderwijs - in hoeverre leerling- en schoolkenmerken een rol spelen in docentoordelen.

Het bepalen van de kwaliteit van docentoordelen is lastig, met name omdat deze oordelen vaak impliciet zijn. In **hoofdstuk vijf** introduceer ik een zogenaamde ‘elicitiemethode’ (O’Hagan, 2006) die het mogelijk maakt impliciete docentoordelen expliciet te maken. Het voordeel is dat docentoordelen vervolgens formeel getest en geëvalueerd kunnen worden. Bovendien maakt deze elicitiemethode het mogelijk docentoordelen en toetsresultaten op statistische wijze samen te voegen (zie hiervoor hoofdstuk 7 en de discussiesectie van deze dissertatie).

Deel III: Leerkracht versus toets

De hamvraag in de context van de overgang van primair naar voortgezet onderwijs is of de eindtoets *of* het docentoordeel geschikter is. In **hoofdstuk zes** probeer ik antwoord te geven op deze vraag, gebruikmakend van een grote dataset van het Centraal Bureau voor de Statistiek (CBS).

Het laatste hoofdstuk, **hoofdstuk zeven**, resulteerde vanuit het idee dat de discussie omtrent de eindtoets mogelijk te zwart-wit gevoerd wordt. In plaats van ons te focussen op de vraag wie er ‘gelijk’ heeft, de docent *of* de eindtoets, zouden we onszelf beter kunnen afvragen hoe docentoordeel en toetsresultaat zo optimaal mogelijk gecombineerd kunnen worden. Wanneer docentoordeelen zijn uitgedrukt zoals uitgelegd in hoofdstuk 5, is het mogelijk om deze te combineren met toetsresultaten gebruikmakend van zogenaamde ‘Bayesiaanse statistiek’ (voor een niet-technische introductie zie Van De Schoot et al., 2014). Het resultaat is een compromis tussen de docent en de toets. Zo’n compromis werkt zolang docentoordeel en toetsresultaat niet ‘teveel’ van elkaar verschillen. Hoofdstuk 7 legt uit hoe getest kan worden of de Bayesiaanse compromis tussen docent en toets betekenisvol is. Specifiek richt dit hoofdstuk zich op twee verschillende checks en hoe robuust deze checks zijn wanneer een van de ingrediënten van deze checks wordt gewijzigd.

Alle extra materialen en appendices, behorende bij de hierboven genoemde hoofdstukken, kunnen geraadpleegd worden op mijn Open Science Framework (OSF) pagina: <https://osf.io/qrg4e/>.

Veel leesplezier toegewenst!

Curriculum Vitae

Kimberley Lek was born to Henriëtte and Tino Lek on June, 25, 1990. Together with her younger brother, she was raised in Aarlanderveen, a small farmer village without any supermarkets but with a curtain shop and four windmills. Determined to become a primary school teacher, she went to the University of Applied sciences ‘Domstad’ (referring to the area in Utrecht, not to the intellectual capacities of their students) in Utrecht from 2007-2011 and finished her teachers’ degree ‘cum laude’. After that, she went to Utrecht University to first complete a pre-master and after that a research master in educational sciences. As one research master is not that hard (it is), she decided to combine this research master with the research master Methodology and Statistics, taught at the same university. In 2015, she completed both research masters cum laude and was the Valedictorian for both research masters. In the same year, she received the Peter G. Swanborn price for the best researchmaster thesis (see <https://www.uu.nl/organisatie/faculteit-sociale-wetenschappen/peter-g-swanbornprijs>) and obtained a ‘NWO talent grant’ (nr. 406-15-062) for a four-year PhD trajectory. 14 April 2016, Kimberley married Guido Bosch.

2 September 2019, Kimberley started working at testing institute Cito in Arnhem.

Publications

- Van De Schoot, R, Schmidt, P, De Beuckelaer, A, **Lek, K**, & Zondervan-Zwijnenburg, M (2015). Editorial: Measurement invariance. *Frontiers in Psychology*, 6: article 1064.
- **Lek, K**, Van De Schoot-Hubeek, W, Kroesbergen, E, & Van De Schoot, R (2017). Hoe zat het ook alweer? Het betrouwbaarheidsinterval in intelligentietests. *De Psycholoog*: 10.
- Zondervan-Zwijnenburg, M, Van De Schoot-Hubeek, W, **Lek, K**, [...], & Van De Schoot, R (2017). Application and evaluation of an expert judgment elicitation procedure for correlations. *Frontiers in Psychology*, 8: article 90.
- **Lek, K**, & Van De Schoot, R (2018). A comparison of the single, conditional and person-specific standard error of measurement: What do they measure and when to use them? *Frontiers in Applied Mathematics and Statistics* 4: article 40.
- **Lek, K**, & Van De Schoot, R (2018). Development and evaluation of a digital expert elicitation method aimed at fostering elementary school teachers’ diagnostic competence. *Frontiers in Education* 3: article 82.

- **Lek, K,** & Van De Schoot, R (2019). How the choice of distance measure influences the detection of prior-data conflict. *Entropy* 21(5): 446.
- Duinhof, E, **Lek, K,** De Looze, M E, & Stevens, G W J M (2019). Revising the self-report strengths and difficulties questionnaire for cross-country comparisons of adolescent mental health problems: the SDQ-R. *Epidemiology and Psychiatric Sciences*: 1.
- **Lek, K,** Oberski, D, Davidov, E, Cieciuch, J, Seddig, D, & Schmidt, P (2019). Approximate measurement invariance. In T P Johnson, B-E Pennell, I A L Stoop & B Dorer (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)* (pp. 911-930). John Wiley & Sons.
- **Lek, K,** & Van De Schoot, R (2019). Bayesian approximate measurement invariance. In F J R Van De Vijver (Ed.), *Invariance analyses in large-scale studies*, OECD Education Working Paper, No. 201.
- **Lek, K,** & Van De Schoot, R (2019). Wie weet het beter, de docent of de centrale eindtoets? *De Psycholoog*, 54(4): 10-21.
- **Lek, K,** & Arts, I (in press). How to improve the estimation of a specific examinee's ($n = 1$) math ability when test data is limited. In R Van De Schoot & M Miočević (Eds.), *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners*.

Acknowledgements

First and foremost, I would like to thank my husband Guido Bosch for all his love, support and patience. In my (many) moments of doubt, stress and insecurity, you managed to put things into perspective for me. I truly could not have done any of this without you. Mum and dad, thank you for believing in me. From a young age, you learned me the value of hard work. This work attitude has helped me through my pre-master, my researchmasters and my PhD; without it I would not be where I stand today. Margreet and Peter, thank you for the many conversations and pep-talks we had during the past ten years. You have always shown great interest in my studies and I count myself very lucky to have such fantastic parents-in-law.

Timothy, Jochem, Wytke, Belinda, Thomas, Daniël, Femke, thank you for all the fun and laughter in the past years. You are the best brothers(-in-law), sisters-in-law, nephew and niece I could have hoped for. Grandma Jopie and (bonus) grandma Anna, thank you for always listening to me and for your efforts in understanding what a PhD entails exactly. Grandma Jopie, you were my first and most important reader. Lianne, Marga, Bert, Leonieke and Marisa, thank you for being such great friends. I could not have finished this PhD without our tea talks and walks. Ans, Henk, José, Piet, Ramona and all other family, in-laws and friends, you are great and I am sorry for the birthdays and other celebrations I've missed over the past couple of years.

Rens, thank you for all your support over the last four years. Whenever I thought that a project was doomed to fail, you were optimistic and showed me opportunities. Thank you for the training days you have organized to guard me and the rest of your team from the mental health 'dangers' that are often associated with doing a PhD. Ruud, thank you for being such a motivating trainer at these training days. As a person, I feel that I am stronger and more confident as a result. Mariëlle, Sanne, Duco, Sonja, Marthe, Evelien, Ingrid, and all other former and current PhD-team members, I had a great time with you. I loved our working-trip to South-Africa and the many early morning and afternoon safaris we went on. Mariëlle, I was lucky to have you as my roommate for the past four years. You are a truly great and funny person, fantastic mother and talented researcher. If I turn fat in the coming years, it is because of all the cake and cookies I have to eat now that you are not there to steal them from me. Wenneke, it was great working with you. Due to your efforts, my PhD research became more useful for teachers and psychologists in practice. I will never forget the trip to your former workplace, although I am glad that Mariëlle and I 'survived' the flying chairs and threats. Remco, thank you very much for your help and feedback in the final stages of my PhD. You are an inspiring person to collaborate with, and I very much enjoyed our projects during my PhD and the current projects we are working on now that I work at Cito.

As it is impossible to thank everyone personally, I would like to say a general 'thank

you' to all other persons who have played a role during my PhD. For instance to all former UU-colleagues who made the university an enjoyable place to work, to all former colleagues of the UU summerschool and all other persons I had the opportunity to meet during my PhD. I could not have finished my PhD without you.

