Item-score reliability

Zijlmans, Eva A.O.

Publication date:
2019

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

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Zijlmans, E. A. O. (2019). *Item-score reliability: Estimation and evaluation*. Gildeprint.

ITEM-SCORE RELIABILITY

ESTIMATION AND EVALUATION

EVA A.O. ZIJLMANS

ITEM-SCORE RELIABILITY

ESTIMATION AND EVALUATION

EVA A.O. ZIJLMANS

Colophon

Printing was financially supported by Tilburg University.

Printed by:	Gildeprint, Enschede, The Netherlands
Cover design:	Rachel van Esschoten, DivingDuck Design - www.divingduckdesign.nl
ISBN:	978-94-6323-482-5

Item-Score Reliability

Estimation and Evaluation

Proefschrift ter verkrijging van de
graad van doctor aan Tilburg University
op gezag van de rector magnificus,
prof. dr. E. H. L. Aarts,
in het openbaar te verdedigen
ten overstaan van een door het college
voor promoties aangewezen commissie
in de Aula van de Universiteit op
vrijdag 15 februari 2019 om 13:30 uur

door

Eva Adriana Oda Zijlmans

geboren op 15 november 1990
te Raamsdonk

Promotores:	Prof. dr. K. Sijtsma
	Prof. dr. L. A. van der Ark
Copromotor:	Dr. J. Tijmstra
Promotiecommissie:	Prof. dr. R. R. Meijer
	Prof. dr. B. P. Veldkamp
	Prof. dr. J. M. Wicherts
	Dr. S. Bouwmeester
	Dr. L. G. M. T. Keijsers

Table of Contents

1 Introduction **1**

1.1 Test-Score Reliability 2

1.2 Item-Score Reliability 2

1.3 Item-Score Reliability in the Literature 3

1.4 Usability 7

1.5 Item-Score Reliability Methods 7

1.6 Outline of the Dissertation 8

2 Methods for Estimating Item-Score Reliability **11**

2.1 Introduction 12

2.2 A Framework for Item-Score Reliability 13

2.3 Methods for Approximating Item-Score Reliability 15

2.4 Simulation Study 20

2.5 Real-Data Example 25

2.6 Discussion 26

3 Item-Score Reliability in Empirical-Data Sets and Its Relationship With Other Item Indices **29**

3.1 Introduction 30

3.2 Method 31

3.3 Results 40

3.4 Discussion 45

4 Investigating the Relationship between Item-Score Reliability, its Estimation Methods, and Other Item Indices **51**

4.1 Introduction 52

4.2 Method 53

4.3 Results 57

4.4 Discussion 62

5 Item-Score Reliability as a Selection Tool in Test Construction **65**

5.1 Introduction 66

5.2 Item Selection in Test Construction 67
5.3 Simulation Study 73
5.4 Discussion 80

6 Epilogue **83**

Appendix A Item-Score Reliability Methods Based on Cronbach's α and Guttman's λ_2 **89**

Appendix B Deriving the Item-Score Reliability and Test-Score Reliability from the Two-Parameter Logistic Model **93**

References **95**

Summary **105**

Acknowledgements **107**

Chapter 1

Introduction

Measuring attributes of people, objects, or systems is an important aspect of scientific research. In the physical sciences, attributes are sometimes observable, making measurement easy. An example is length, which can be simply measured using a measuring tape. However, also in physics and other scientific areas as well, attributes are often unobservable, or only observable through other phenomena that are related to the attributes but do not coincide with them. One may think of radioactivity, electrical current, and hardness of materials. For example, one may feel a weak electrical current as tickling to the skin, but that feeling is not identical to a measurement procedure and the attribute of electrical current itself remains unobservable. Measurement in the social sciences is difficult for the same reason, because the attributes of interest often are unobservable and the only things noticeable are symptoms of attributes. Measurements of a person's intelligence, extraversion, or neuroticism cannot be obtained by simply reading off a scale that is readily available, thereby introducing a challenge for measurement in the social sciences.

Frequently used instruments to measure unobservable psychological attributes are tests or questionnaires. The manifestations resulting from the attribute of interest are recorded by means of different so-called items. Scores on various items are combined to obtain a quantitative measurement of the attribute. It is of great importance that the tests and questionnaires employed are of high quality, because this will benefit the quality of the measurements.

The research area concerned with the theory and methods of measurement in the social sciences is psychometrics. Psychometricians work on improving the quality of the measurements obtained by means of tests and questionnaires by improving the measurement instrument and the statistical methods used to analyze the obtained measurements. An important aspect of the quality of a test is test-score reliability.