Kimberley Lek

References

- Agirdag, O, P Van Avermaet, and M Van Houtte. 2013. "School Segregation and Math Achievement: A Mixed-Method Study on the Role of Self-Fulfilling Prophecies." *Teachers College Record* 115 (3): 1–50.
- Albers, C J, R R Meijer, and J N Tendeiro. 2016. "Derivation and Applicability of Asymptotic Results for Multiple Subtests Person-Fit Statistics." *Applied Psychological Measurement* 40 (4): 274–88.
- Ali, S M, and S D Silvey. 1966. "A General Class of Coefficients of Divergence of One Distribution from Another." *Journal of the Royal Statistical Society Series B* 28 (1): 131–42.
- Angoff, W H. 1993. "Perspectives on Differential Item Functioning Methodology." In *Differential Item Functioning*, by P W Holland and H Wainer, 3–23. NJ, US: Lawrence Erlbaum Associates Inc.
- Artelt, C, and T Rausch. 2014. "Accuracy of Teacher Judgment." In *Teachers' Professional Development*, by S Krolak-Schwerdt, 27–43. Rotterdam, the Netherlands: Sense Publishers.
- Asparouhov, T, and B Muthén. 2010. "Bayesian Analysis Using Mplus: Technical Implementation." Technical appendix. Los Angeles: Muthen & Muthen. www.statmodel.com.
- . 2014. "Multi-Group Factor Analysis Alignment." *Structural Equation Modeling* 21: 1–14.
- Bakker, B F M, J Van Rooijen, and L Van Toor. 2014. "The System of Social Statistical Datasets of Statistics Netherlands: An Integral Approach to the Production of Register-Based Social Statistics." *Statistical Journal of the IAOS* 30, 411–24.
- Barnett, A G, J Van Der Pols, and A J Dobson. 2005. "Regression to the Mean: What It Is and How to Deal with It." *International Journal of Epidemiology* 34: 215–20.
- Baron, R M, D Y H Tom, and H M Cooper. 1985. "Social Class, Race, and Teacher Expectations." In *Teacher Expectancies*, by J B Dusek, 251–69. Hillsdale, NJ: Erlbaum.
- Baumert, J, and M Kunter. 2013. "The COACTIV Model of Teachers' Professional Competence." In *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers: Results from the COACTIV Project*, by M Kunter, J Baumert, W Blum, S Klusmann, S Kruss, and M Neubrand, 25–48. New York: Springer.
- Baysu, G, A Alanya, and H A G De Valk. 2018. "School Trajectories of the Second Generation of Turkish Immigrants in Sweden, Belgium, Netherlands, Austria and

- Germany: The Role of School Systems.” *International Journal of Comparative Sociology* 59 (5): 451–79.
- Becker, G S. 2010. *The Economics of Discrimination*. University of Chicago press.
- Becker, R, F Jäpel, and M Beck. 2013. “Diskriminierung Durch Lehrpersonen Oder Herkunftsbedingte Nachteile von Migranten Im Detuschweizer Schulsystem.” *Swiss Journal of Sociology*, no. 39: 517–49.
- Bentler, P M, and D G Bonett. 1980. “Significance Tests and Goodness of Fit in the Analysis of Covariance Structures.” *Psychological Bulletin* 88: 588–606.
- Berger, J O, J M Bernardo, and D Sun. 2009. “The Formal Definition of Reference Priors.” *Annual Statistics* 37: 905–38.
- Bernardo, J M. 1979. “Reference Posterior Distributions for Bayesian Inference.” *Journal of the Royal Statistical Society B Methodology* 41: 113–47.
- Bernardo, J M, and A F M Smith. 2000. *Bayesian Theory*. New York: Wiley Series in Probability; Statistics; John Wiley & Sons.
- Bier, V. 2004. “Implications of the Research on Expert Overconfidence and Dependence.” *Reliability Engineering and System Safety* 85: 321–29.
- Biesta, G.J.J. 2012. *Goed Onderwijs En de Cultuur van Het Meten. Ethiek, Politie En Democratie*. Den Haag: Boom Lemma.
- Billiet, J. 2003. “Cross-Cultural Equivalence with Structural Equation Modeling.” In *Cross-Cultural Survey Methods*, by J A Harkness, F Van De Vijver, and P P Mohler, 247–64. New York: John Wiley.
- Blauw, S. 2018. *Het Bestverkochte Boek Ooit (Met Deze Titel)*. Amsterdam: De Correspondent.
- Boone, S, and M Van Houtte. 2013. “Why Are Teacher Recommendations at the Transition from Primary to Secondary Education Socially Biased? A Mixed-Method Research.” *British Journal of Sociology of Education* 34 (1): 20–38.
- Borghans, L, R Diris, and T Schils. 2018. “Sociale Ongelijkheid in Het Onderwijs Is Hardnekkig.” *ESB Onderwijs En Wetenschap* 103.
- Borsboom, D. 2006a. “Can We Bring About a Velvet Revolution in Psychological Measurement? A Rejoinder to Commentaries.” *Psychometrika* 71 (3): 463–67.
- . 2006b. “The Attack of the Psychometricians.” *Psychometrika* 71 (3): 425–40.
- Borsboom, D, and J Van Heerden. 2003. “The Theoretical Status of Latent Variables.” *Psychological Review* 110 (2): 203–219.
- Borsboom, D, G J Mellenbergh, and J Van Heerden. 2002. “Different Kinds of DIF: A Distinction Between Absolute and Relative Forms of Measurement Invariance and Bias.” *Applied Psychological Measurement* 26 (4): 433–50.
- . 2004. “The Concept of Validity.” *Psychological Review* 111 (4): 1061–71.
- Borsboom, D, J-W Romeijn, and J M Wicherts. 2008. “Measurement Invariance Versus Selection Invariance: Is Fair Selection Possible?” *Psychological Methods* 13 (2): 75–98.
- Bosch, M, J Konermann, C De Wit, M Rutten, and M Amsing. 2008. “Passende

- Overgang. Een Verkenning Naar de Stand van Zaken Rond de Overgang Tussen Primair En Voortgezet Onderwijs.” ’s Hertogenbosch: KPC Groep.
- Boudett, K P, E A City, and R J Murnane. 2013. *Data Wise: A Step by Step Guide to Using Assessment Results to Improve Teaching and Learning*. MA: Cambridge University Press.
- Bousquet, N. 2008. “Diagnostics of Prior-Data Agreement in Applied Bayesian Analysis.” *Journal of Applied Statistics* 35: 1011–29.
- Box, G. 1980. “Sampling and Bayes’ Inference in Scientific Modelling and Robustness.” *Journal of the Royal Statistical Society A* 143: 383–430.
- Brannick, M T. 1995. “Critical Comments on Applying Covariance Structure Modeling.” *Journal of Organizational Behavior* 16: 201–13.
- Breen, R, and J O Jonsson. 2000. “Analyzing Educational Careers: A Multinomial Transition Model.” *American Sociological Review* 65 (5): 754–72.
- Brennan, R L. 2006. *Educational Measurement*. 4th edition. Praeger Publishers.
- . 2010. “Generalizability Theory and Classical Test Theory.” *Applied Measurement in Education* 24 (1): 1–21.
- Brennan, R L, and W Lee. 1999. “Conditional Scale-Score Standard Errors of Measurement Under Binomial and Compound Binomial Assumptions.” *Educational and Psychological Measurement* 59: 5–24.
- Brown, T A. 2015. *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Publications.
- Brown, W. 1910. “Some Experimental Results in the Correlation of Mental Abilities.” *British Journal of Psychology* 3: 296–322.
- Byrne, B M, R J Shavelson, and B Muthen. 1989. “Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance.” *Psychological Bulletin* 105 (3): 456–66.
- Carmichael, C. 2015. “Discrepancies Between Standardized Testing and Teacher Judgements in an Australian Primary School Context.” *Mathematics Teacher Education and Development* 17 (1): 62–75.
- Cha, S-H. 2007. “Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions.” *Int. J. Math. Model. Methods Appl. Sci.* 1.
- Chang, W, J Cheng, J J Allaire, Y Xie, and J McPherson. 2017. *Shiny: Web Application Framework for R. R Package Version 1.0.5*. <https://CRAN.R-project.org/package=shiny>.
- Charter, R A, and L S Feldt. 2001. “Confidence Intervals for True Scores: Is There a Correct Approach?” *Journal of Psychoeducational Assessment* 19: 350–64.
- Chen, F F. 2007. “Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance.” *Structural Equation Modeling* 14 (3): 464–504.
- Cheung, G W, and R B Rensvold. 2002. “Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance.” *Structural Equation Modeling* 9 (2): 233–55.
- Cheung, M W-L, and K Au. 2005. “Applications of Multilevel Structural Equation