1.1 Test-Score Reliability

The test-score reliability indicates the repeatability of the test score. This means that when a test is administered to a person, and we would be able to erase this person's memory and let him or her retake the test, a reliable test would result in the same score or almost the same score for this person. Researchers commonly investigate the reliability of the test score, denoted $\rho_{XX'}$, which stands for the product-moment correlation between two independent administrations of the same test in the same group of people. Of course, researchers will not be able to erase the memory of their subjects, and therefore psychometricians have developed methods to approximate test-score reliability, often based on just one administration of the test. The most commonly used test-score reliability method is coefficient alpha (Cronbach, 1951), which is a lower bound to the test-score reliability. Even though other lower bounds to the test-score reliability are available that better approximate test-score reliability (Sijtsma, 2009), in practice coefficient alpha is the most used method for test-score reliability. However, it is not very common to investigate the reliability at the level of the item score. Even though psychometric, psychological, marketing, management and human resource studies, and other research areas have given attention to the reliability of individual items, a thorough investigation of this subject has not been carried out before. This thesis fills the gap and studies methods for approximating and estimating the reliability of individual item scores, and discusses the practical need for such estimates.

1.2 Item-Score Reliability

Because the reliability of a test score is an important aspect of test quality, assessing the reliability of the individual item scores could also be interesting and is worthy of further investigation. Currently, various item indices assessing aspects of item quality other than reliability are available. Researchers investigate aspects of item quality by means of the corrected item-total correlation, the item-factor loading, the item scalability, or the item discrimination. Even though these indices all measure some aspect of item quality, none explicitly touch upon the repeatability of an item score. However, there are some studies available that mention or study item-score reliability or apply some method to estimate item-score reliability in the analysis of real test data.

1.3 Item-Score Reliability in the Literature

Several studies have touched upon the concept of item-score reliability, but a thorough investigation of its use, applications and methods had not been executed yet. In a literature review, we attempted to gather the research and information available in the current literature about item reliability, so as to create a concise overview of what has been known so far with respect to item-score reliability. This literature review identified the gaps with respect to item-score reliability so that these gaps can be further investigated. Even though we attempted to be thorough, it might be the case that not all methods ever mentioned in the literature are part of this review.

Guttman (1946) described a universe where indefinitely many trials of a person responding to an item take place, which leads to the definition of the reliability coefficient of the item. He argued that an item is unreliable for a person to the extent to which his or her response varies across repeated experiments under the same conditions. Because repeated experiments under the same conditions are difficult, if not impossible to realize, this specific reliability coefficient of the item cannot be computed in practice. In Guttman's study, definitions appropriate for test-retest reliability for items at both the level of an individual responding to an item and the population responding to an item are described. The derivations presented in his study can be used in practice to compute a lower and an upper bound to the population reliability coefficient from a single trial of a large population. However, it is not possible to compute a point estimate of the reliability of an item using the derivations presented by Guttman, which is the reason we did not consider his definitions in this study.

The purpose of the study by Knapp (1977) was to develop from first principles the concept of the reliability of a dichotomous one-item cognitive test, and to suggest procedures for estimating reliability for such one-item tests. Knapp (1977) investigated an approach without using correlations towards estimating the reliability of a dichotomous item based on the proportion of individuals that know the correct answer and answer correctly, and the proportion of individuals that do not know the answer and answer incorrectly. These proportions are defined as reliable scores. In practice, there is only one option available for estimating these proportions and that is the "test-retest" technique, which involves the administration of the item on two or more occasions to members of the population of interest. There are two reasons we did not consider this method in our study. First, respondents will remember what they answered during the first administration of the test, and they might change over time. Second, because data often does not have a second measurement at a different time point, which is necessary for the "test-retest" technique, this method cannot always be applied.

For the evaluation of item quality during test construction using nonparametric item response theory methods (Mokken, 1971), Meijer, Sijtsma, and Molenaar (1995) studied the estimation of the reliability of a single dichotomous item score. They discussed three methods for the estimation of item-score reliability for dichotomous items, each based on the assumptions of nondecreasing and nonintersecting item response functions (e.g., Embretson & Reise, 2000). By means of analytical and Monte Carlo studies, Meijer et al. (1995) found one method to be superior over the other two, because it had smaller bias and smaller sampling variance. The methods investigated in this study were studied only for dichotomous items, and therefore needed some adjustments before they could also be used for the estimation of item-score reliability for polytomous items. The adjusted method was considered in this study.