- Modeling to Cross-Cultural Research.” *Structural Equation Modeling* 12 (4): 598–619.
- Chmielewski, A K. 2017. “The Global Increase in the Socioeconomic Achievement Gap, 1964-2015.” 1704.
- Cieciuch, J, E Davidov, and P Schmidt. 2018. “Alignment Optimization: Estimation of the Most Trustworthy Means in Cross-Cultural Studies Even in the Presence of Noninvariance.” In *Cross Cultural Analysis: Methods and Applications*, by E Davidov, P Schmidt, and J Billiet, 571–92. New York: Routledge.
- Cieciuch, J, E Davidov, P Schmidt, R Algesheimer, and S H Schwartz. 2014. “Comparing Results of an Exact Versus an Approximate (Bayesian) Measurement Invariance Test: A Cross-Country Illustration with a New Scale to Measure 19 Human Values.” *Frontiers in Psychology* 5: 982.
- Claassen, A, and L Mulders. 2003. “Leerlingen Na de Overstap. Een Vergelijking van Vier Cohorten Leerlingen Na de Overgang van Basisonderwijs Naar Voortgezet Onderwijs Met Nadruk Op de Positie van Doelgroepopleerlingen van Het Onderwijsachterstanden Beleid.” Nijmegen, the Netherlands: ITS, Radboud Universiteit Nijmegen.
- Clemen, R T, and T Reilly. 2001. *Making Hard Decisions with Decision Tools*. Pacific Grove, CA: Duxbury Press.
- Clemen, R T, and R L Winkler. 1999. “Combining Probability Distributions from Experts in Risk Analysis.” *Risk Analysis* 19 (2): 187–203.
- Colton, D A, X Gao, and M J Kolen. 1996. “Assessing the Reliability of Performance Level Scores Using Bootstrapping.” New York.
- Cornwell, C, D B Mustard, and J Van Parys. 2012. “Noncognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School.” *The Journal of Human Resources* 48 (1): 236–64.
- Cox, D R, and D V Hinkley. 1974. *Theoretical Statistics*. London, UK: Chapman & Hall.
- Crawford, J R. 2004. “Psychometric Foundations of Neuropsychological Assessment.” In *Clinical Neuropsychology: A Practical Guide to Assessment and Management for Clinicians*, by L H Goldstein and J E McNeil, 121–40. John Wiley & Sons.
- Crawford, J R, and P H Garthwaite. 2009. “Percentiles Please: The Case for Expressing Neuropsychological Test Scores and Accompanying Confidence Limits as Percentile Ranks.” *The Clinical Neuropsychologist* 23 (2): 193–204.
- Creighton, T B. 2007. *School and Data: The Educator’s Guide for Using Data to Improve Decision Making*. London: Sage.
- Cronbach, L J. 1991. “Methodological Studies - a Personal Retrospective.” In *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach*, by R E Snow and D E Wiley, 385–400. Hillsdale, NJ: Erlbaum.
- . 2004. “My Current Thoughts on Coefficient Alpha and Successor Procedures.” *Educational Psychological Measurements* 64: 391–418.
- Cronbach, L J, and P E Meehl. 1995. “Construct Validity in Psychological Tests.” *Psychological Bulletin* 52 (4): 281–302.
- Cronbach, L J, G C Gleser, H Nanda, and N Rajaratnam. 1972. *The Dependability of*

Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley.

Davidov, E. 2008. "A Cross-Country and Cross-Time Comparison of the Human Values Measurements with the Second Round of the European Social Survey." *Survey Research Methods* 2 (1): 33–46.

———. 2010. "Testing for Comparability of Human Values Across Countries and Time with the Third Round of the European Social Survey." *International Journal of Comparative Sociology* 51 (3): 171–91.

Davidov, E, J Cieciuch, B Meuleman, P Schmidt, R Algesheimer, and M Hausherr. 2015. "The Comparability of Measurements of Attitudes Toward Immigration in the European Social Survey Exact Versus Approximate Measurement Equivalence." *Public Opinion Quarterly* 79 (S1): 244–66.

Davidov, E, H Dülmer, J Cieciuch, A Kuntz, D Seddig, and P Schmidt. 2018. "Explaining Measurement Non-Equivalence Using Multilevel Structural Equation Modeling." *Sociological Methods & Research* 47 (4): 729–60.

Davidov, E, H Dülmer, E Schlüter, P Schmidt, and B Meuleman. 2012. "Using a Multilevel Structural Equation Modeling Approach to Explain Cross-Cultural Measurement Noninvariance." *Journal of Cross-Cultural Psychology* 43 (4): 558–75.

Davidov, E, B Meuleman, J Cieciuch, P Schmidt, and J Billiet. 2014. "Measurement Equivalence in Cross-National Research." *Annual Review of Sociology* 40 (1): 55–75.

Davidov, E, P Schmidt, and J Billiet. 2010. *Cross-Cultural Analysis: Methods and Applications*. New York: Taylor & Francis.

Davidov, E, P Schmidt, and S H Schwartz. 2008. "Bringing Values Back in: The Adequacy of the European Social Survey to Measure Values in 20 Countries." *Public Opinion Quarterly* 72 (3): 420–45.

Davis, J A. 1966. "The Campus as a Frog Pond: An Application of the Theory of Relative Deprivation to Career Decisions of College Men." *American Journal of Sociology* 72 (1): 17–31.

De Boeck, P. 2008. "Random Item IRT Models." *Psychometrika* 73: 533–59.

De Groot, A. D. 1966. *Vijven En Zessen*. Groningen: Wolters.

"De Staat van Het Onderwijs 2019." 2019. Inspectie van het Onderwijs.

Deci, E L, and R M Ryan. 2016. "Optimizing Students' Motivation in the Era of Testing and Pressure: A Self-Determination Theory Perspective." In *Building Autonomous Learners*, by W Lie, J Wang, and R Ryan, 9–29. Singapore: Springer.

Depaoli, S, and R Van De Schoot. 2015. "Improving Transparency and Replication in Bayesian Statistics: The WAMBS-Checklist." *Psychological Methods, Advance Online Publication*.

DeVellis, R F. 2006. "Classical Test Theory." *Medical Care* 44 (11): 50–59.

Deza, M M, and E Deza. 2009. *Encyclopedia of Distances*. Berlin, Germany: Springer.

Ditton, H, J Krüsken, and M Schauenberg. 2005. "Bildungsungleichheit - Der Beitrag von Familie Und Schule." *Zeitschrift Für Erziehungswissenschaft* 8: 285–304.

Dolin, J, P Black, W Harlen, and A Tiberghien. 2018. "Exploring Relations Between

- Summative and Formative Assessment.” In *Transforming Assessment: Through Interplay Between Practice, Research and Policy*, by J Dolin and R Evans, 53–80. Springer.
- Dorans, N J, and K Zeller. 2004. “Examining Freedle’s Claims About Bias and His Proposed Solution: Dated Data, Inappropriate Measurement, and Incorrect and Unfair Scoring.” Research report RR-04-26. ETS.
- Driessen, G. 1991. “Discrepancies Tussen Toetsresultaten En Doorstroomniveau. Positieve Discriminatie Bij de Overgang Basisonderwijs - Voortgezet Onderwijs?” *Pedagogische Studiën* 68: 27–35.
- Driessen, G, P Slegers, and F Smit. 2008. “The Transition from Primary to Secondary Education: Meritocracy and Ethnicity.” *European Sociological Review* 24 (4): 527–42.
- Dronkers, J. 2013. “Citotoets Is Helaas Nog Lang Niet Overbodig.” *De Volkskrant*, March.
- Dumont, H, D Klinge, and K Maaz. 2019. “The Many (Subtle) Ways Parents Game the System: Mixed-Method Evidence on the Transition into Secondary-School Tracks in Germany.” *Sociology of Education* 92 (2): 199–228.
- Dunlosky, J, and J Metcalfe. 2009. *Metacognition*. Thousand Oaks, CA: Sage.
- Dusek, J B, and G Joseph. 1983. “The Bases of Teacher Expectancies: A Meta-Analysis.” *Journal of Educational Psychology* 75 (3): 327.
- Edlmann, K, J Bensabat, A Niemi, R S Haszeldine, and C I McDermott. 2016. “Lessons Learned from Using Expert Elicitation to Identify, Assess and Rank the Potential Leakage Scenarios at the Heletz Pilot CO₂ Injection Site.” *International Journal of Greenhouse Gas Control* 49: 473–87.
- Embretson, S E. 1996. “The New Rules of Measurement.” *Psychological Assessment* 8 (4): 341–49.
- Erikson, R, J H Goldthorpe, and L Portocarero. 1979. “Intergenerational Class Mobility in Three West European Countries: England, France and Sweden.” *British Journal of Sociology* 30: 415–41.
- Espin, C A, M M Wayman, S L Deno, K L McMaster, and M de Rooij. 2017. “Data-Based Decision Making: Developing a Method for Capturing Teachers’ Understanding of CBM Graphs.” *Learning Disability Research Practice* 32: 8–21. doi:10.1111/ldrp.12123.
- Evans, J S B. 2008. “Dual-Processing Accounts for Reasoning, Judgment, and Social Cognition.” *Annual Review of Psychology* 59 (1): 255–78.
- Evans, M. 2015. *Measuring Statistical Evidence Using Relative Belief*. Didcot, UK: Taylor & Francis.
- Evans, M, and G-H Jang. 2011. “A Limit Result for the Prior Predictive Applied to Checking for Prior-Data Conflict.” *Statistical Probability Letters* 81: 1034–8.
- Evans, M, and H Moshonov. 2006. “Checking for Prior-Data Conflict.” *Bayesian Analysis* 1: 893–914.
- . 2007. “Checking for Prior-Data Conflict with Hierarchically Specified Priors.” In *Bayesian Statistics and Its Applications*, by A Upadhyay, U Singh, and D Dey, 145–59. New Delhi, India: Anamaya Publishers.
- Evers, A, W Lucassen, R Meijer, and K Sijtsma. 2010. “COTAN Beoordelingssysteem

Voor de Kwaliteit van Tests (Geheel Herziene Versie).” FMG: Psychology Research Institute.

Exalto, R, D Sipkens, T Klein, and D Kooij. 2019. “Terecht Overstaprecht? Doorstroom Havo-Vwo.” <https://www.rijksoverheid.nl/documenten/rapporten/2019/01/10/terecht-overstaprecht-doorstroom-havo-vwo>.

Faber, M, M Van Geel, and A Visscher. 2013. “Digitale Leerlingvolgsystemen Als Basis Voor Opbrengstgericht Werken in Het Primair Onderwijs: Een Analyse van de Wijze Waarop Scholen En Besturen de Mogelijkheden van Digitale Leerlingvolgsystemen Kunnen Benutten.” Kennisnet.

Feldt, L S, and A L Qualls. 1996. “Estimation of Measurement Error Variance at Specific Score Levels.” *Journal of Educational Measurement* 33 (2): 141–56.

———. 1998. “Approximating Scale Score Standard Error of Measurement from the Raw Score Standard Error.” *Applied Measurement in Education* 11: 159–77.

Feron, E. 2018. “The Role of Cognitive Tests and Teachers in the Transition from Primary to Secondary Education.” PhD thesis, Maastricht, Netherlands: Maastricht University. 10.26481/dis.20180629ef.

Feron, E, T Schils, and B Ter Weel. 2015. “Does the Teacher Beat the Test? The Additional Value of Teacher Assessment in Predicting Student Ability.” <https://www.cpb.nl/sites/default/files/publicaties/download/cpb-discussion-paper-300-does-teacher-beat-test.pdf>.

Finch, W H, J E Bolin, and K Kelley. 2014. *Multilevel Modeling Using R*. Statistic in the Social and Behavioral Sciences Series. Boca Raton: Chapman & Hall.

Fox, J-P, and S Mariani. 2017. “Person-Fit Statistics for Joint Models for Accuracy and Speed.” *Journal of Educational Measurement* 54 (2): 243–62.

Fox, J-P, and A J Verhagen. 2010. “Random Item Effects Modeling for Cross-National Survey Data.” In *Cross-Cultural Analysis: Methods and Applications*, by E Davidov, P Schmidt, and J Billiet, 467–88. London, UK: Routledge Academic.

Freedle, R O. 2003. “Correcting the SAT’s Ethnic and Social Bias: A Method for Reestimating SAT Scores.” *Harvard Educational Review* 73: 1–43.

Gabriele, A J, and K H Park. 2016. “Elementary Mathematics Teachers’ Judgment Accuracy and Calibration Accuracy: Do They Predict Students’ Mathematics Achievement Outcomes?” *Learning and Instruction* 45: 49–60. doi:10.1016/j.learninstruc.2016.06.008.

Gallagher, C J. 2003. “Reconciling a Tradition of Testing with a New Learning Paradigm.” *Educational Psychology Review* 15 (1): 83–99.

Ganzeboom, H B G, P M De Graaf, and D J Treiman. 1992. “A Standard International Socio-Economic Index of Occupational Status.” *Social Science Research* 21: 1–56.

Gardner, J. 2013. “The Public Understanding of Error in Educational Assessment.” *Oxford Review of Education* 39 (1): 72–92.

Gelman, A, J B Carlin, H S Stern, and D B Rubin. 2003. *Bayesian Data Analysis*. Second Edition. CRC Press.

Gelman, A, W R Gilks, S Richardson, and D J Spiegelhalter. 1996. “Inference and Monitoring Convergence.” In *Markov Chain Monte Carlo in Practice*. London, UK:

Chapman; Hall.

Geven, S A J, A H Batruch, and H Van De Werfhorst. 2018. "Inequality in Teacher Judgements, Expectations and Track Recommendations: A Review Study." Rijksoverheid.

Glock, S, and S Krolak-Schwerdt. 2013. "Does Nationality Matter? The Impact of Stereotypical Expectations on Student Teachers' Judgments." *Social Psychology of Education* 16: 111–27.

Glutting, J, and T Oakland. 1993. *GATSB Guide to the Assessment of Test Session Behavior for the WISC-III and the WIAT*. Psychological Corporation.

Goldstein, D G, and D Rothschild. 2014. "Lay Understanding of Probability Distributions." *Judgements in Decision Making* 9: 1–14.

Goldstein, D G, E J Johnson, and W F Sharpe. 2008. "Choosing Outcomes Versus Choosing Products: Consumer-Focused Retirement Investment Advice." *Journal of Consumer Research* 35: 440–56.

Good, T L, and J E Brophy. 1997. *Looking in Classrooms (7th Ed.)*. New York: Longman.

Goossens, L H J, R M Cooke, A R Hale, and L Rodic-Wiersma. 2008. "Fifteen Years of Expert Judgement at TUDelft." *Safety Science* 46: 234–44.

Gore, S. 1987. "Biostatistics and the Medical Research Council." *Medical Research Council News* 35: 19–20.

Gross, C, A Gottburgsen, and A Phoenix. 2016. "Education Systems and Intersectionality." In *Education Systems and Inequalities: International Comparisons*, by A Hadjar and C Gross, 15–72. Bristol University Press, Policy Press.

Guilford, J P. 1936. *Psychometric Methods*. New York: McGraw-Hill.

Gulliksen, H. 1950. *Theory of Mental Tests*. New York: Wiley.

Gwet, K L. 2008. "Intrarater Reliability." *Wiley Encyclopedia of Clinical Trials*, 1–14.

Hachfeld, A, Y Anders, S Schroeder, P Stanat, and M Kunter. 2010. "Does Immigration Background Matter? How Teachers' Predictions of Students' Performance Relate to Student Background." *International Journal of Educational Research* 49: 78–91.

Haertel, E H. 2006. "Reliability." In *Educational Measurement*, by R L Brennan, 65–110. Westport, CT: American Council on Education/Praeger.

Haladyna, T, N Haas, and J Allison. 1998. "Continuing Tensions in Standardized Testing." *Childhood Education* 74 (5): 262–73.

Hald, T, W Aspinall, B Devleesschauwer, R M Cooke, T Corrigan, A H Havelaar, and S Hoffman. 2016. "World Health Organization Estimates of the Relative Contributions of Food to the Burden of Disease Due to Selected Foodborne Hazards: A Structured Expert Elicitation." *PLOS Medicine* 11 (1): 1–35.

Hambleton, R K, and R W Jones. 1993. "Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development." *Instructional Topics in*

Educational Measurement, 38–47.

Hanan, U, D A Moore, and C K Morewedge. 2010. “A Simple Remedy for Overprecision in Judgment.” *Judgment and Decision Making* 5 (7): 467–76.

Havelaar, A H, A V Galindo, D Kurowicka, and R M Cooke. 2008. “Attribution of Foodborne Pathogens Using Structured Expert Elicitation.” *Foodborne Pathogens and Disease* 5 (5): 649–59.

He, Q. 2009. “Estimating the Reliability of Composite Scores.” Ofqual.

Heisser, W J. 2006. “Measurement Without Copper Instruments and Experiment Without Complete Control.” *Psychometrika* 71 (3): 457–61.

Hendrickson, A. 2007. “An NCME Instructional Module on Multistage Testing.” *Educational Measurement: Issues and Practice* 26: 44–52.

Herppich, S, A K Praetorius, N Forster, I Glogger-Frey, K Karst, D Leutner, and A Südkamp. 2018. “Teachers’ Assessment Competence: Integrating Knowledge-, Process- and Product-Oriented Approaches into a Competence-Oriented Conceptual Model.” *Journal of Teaching and Teacher Education* 76: 181–93.

Hirsch, S. 2017. “Measurement in Education: Test Scores and Beyond.” PhD thesis, Maastricht, Netherlands: Maastricht University. 10.26481/dis.20170111sh.

Hoekstra, R, R D Morey, J N Rouder, and E-J Wagenmakers. 2014. “Robust Misinterpretation of Confidence Intervals.” *Psychonomic Bulletin & Review* 21 (5): 1157–64.

Hoijsink, H, and R Van De Schoot. 2018. “Testing Small Variance Priors Using Prior-Posterior Predictive P Values.” *Psychological Methods* 23: 561–69.

Holden, R R. 2010. “Face Validity.” In *Corsini Encyclopedia of Psychology*, by I B Weiner and W E Craighead, 637–38. New York: Wiley.

Holder, K, and U Kessels. 2017. “Gender and Ethnic Stereotypes in Student Teachers’ Judgments: A New Look from a Shifting Standards Perspective.” *Social Psychology of Education* 20: 471–90.

Holland, P W, and H Wainer. 2012. *Differential Item Functioning*. New York: Routledge.

Hoogstad, M. 2012. “Trainen Voor de Heilige Cito-Toets.” *NRC*, February.