Wanous and Reichers (1996) described how to estimate the reliability of a single item score using an adapted version of the correction for attenuation (Lord & Novick, 1968; Nunnally & Bernstein, 1994; Spearman, 1904), from now on referred to as method CA. Their focus was on measuring the reliability of a single-item measure for job satisfaction. In a follow-up study, Wanous, Reichers, and Hudy (1997) conducted a meta-analysis of single-item measures of overall job satisfaction, thereby investigating how good single-item measures are for measuring an allegedly *simple* construct like job satisfaction. The reliability of the single-item measure for job satisfaction was again investigated by means of method CA. These authors concluded that, depending on the assumptions made, the item-score reliability for the single-item measure for job satisfaction was between $.45$ and $.69$, and it was therefore deemed acceptable to use this single-item measure. The second method proposed by Wanous and Hudy (2001) to estimate the item-score reliability of a single-item measure is based on the factor model, originally proposed by Weiss (1976, pp. 351–352) and first used in a monograph by Arvey, Landon, Nutting, and Maxwell (1992, p. 1000). Harman (1976, pp. 18–19) described how the variance of a variable can be split into the communality, the specificity, and the error or unreliability. The complement of the error variance can be seen as the true-score variance, equal to the sum of the communality and specificity, and relevant to the reliability of the variable. Communality can thus be regarded as a conservative estimate of item true-score variance, and the estimate of communality can be obtained by means of a factor model. This factor analytic method was not considered in this study, because we focused on the classical definition of reliability, where the variance can be split in true score variance and error score variance, whereas in the factorial analytic method the variance is split into three components.

Fuchs and Diamantopoulos (2009) provided researchers in the field of management studies with concrete guidelines for using single-item measures. With regard to the reliability of a single-item measure, they concluded that one should not reject single-item measures because of concerns about how to estimate their reliability, because adequate methods to estimate item-score reliability do exist, or reject the value of the resulting estimate, because these are often within acceptable levels. Next to method CA developed by Wanous and Reichers (1996), Fuchs and Diamantopoulos (2009) proposed applying the Spearman-Brown formula in reverse (see Nunnally & Bernstein, 1994, pp. 263–264, for the Spearman-Brown formula), meaning that one solves the theoretical reliability of one item from knowledge of the test-score reliability and the number of items. The authors described that on the one hand Spector (1992, p. 4) argued that “yes” or “no” single-item measures are *notoriously* unreliable, because the responses are not consistent over time; thus respondents may answer differently the next day. Also, Churchill (1979) described how the reliability tends to increase and measurement error decreases as the number of items in a combination increases. On the other hand, Drolet and Morrison (2001, p. 200) argued that adding more items to the measure results in minimal extra information compared to the single-item measure, and is therefore perhaps not worth the effort. Because of the aforementioned reasons, Fuchs and Diamantopoulos (2009) concluded that for ability tests a single item cannot provide a reliable estimate of the individual’s ability (Rossiter, 2002, p. 321), but for business-related constructs a good single-item measure instead of a multi-item measure will not change theoretical tests and empirical findings (Bergkvist & Rossiter, 2007).

In the literature, there are several examples of applied research where item-score reliability is investigated. Russell, Weiss, and Mendelsohn (1989) introduced a single-item scale, the Affect Grid, to quickly assess affect along the dimensions of pleasure-displeasure and arousal-sleepiness. The authors used the fact that reliability sets an upper bound on validity, and conversely claimed that and index of convergent validity estimates a lower bound on reliability. By means of this method they obtained an item-score reliability that ranged from $.74$ to $.94$ for the pleasure dimension, and from $.63$ to $.92$ for the arousal dimension. They concluded that the average subject therefore yielded scores sufficiently reliable to be useful. Robins, Hendin, and Trzesniewski (2001) investigated the reliability of a single-item self-esteem scale (SISE) measure. They used the approach by Heise (1969, Equation 9), that provides an estimate of test-retest reliability. The method is based on the pattern of autocorrelations of the single-item measure over three points in time. They computed the Heise estimate three times using measurements of the SISE from the beginning to the end of college. They found a mean reliability estimate for the SISE of $.75$.