Hox, J, M Moerbeek, and R Van De Schoot. 2017. *Multilevel Analysis: Techniques and Applications*. 3rd ed. Athenaeum Uitgeverij.

Hu, Y, J R Nesselroade, M K Erbacher, S M Boker, S A Burt, P K Keel, M C Neale, C L Sisk, and K Klump. 2016. “Test Reliability at the Individual Level.” *Structural Equation Modeling* 23: 532–43.

Huff, D. 1993. *How to Lie with Statistics*. USA: New York: Norton paperback.

Jacob, M, and N Tieben. 2007. “Social Selectivity of Track Mobility in Secondary Schools. a Comparison of Intra-Secondary Transitions in Germany and the Netherlands.” 105. University of Mannheim.

Jak, J S, F J Oort, and C V Dolan. 2013. “A Test for Cluster Bias: Detecting Violations of Measurement Invariance Across Clusters in Multilevel Data.” *Structural Equation*

- Modeling* 20 (2): 265–82.
- . 2014a. “Measurement Bias in Multilevel Data.” *Structural Equation Modeling* 21 (1): 31–39.
- . 2014b. “Using Two-Level Factor Analysis to Test for Cluster Bias in Ordinal Data.” *Multivariate Behavioral Research* 49 (6): 544–53.
- Jang, G-H. 2010. “Invariant Procedures for Model Checking, Checking for Prior-Data Conflict and Bayesian Inference.” PhD thesis, Toronto, ON, Canada: University of Toronto.
- Jansen, I. 2015. “Citostress. 2Doc.”
- Janson, H, and U Olsson. 2001. “A Measure of Agreement for Interval or Nominal Multivariate Observations.” *Educational and Psychological Measurement* 61 (2): 277–89.
- Jenrich, R I. 2006. “Rotation to Simple Loadings Using Component Loss Functions: The Oblique Case.” *Psychometrika* 71 (1): 173–91.
- Jimerson, J B. 2014. “Thinking About Data: Exploring the Development of Mental Models for ‘Data Use’ Among Teachers and School Leaders.” *Studies in Educational Evaluation* 42: 5–14.
- Johnson, P C D. 2014. “Extension of Nakagawa & Schielzeth’s R² GLMM to Random Slopes Models.” *Methods in Ecology and Evolution* 5: 944–46.
- Johnson, S R, G A Tomlinson, G A Hawker, J T Granton, H A Grosbein, and B M Feldman. 2010a. “A Valid and Reliable Belief Elicitation Method for Bayesian Priors.” *Journal of Clinical Epidemiology*, 63: 370–83.
- . 2010b. “Methods to Elicit Beliefs for Bayesian Priors: A Systematic Review.” *Journal of Clinical Epidemiology*, 63: 355–69.
- Jussim, L, and K D Harber. 2005. “Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies.” *Personality and Social Psychology Review* 9 (2): 131–55.
- Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus; Giroux.
- Kahneman, D, and S Frederick. 2005. “A Model of Heuristic Judgment.” In *Cambridge Handbook of Thinking and Reasoning*, by J H Keith and R G Morrison, 267–93. Cambridge: Cambridge University Press.
- Kaiser, J, J Reteldorf, A Südkamp, and J Möller. 2013. “Achievement and Engagement: How Student Characteristics Influence Teacher Judgments.” *Learning and Instruction* 28.
- Kaiser, J, A Südkamp, and J Möller. 2017. “The Effects of Student Characteristics on Teachers’ Judgment Accuracy: Disentangling Ethnicity, Minority Status, and Achievement.” *Journal of Educational Psychology* 109 (6): 871–88.
- Kalmijn, M, and G Kraaykamp. 2003. “Dropout and Downward Mobility in the Educational Career: An Event-History Analysis of Ethnic Schooling Differences in the Netherlands.” *Educational Research and Evaluation* 9 (3): 265–87.
- Kamerman, S, and J Vasterman. 2015. “De Leerkracht Weet Het Vaak écht Beter Dan

de Citotoets.” *NRC.next*, March.

“Kansen(on)gelijkheid Bij de Overgangen PO-VO: Bevindingen En Bevorderende En Belemmerende Factoren.” 2018. Inspectie van het Onderwijs.

Kaplan, D, and S Depaoli. 2013. “Bayesian Statistical Methods.” In *Oxford Handbook of Quantitative Methods*, by T D Little, 407–37. Oxford, UK: Oxford University Press.

Karels, Machiel. 2019. “Wij-Leren.nl. 10 Pittige Problemen Met de Eindtoets Basisonderwijs.” April 3. <https://wij-leren.nl/10-pittige-problemen-met-de-centrale-eindtoets.php>.

Kautz, T, J J Heckman, R Diris, B Ter Weel, and L Borghans. 2014. “Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success.” 110. OECD.

Kelley, T L. 1927. *Interpretation of Educational Measurements*. New York: MacMillan.

———. 1947. *Fundamentals of Statistics*. Cambridge: Harvard University Press.

Kincheloe, J, S Steinberg, and A Gresson. 1996. *Measured Lies: The Bell Curve Examined*. New York: St. Martins.

Kirk, M D, S M Pires, R E Black, M Caipo, J A Crump, B Devleeschauwer, and F J Angulo. 2015. “World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal and Viral Diseases 2010: A Data Synthesis.” *PLOS Medicine* 12 (12): 1–21.

Klapproth, F, S Glock, M Bohmer, S Krolak-Schwerdt, and R Martin. 2012. “School Placement Decisions in Luxembourg: Do Teachers Meet the Education Ministry’s Standards?” *Literacy Information and Computer Education Journal* 1 (1): 765–71.

Kloosterman, R, and P M De Graaf. 2010. “Non-Promotion or Enrolment in a Lower Track? The Influence of Social Background on Choices in Secondary Education for Three Cohorts of Dutch Pupils.” *Oxford Review of Education* 36 (3): 363–84.

Kneyber, R, and J Evers. 2013. *Het Alternatief. Weg Met de Afrekencultuur in Het Onderwijs*. Amsterdam: Boom.

Knight, K. 2000. *Mathematical Statistics*. New York: Chapman & Hall.

Knoester, M, and A Wayne. 2017. “Standardized Testing and School Segregation: Like Tinder for Fire?” 20 (1): 1–14.

Knol, A B, P Slottje, J P van der Sluijs, and E Lebet. 2010. “The Use of Expert Elicitation in Environmental Health Impact Assessment: A Seven Step Procedure.” *Environmental Health* 9 (19): 1–16.

Korpershoek, H, C Beijer, M Spithoff, H M Naaijer, and A C Timmermans. 2016. “Overgangen En Aansluitingen in Het Onderwijs. Deelrapportage 1: Reviewstudie Naar de Po-Vo En de Vmbo-Mbo Overgang.”

König, C, and R Van De Schoot. 2017. “Bayesian Statistics in Educational Research – a Look at the Current State of Affairs.” *Educational Review*, 1–24.

Kristen, C. 2000. “Ethnic Differences in Educational Placement: The Transition from Primary to Secondary Schooling.” Mannheim: MZES.

Krolak-Schwerdt, S, M Böhmer, and C Gräsel. 2012. “Leistungsbeurteilung von

- Schulkindern: Welche Rolle Spielen Ziele Und Expertise Der Lehrkraft?" *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, 44: 111–22.
- Kruschke, J. 2011. *Doing Bayesian Data Analysis: A Tutorial Introduction with R*. San Diego, CA: Academic Press.
- Kruschke, J, H Arguinis, and H Joo. 2012. "The Time Has Come! Bayesian Methods for Data Analysis in the Organizational Sciences." *Organizational Research Methods* 15: 722–52.
- Kuhnert, P M, T G Martin, and S P Griffiths. 2010. "A Guide to Eliciting and Using Expert Knowledge in Bayesian Ecological Models." *Ecological Letters* 13: 900–914.
- Kullback, S, and R A Leibler. 1951. "On Information and Sufficiency." *Annual Mathematical Statistics* 1: 79–86.
- Kunda, Z, and S J Spencer. 2003. "When Do Stereotypes Come to Mind and When Do They Color Judgments? A Goal-Based Theoretical Framework for Stereotype Activation and Application." *Psychological Bulletin* 129: 522–44.
- Lee, S-Y. 2007. *Structural Equation Modeling: A Bayesian Approach*. New York: Wiley & Sons.
- Lee, W, R L Brennan, and M J Kolen. 2000. "Estimators of Conditional Scale-Score Standard Errors of Measurement: A Simulation Study." *Journal of Educational Measurement* 37 (1): 1–20.
- Lek, K, and R Van De Schoot. 2018. "Development and Evaluation of a Digital Expert Elicitation Method Aimed at Fostering Elementary School Teachers' Diagnostic Competence." *Frontiers in Education* 3: 82. doi:10.3389/educ.2018.00082.
- Lezak, M D, D B Howieson, D W Loring, H J Hannay, and J S Fisher. 2004. *Neuropsychological Assessment, 4th*. New York: Oxford University Press.
- Lichtenstein, S, B Fischhoff, and L D Philips. 1982. "Calibration of Probabilities: The State of the Art to 1980." In *Judgment Under Uncertainty: Heuristics and Biases*. Great Britain: Cambridge University Press.
- Liese, F, and I Vajda. 2006. "On Divergences and Informations in Statistics and Information Theory." *IEEE Trans. Inf. Theory* 52: 4394–4412.
- Lindahl, E. 2007. "Comparing Teachers' Assessments and National Test Results - Evidence from Sweden." 24. IFAU - Institute for labour market policy evaluation.
- Lindler, F, D Van Roon, and B F M Bakker. 2011. "Combining Data from Administrative Sources and Sample Surveys; the Single Variable Case." In *ESSnet Data Integration. WP4 Case Studies*, 39–97. Luxembourg: Eurostat.
- Linn, R L. 2001. "A Century of Standardized Testing: Controversies and Pendulum Swings." *Educational Assessment* 7 (1): 29–38.
- Little, R J. 1988. "Missing-Data Adjustments in Large Surveys." *Journal of Business & Economic Statistics* 6 (3): 287–96.
- Lommen, M J J, R Van De Schoot, and I M Engelhard. 2014. "The Experience of Traumatic Events Disrupts the Measurement Invariance of a Posttraumatic Stress Scale." *Frontiers in Psychology* 5: 1–7.
- Lord, F M. 1952. "A Theory of Test Scores." Psychometric Monograph No. 7.

Psychometric Society, Center for Educational Research; Evaluation, University of North Carolina at Greenboro.

———. 2012. *Applications of Item Response Theory to Practical Testing Problems*. Routledge.

Lord, F M, and M R Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Publishing Company.

MacCallum, R C, M Roznowski, and L B Necowitz. 1992. “Model Modifications in Covariance Structure Analysis: The Problem of Capitalizing on Chance.” *Psychological Bulletin* 111: 490–504.

Machts, N, J Kaiser, F T C Schmidt, and J Möller. 2016. “Accuracy of Teachers’ Judgments of Students’ Cognitive Abilities: A Meta-Analysis.” *Educational Research Review* 19: 85–103.

Mandinach, E B. 2012. “A Perfect Time for Data Use: Using Data-Driven Decision Making to Inform Practice.” *Educational Psychologist* 47 (2): 71–85.

Mare, R D. 1980. “Social Background and School Continuation Decisions.” *Journal of the American Statistical Association* 75 (370): 295–305.

———. 1981. “Change and Stability in Educational Stratification.” *American Sociological Review* 45: 72–81.

Martin, T G, M A Burgman, F Fidler, P M Kuhnert, S Low Choy, M McBride, and K Mengersen. 2011. “Eliciting Expert Knowledge in Conservation Science.” *Conservation Biology* 26: 29–38.

McManus, I C. 2012. “The Misinterpretation of the Standard Error of Measurement in Medical Education: A Primer on the Problems, Pitfalls and Peculiarities of the Three Different Standard Errors of Measurement” 34 (7): 569–76.

Meissel, K, F Meyer, E S Yao, and C M Rubie-Davies. 2017. “Subjectivity of Teacher Judgments: Exploring Student Characteristics That Influence Teacher Judgments of Student Ability.” *Teaching and Teacher Education* 65: 48–60.

Meitinger, K. 2014. “Necessary but Insufficient: Why Measurement Invariance Tests Need Online Probing as a Complementary Tool.” Yokohama, Japan.

Mellenbergh, G J. 1989. “Item Bias and Item Response Theory.” *International Journal of Educational Research* 13 (2): 127–43.

———. 2011. *A Conceptual Introduction to Psychometrics: Development, Analysis and Application of Psychological and Educational Tests*. The Hague, Netherlands: Eleven international publishing.

Merkle, E, and Y Rosseel. 2016. *Blavaan (R Package Version 0.1-3, Pp. 1-16)*. Vienna, Austria: The Comprehensive R Archive Network.

Mertens, F J H. 2016. *’Vijven En Zessen’ van Prof. Adriaan de Groot, Een Boekje Dat Geschiedenis Maakte*. Oosterwijk, the Netherlands: Wolf Legal Publishers.

Meuleman, B. 2012. “When Are Item Intercept Differences Substantively Relevant in Measurement Invariance Testing?” In *Methods, Theories and Empirical Applications in*

- the Social Sciences: Festschrift for Peter Schmidt*, by S Salzbom, E Davidov, and J Reinecke, 97–104. Heidelberg, Germany: Springer VS.
- . 2016. “Explaining Cross-National Inequivalence in Factor Loadings and Intercepts: A Bayesian Multilevel SEM Approach.” Chicago.
- Millsap, R E. 2011. *Statistical Approaches to Measurement Invariance*. New York: Routledge.
- Mizala, A, F Martinez, and S Martinez. 2015. “Pre-Service Elementary School Teachers’ Expectations About Student Performance: How Their Beliefs Are Affected by Their Mathematics Anxiety and Student’s Gender.” *Teaching and Teacher Education* 50: 70–78.
- Molenaar, P C M. 2004. “A Manifesto on Psychology as Idiographic Science: Bringing the Person Back into Scientific Psychology- This Time Forever.” *Measurement* 2: 201–18.
- . 2016. “Person-Oriented and Subject-Specific Methodology: Some Additional Remarks.” *Journal of Person-Oriented Research* 2 (2): 16–19.
- Moshonov, H. 2006. “Checking for Prior-Data Conflict.” *Bayesian Analysis* 1: 893–914.
- Muthén, B, and T Asparouhov. 2012. “Bayesian Structural Equation Modeling: A More Flexible Representation of Substantive Theory.” *Psychological Methods* 17 (3): 313–35.
- . 2013. “BSEM Measurement Invariance Analysis.” No. 17. <http://www.statmodel.com>.
- Muthén, L K, and B Muthén. 1998. *Mplus User’s Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.
- Nakagawa, S, P C D Johnson, and H Schielzeth. 2017. “The Coefficient of Determination R² and Intra-Class Correlation Coefficient from Generalized Linear Mixed-Effects Models Revisited and Expanded.” *Journal of the Royal Society Interface* 14 (134): 20170213.
- Neill, M. 2016. “The Testing Resistance and Reform Movement.” *Monthly Review*, March, 8–29.
- Niessen, S, and R Meijer. 2016. “De Leerkracht Is Geen Meetinstrument.” *NRC.next*, April.
- Nikulin, M S. 2001. “Hellinger Distance.” In *Encyclopedia of Mathematics*. Kluwer Academic Publishers.
- Nott, D J, W Xueou, M Evans, and B-G Englert. 2016. “Checking for Prior-Data Conflict Using Prior to Posterior Divergences.” *arXiv*. doi:arXiv:1611.00113.
- Novick, M R. 1966. “The Axioms and Principal Results of Classical Test Theory.” *Journal of Mathematical Psychology* 3: 1–18.
- NRC. 2019. “Al Dat Toetsen Maakt Nerveus En Biedt Schijnzekerheid.” *NRC Handelsblad*, January.
- Nunnally, J C, and I H Bernstein. 1994. *Psychometric Theory*. New York: McGraw-Hill.
- Oakley, J. 2017. *SHELF: Tools to Support the Sheffield Elicitation Framework. R Package Version 1.2.3*. <https://CRAN.R-project.org/package=SHELF>.
- Oberski, D L. 2014. “Evaluating Sensitivity of Parameters of Interest to Measurement

- Invariance in Latent Variable Models.” *Political Analysis* 22 (1): 45–60.
- Oberski, D L, J K Vermunt, and G B D Moors. 2015. “Evaluating Measurement Invariance in Categorical Data Latent Variable Models with the EPC-Interest.” *Political Analysis* 23 (4): 550–63.
- OECD. 2016. “Selecting and Grouping Students.” In *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*. Paris: OECD publishing. <https://doi.org/10.1787/9789264267510-9-en>.
- OESO. 2016. *Netherlands 2016: Foundations for the Future*. Parijs: OESO.
- Onderwijs, Inspectie van het. 2007. “Onderadvisering in Beeld.”
- . 2018. “De Staat van Het Onderwijs 2016-2017.”
- Onderwijsraad. 2018. “Advies Gelijke Kans Op Doorstroom Vmbo-Havo.” <https://www.onderwijsraad.nl/upload/documents/publicaties/volledig/Advies-doorstroom-vmbo-havo.pdf>.
- Oomens, M, F Scholten, and H Luyten. 2016. “Evaluatie Wet Eindtoetsing Po, Tussenrapportage.” <https://www.rijksoverheid.nl/documenten/rapporten/2017/01/27/tussenrapport-age-evaluatiewet-eindtoetsing-po>.
- O’Hagan, A. et al. 2006. *Uncertain Judgements: Eliciting Experts’ Probabilities*. John Wiley & Sons.
- O’Leary, R, S Low Choy, J V Murray, M Kynn, R Denham, T G Martin, and K Mengersen. 2009. “Comparison of Three Expert Elicitation Methods for Logistic Regression on Predicting the Presence of the Threatened Brush-Tailed Rock-Wallaby.” *Petrogale Penicillata. Environmetrics* 20: 379–98.
- Pameijer, N. 2014. “Waarom Een Ontwikkelingsperspectief Meer Is Dan IQ En Leerrendement.” <http://wijleren.nl/intelligentietest-passend-onderwijs.php>.
- Pit-Ten Cate, I, S Krolak-Schwerdt, and S Glock. 2016. “Accuracy of Teachers’ Tracking Decisions: Short- and Long-Term Effects of Accountability.” *European Journal of Psychology of Education* 31 (2): 225–43.
- Pit-Ten Cate, I, S Krolak-Schwerdt, S Glock, and M Markova. 2014. “Improving Teachers’ Judgments: Obtaining Change Through Cognitive Processes.” In *Teachers’ Professional Development*, by S Krolak-Schwerdt, 27–43. Rotterdam, the Netherlands: Sense Publishers.
- Pit-Ten Cate, I, S Krolak-Schwerdt, T Horstmann, and S Glock. 2013. “Better Decisions Through Science - Changing Decision Making Processes by Applying Formal Decision Rules.” Munich, Germany.
- PORaad. 2018. “Overgang PO-VO: Sectoraal Standpunt.”
- Praetorius, A K, T Koch, A Scheunpflug, H Zeinz, and M Dresel. 2017. “Identifying Determinants of Teachers’ Judgment (in)accuracy Regarding Students’ School-Related Motivations Using a Bayesian Cross-Classified Multi-Level Model.” *Learning and Instruction* 52: 148–60.
- Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedagogiske Institut.
- Remie, M, and M Huygen. 2019. “Een Leven Lang Getoetst. Waarom Eigenlijk?” *NRC*,

January.

Resing, W C M, and J B Blok. 2002. "De Classificatie van Intelligentiescores: Voorstel Voor Een Eenduidig Systeem." *De Psycholoog* 37: 244–49.

Rose, C, and M D Smith. 2002. *Mathematical Statistics with Mathematica*. New York: Springer-Verlag.

Rosenthal, R, and L Jacobson. 1968. "Pygmalion in the Classroom." *The Urban Review* 3 (1): 16–20.

Rubie-Davies, C M, and E R Peterson. 2016. "Relations Between Teachers' Achievement, over- and Underestimation, and Students' Beliefs for Maori and Pakeha Students." *Contemporary Educational Psychology* 47: 72–83.

Ruckdeschel, P, M Kohl, T Stabla, and F Camphausen. 2006. "S4 Classes for Distributions." *R News* 6 (2): 2–6. <https://CRAN.R-project.org/doc/Rnews/>.

Rutkowski, D, and L Rutkowski. 2013. "Measuring Socioeconomic Background in PISA: One Size Might Not Fit All." *Research in Comparative and International Education* 8 (3): 259–78.

Sanders, P F, and A J Verschoor. 1996. "Parallel Test Construction Using Classical Item Parameters." *Applied Psychological Measurement* 22 (3): 212–23.

Santelices, M V, and M Wilson. 2010. "Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning." *Harvard Educational Review* 80 (1): 106–33.

Scheerens, J, and G Van Der Werf. 2018. "Immigrant Student Achievement and Education Policy in the Netherlands." In *Immigrant Student Achievement and Education Policy: Cross-Cultural Approaches*, by L Volante, D Klinger, and O Bilgili. Switzerland: Springer.

Schenke, W, and J Meijer. 2018. "Datagebruik in Het Onderwijs: Problematiek Uiteengezet." 40782. Amsterdam: Kohnstamm Instituut.

Schildkamp, K, and C L Poortman. 2015. "Factors Influencing the Functioning of Data Teams." *Teachers College Record* 117 (4): n4.

Schildkamp, K, A Handelzalts, C L Poortman, H Leusink, M Meerdink, M Smit, and M D Hubers. 2018. *The Data Team Procedure: A Systematic Approach to School Improvement*. Springer texts in education.

Schildkamp, K, C L Poortman, H Luyten, and J Ebbeler. 2017. "Factors Promoting and Hindering Data-Based Decision Making in Schools." *School Effectiveness and School Improvement* 28 (2): 242–58.

Schouws, S. 2015. "Zin En Onzin van Het Meten van Intelligentie." *Psychopraktijk* 3: 34–36.

Shealy, R, and W Stout. 1993. "A Model-Based Standardization Approach That Separates True Bias/DIF from Group Ability Differences and Detects Test Bias/DIF as Well as Item Bias/DIF." *Psychometrika* 58 (2): 159–94.

Siegel, J, E A Gilmore, N Gallagher, and S Fetter. 2017. "An Expert Elicitation of the Proliferation Resistance of Using Small Modular Reactors (SMR) for the Expansion of

- Civilian Nuclear Systems.” *Risk Analysis* 38 (2): 242–54.
- Sijtsma, K. 2009. “On the Use, the Misuse, and the Very Limited Usefulness of Cronbach’s Alpha.” *Psychometrika* 74: 107–20.
- . 2012. “Future of Psychometrics: Ask What Psychometrics Can Do for Psychology” 77 (1): 4–20.
- Singh, G, J Sinner, J Ellis, M Kandlikar, and B Halpern. 2017. “Mechanisms and Risk of Cumulative Impacts to Coastal Ecosystem Services: An Expert Elicitation Approach.” *Journal of Environmental Management* 199: 229–41.
- Smeets, E, J Van Kuijk, and G Driessen. 2014. “Handreiking Bij Het Opstellen van Het Basisschooladvies.” 34001792. ITS, Radboud Universiteit Nijmegen.
- Smith, William C. 2014. “The Global Transformation Toward Testing for Accountability.” *Education Policy Analysis Archives* 22 (116): 1–34.
- Sneyers, E, J Vanhoof, and P Mahieu. 2018. “Primary Teachers’ Perceptions That Impact Upon Track Recommendations Regarding Pupils’ Enrolment in Secondary Education: A Path Analysis.” *Social Psychology of Education* 21: 1153–73.
- Spearman, C. 1904. “The Proof and Measurement of Association Between Two Things.” *American Journal of Psychology* 15: 72–101.
- . 1910. “Correlation Calculated from Faulty Data.” *British Journal of Psychology* 3: 271–95.
- Speirs-Bridge, A, F Fidler, M McBride, L Flander, G Cumming, and M Brugman. 2010. “Reducing Overconfidence in the Interval Judgments of Experts.” *Risk Analysis* 30 (3): 512–23.
- Staatssecretaris van Onderwijs, Cultuur en Wetenschap. 2014. “Toetsbesluit PO.”
- . 2015a. “Eerste Inzichten Wet Eindtoetsing PO.”
- . 2015b. “Overgang van Primair Naar Voortgezet Onderwijs.”
- . 2016. “Inzichten Inzake de Wet Eindtoetsing Primair Onderwijs.”
- . 2017. “Tussenevaluatie Wet Eindtoetsing PO.”
- Steenkamp, J B, and H Baumgartner. 1998. “Assessing Measurement Invariance in Cross-National Consumer Research.” *Journal of Consumer Research* 25 (1): 78–90.
- Steiger, J H, and J M Lind. 1980. “Statistically Based Tests for the Number of Common Factors.” Iowa City, IA.
- Südkamp, A, J Kaiser, and J Möller. 2012. “Accuracy of Teachers’ Judgments of Students’ Academic Achievement: A Meta-Analysis.” *Journal of Educational Psychology* 104 (3): 743–62.
- . 2014. “Teachers’ Judgments of Students’ Academic Achievement.” In *Teachers’ Professional Development: The Future of Education Research*. Rotterdam, the Netherlands: Sense Publishers.
- Team, R Core. 2017. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- team, Rstudio. 2016. *Rstudio: Integrated Development for R*. Boston, MA, USA: RStudio

Inc. <http://www.rstudio.com>.

Tellegen, P. 2002. “De Handleiding van de WISC-III NL: Correcties, Opmerkingen En Suggesties.” <http://www.testresearch.nl/wisc/wiscopm.html>.

———. 2004. “De Waan van ‘Het’ IQ.” <http://www.testresearch.nl/wisc/wiscopm.html>.

Thorndike, R L. 1951. “Reliability.” In *Educational Measurement*, by E F Lindquist, 560–620. Washington, DC: American Council of Education.

Thys, S, and M Van Houtte. 2016. “Ethnic Composition of the Primary School and Educational Choice: Does the Culture of Teacher Expectations Matter?” *Teaching and Teacher Education* 59: 383–91.

Tieben, N. 2009. “Parental Resources and Relative Risk Aversion in Intra-Secondary Transitions in the Netherlands: A Trend Analysis of Non-Standard Educational Decision Situations.” *European Sociological Review*, 1–12.

Tieben, N, P M De Graaf, and N D De Graaf. 2010. “Changing Effects of Family Background on Transitions to Secondary Education in the Netherlands: Consequences of Educational Expansion and Reform.” *Research in Social Stratification and Mobility* 28: 77–90.

Timmermans, A C, H De Boer, H T A Amsing, and M P C Van Der Werf. 2018. “Track Recommendation Bias: Gender, Migration Background and SES Bias over a 20-Year Period in the Dutch Context.” *British Educational Research Journal* 44 (5): 847–74.

Timmermans, A C, H De Boer, and M P C Van Der Werf. 2016. “An Investigation of the Relationship Between Teachers’ Expectations and Teachers’ Perceptions of Student Attributes.” *Social Psychology Education* 19: 217.

Timmermans, A C, H Kuyper, and G Van Der Werf. 2013. “Schooladviezen En Onderwijsloopbanen: Voorkomen, Risicofactoren En Gevolgen van Onder- En Overadviesing.” Gronings Instituut voor Onderzoek van het Onderwijs (GION).

———. 2015. “Accurate, Inaccurate, or Biased Teacher Expectations: Do Dutch Teachers Differ in Their Expectations at the End of Primary Education?” *British Journal of Educational Psychology* 85 (4): 459–78.

Tolsma, J, M Coenders, and M Lubbers. 2007. “Trends in Ethnic Educational Inequalities in the Netherlands: A Cohort Design.” *European Sociological Review* 23 (3): 325–39.

Traub, R. 1997. “Classical Test Theory in Historical Perspective.” *Educational Measurement* 16: 8–14.

Tversky, A, and D Kahneman. 1974. “Judgment Under Uncertainty: Heuristics and Biases.” *Science* 185 (4157): 1124–31.

Van Aarsen, E. 2013. “Voorspellende Waarde van LOVS Toetsen Voor Schoolsucces.” Utrecht: Oberon.

Van Buuren, S, and K Groothuis-Oudshoorn. 2011. “Mice: Multivariate Imputation by Chained Equations in R.” *Journal of Statistical Software* 45 (3): 1–67.

Van De Schoot, R, and S Depaoli. 2014. “Bayesian Analyses: Where to Start and What to Report.” *European Health Psychologist* 16 (2): 75–84.

Van De Schoot, R, D Kaplan, J Denissen, J B Asendorpf, F J Neyer, and M A Van Aken.

2014. "A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research." *Child Development* 85 (3): 842–60.
- Van De Schoot, R, A Kluytmans, L Tummers, P Lugtig, J Hox, and B Muthen. 2013. "Facing Off with Scylla and Charybdis: A Comparison of Scalar, Partial and the Novel Possibility of Approximate Measurement Invariance." *Frontiers in Psychology* 4: 770.
- Van De Schoot, R, P Schmidt, A De Beuckelaer, K Lek, and M Zondervan-Zwijnenburg. 2015. "Editorial: Measurement Invariance." *Frontiers in Psychology* 6: 1064.
- Van De Vijver, F. et al. 2019. "Invariance Analyses in Large-Scale Studies." *OECD Publishing, Paris*, OECD education working papers,, no. 201.
- Van De Werfhorst, H, and J J B Mijs. 2010. "Achievement Inequality and the Institutional Structure of Educational Systems: A Comparative Perspective." *Annual Review of Sociology* 36: 407–28.
- Van De Werfhorst, H, L Elfers, and S Karsten. 2015. "Onderwijsstelsels Vergeleken: Leren, Werken En Burgerschap."
- Van Den Bergh, L, E Denessen, L Hornstra, M Voeten, and R W Holland. 2010. "The Implicit Prejudiced Attitudes of Teachers: Relations to Teacher Expectations and the Ethnic Achievement Gap." *American Educational Research Journal* 47: 497–527.
- Van Den Bosch, R M, C A Espin, S Chung, and N Saab. 2017. "Data-Based Decision Making: Teachers' Comprehension of Curriculum-Based Measurement Progress-Monitoring Graphs." *Learning Disabilities Research and Practice* 32 (1): 46–60.
- Van Der Linden, W J. 1979. "Binomial Test Models and Item Difficulty." *Applied Psychological Measurement* 3: 401–11.
- Van Erp, S, J Mulder, and D L Oberski. 2018. "Prior Sensitivity in Default Bayesian Structural Equation Modeling." *Psychological Methods* 23 (2): 363–88.
- Van Geel, M, and A Visscher. 2013. "Opbrengstgericht Werken Met Het Digitale Leerlingvolgsysteem." *Weten Wat Werkt En Waarom; Wetenschappelijk Tijdschrift over Opbrengsten En Werking van Ict in Het Onderwijs* 2 (1): 22–29.
- Van Ravenzwaaij, D, and R Hamel. 2006. "De Nederlandstalige WAIS-III Na Hernormering." *De Psycholoog*, 268–71.
- Van Rens, M, C Haelermans, W Groot, H Maassen, and H Van Den Brink. 2018. "Facilitating a Successful Transition to Secondary School: (How) Does It Work? A Systematic Literature Review." *Adolescent Research Review* 3: 43–56.
- Van Rooijen, M, H Korpershoek, J Vugteveen, and M C Opdenakker. 2017. "Transition from Primary to Secondary Education and the Continuing School Career." *Pedagogische Studiën* 94 (2): 110–34.
- Van Spijker, F, K Van der Houwer, and R Van Gaalen. 2017. "Invloed Ouderlijk Opleidingsniveau Reikt Tot Ver in Het Voortgezet Onderwijs." *ESB Onderwijs En Wetenschap* 102: 234–36.
- Van Strien, Pieter J, and Willem K B Hofstee. 1995. "An Interview with Adriaan d. de Groot." *New Ideas in Psychology* 13 (3): 341–56.
- Vandenberg, R J. 2002. "Toward a Further Understanding of and Improvement in Measurement Invariance Methods and Procedures." *Organizational Research Methods*

5 (2): 139–58.

Vandenberg, R J, and C E Lance. 2000. “A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research.” *Organizational Research Methods* 3 (1): 4–70.

Vanlommel, K, and K Schildkamp. 2018. “How Do Teachers Make Sense of Data in the Context of High-Stakes Decision Making?” *American Educational Research Journal*.

Vanlommel, K, R Van Gasse, J Vanhoof, and P Van Petegem. 2018. “Teachers’ High-Stakes Decision Making. How Teaching Approaches Affect Rational and Intuitive Data Collection.” *Teaching and Teacher Education* 71: 108–19.

Veen, D, D Stoel, N Schalken, K Mulder, and R Van De Schoot. 2018. “Using the Data Agreement Criterion to Rank Experts’ Beliefs.” *Entropy* 20: 1–17.

Verger, Antoni, Clara Fontdevilla, and Lluís Parcerisa. 2019. “Reforming Governance Through Policy Instruments: How and to What Extent Standards, Tests and Accountability in Education Spread Worldwide” 40 (2): 248–70.

Wagenmakers, E-J, R D Morey, and M D Lee. 2016. “Bayesian Benefits for the Pragmatic Researcher.” *Current Directions in Psychological Science* 25 (3): 169–76.

Wahl, D, F E Weinert, and G L Huber. 2007. *Psychologie Für Die Schulpraxis*. Belm-Vehrte: Sozio-Publishing.

Wainer, H. 2000. “Kelley’s Paradox.” *Chance* 13: 47–48.

Walther, A B, and J L Moore. 2005. “The Concepts of Bias, Precision and Accuracy, and Their Use in Testing the Performance of Species Richness Estimators, with a Literature Review of Estimator Performance.” *Ecography* 28: 815–29.

Wang, S, C M Rubie-Davies, and K Meissel. 2018. “A Systematic Review of the Teacher Expectation Literature over the Past 30 Years.” *Educational Research and Evaluation* 24 (3): 124–79.

Wang, T, M J Kolen, and D J Harris. 1996. “Conditional Standard Errors, Reliability and Decision Consistency of Performance Levels Using Polytomous IRT.” New York.

Wang, Z, and S J Osterlind. 2013. “Classical Test Theory.” In *Handbook of Quantitative Methods for Educational Research*, by T Theo, 31–44. Sense Publishers.

Winkler, R L. 1967. “The Assessment of Prior Distributions in Bayesian Analysis” 62: 776–880.

Wolvers, R J, and P Lugtig. 2016. “A Comparison of Four Methods for Testing Measurement Invariance Across Many Groups.” Masterthesis, Utrecht, the Netherlands: Utrecht University.

Woodruff, D J. 1990. “Conditional Standard Error of Measurement in Prediction.” *Journal of Educational Measurement*, no. 27: 191–208.

Zachary, R A, and R L Gorsuch. 1985. “Continuous Norming: Implications for the WAIS-R.” *Journal of Clinical Psychology* 41 (1): 86–94.

Zeedijk, M S, J Gallacher, M Henderson, G Hope, B Husband, and K Lindsay. 2003. “Negotiating the Transition from Primary to Secondary School. Perceptions of Pupils,

Parents and Teachers.” *School Psychology International* 24 (1): 67–79.

Zercher, F, P Schmidt, J Ciecuch, and E Davidov. 2015. “The Comparability of the Universalism Value over Time and Across Countries in the European Social Survey: Exact Versus Approximate Measurement Invariance.” *Frontiers in Psychology* 6: 207–17.

Zondervan-Zwijnenburg, M, W Van De Schoot-Hubeek, K Lek, H Hoijtink, and R Van De Schoot. 2017. “Application and Evaluation of an Expert Judgement Elicitation Procedure for Correlations.” *Frontiers in Psychology* 8 (90): 1–15.