Wanous and Hudy (2001) estimated the reliability of a College Teaching Effectiveness single-item measure using both the factor analytic method and method CA. They found a higher item-score reliability value for the factor analytic method ($.88$) than for method CA ($.64$). Ginns and Barrie (2004) continued this research, by investigating single-item ratings of the quality of instructors or subjects, used by higher education institutions. They also used both the factor analytic method and method CA and found item-score reliability values of $.94$ and $.96$, respectively. Shamir and Kark (2004) offered a single-item measure for identification with organizations and organizational units. Item-score reliability was assessed by means of test-retest correlations over a period of two weeks, which resulted in values of $.73$ and $.80$ in the different samples in their study. They assessed these values to provide evidence for reliability of the single-item measure. Dolbier, Webster, McCalister, Mallon, and Steinhardt (2005) investigated, following Wanous et al. (1997), the reliability of a single-item measure measuring job satisfaction. They used method CA and concluded that the item-score reliability was $.73$ when the correlation between the single- and the multiple-item job satisfaction measures was assumed to be perfect, and $.90$ when the correlation was assumed to be more conservative. Zimmerman et al. (2006) developed single-item measures for three domains that are important to consider when treating depressed patients. For two of those measures, they used intraclass correlation coefficients for test-retest reliability to determine the item-score reliability. For the psychosocial functioning measure they found a value of $.76$, and for the quality of life measure, they found a value of $.81$. They concluded that both values were high. Postmes, Haslam, and Jans (2012) introduced a single-item social identification measure assessing (SISI) that assessed the respondent's identification with his or her group or category, on a 7-point scale. They investigated the reliability of this single-item measure by means of method CA and test-retest reliability, and found values of $.76$, and $.64$, respectively. Based on other studies that report the reliability of single-item measures, Postmes et al. (2012) conclude that the reliability of their SISI measure exceeds most other reliabilities of single-item measures, which strengthens their confidence in the robustness of this scale for use in psychological research. Melián-González, Bulchand-Gidumal, and López-Valcárcel (2015) investigated the relationship between employee satisfaction and organizational performance, the former being assessed by a single-item measure. Using the factor analytical method, they obtained a communality of 0.95, and classified this as the item-score reliability of this measure. Williams, Thomas, and Smith (2017) used the Well-Being Process Questionnaire (WPQ) to investigate stress levels and the well-being of university staff. Reliability of this measure was assessed by means of method CA by

Wanous et al. (1997), with a value of $.90$ as the assumed true correlation between the single-item and the multi-item measure. They found an item-score reliability value of $.64$ and therefore concluded that multi-item measures would be more suitable for research purposes, because they provide more consistently high reliability scores. Gignac and Wong (2018) investigated the item-score reliability of a single-anagram version of the Anagram Persistence Task (APT) via its communality within the relevant factor analytic solution. The result of their analysis showed an item-score reliability value of $.42$ for the single-anagram version of the APT, which was assessed as unacceptably low for researcher purposes.

The applied research examples described above indicate that the method by Wanous et al. (1997) is used frequently, as is test-retest reliability. Also, the studies by Wanous and Reichers (1996), Wanous et al. (1997), and Wanous and Hudy (2001) are often cited (Google Scholar cited Wanous et al. (1997) 2400+ times (November 16 2017)) to motivate that single-item measures are, given certain conditions, reliable measurement instruments. Other applications of item-score reliability, besides single-item measures, are not very common yet, which is a reason to further explore its usability.

1.4 Usability

Currently, item-score reliability is mainly used for motivating that a single-item measure, for example, measuring job satisfaction, is a reliable measure (Wanous & Reichers, 1996; Wanous et al., 1997), but there are many more situations in which item-score reliability can be a useful tool. Having information about the reliability of a single item gives researchers the opportunity to identify unreliable item scores and remove these items from the test in order to obtain a test of higher quality. Other possibilities that arise are selecting the most reliable item from a test to use as a single-item measure. Also, one could use the item-score reliability in test construction as a selection tool to decide which items to add or omit from a test to increase its quality and obtain better measurement instruments.

1.5 Item-Score Reliability Methods

Considering what has been investigated with respect to item-score reliability, at the start of this research, we found two methods available to estimate item-score reliability. The first method is the method proposed by Wanous and Reichers (1996), which we called method CA. The second method is the method proposed by Meijer et al. (1995), but this method is defined for dichotomous items only. The availability of only two methods, of which the latter method cannot be used for polytomous items, shows a need for new, perhaps better, methods. Also, the performance of method CA has not been investigated with regard to bias and accuracy.

In this dissertation, the starting point for the development of methods for determining item-score reliability has been considering the methods used for the determination of test-score reliability. The latter methods were adjusted such that they fit in a general framework for estimating item-score reliability, thereby making the differences clear between how the different methods approximate item-score reliability.

The first method that was adjusted for estimating the reliability of an item score instead of a test score is the Molenaar-Sijtsma method, already employed by Meijer et al. (1995) for estimating the reliability of single dichotomous items. Molenaar and Sijtsma (1988; also Sijtsma & Molenaar, 1987; Van der Ark, 2010) proposed method MS to estimate the reliability of a test-score using a single-administration. The theoretical basis of this method was used to develop a new method for estimating item-score reliability, called method MS. The second method was based on coefficient λ_6, developed by Guttman (1945), which led to method λ_6 for estimating item-score reliability. Finally, the latent class reliability coefficient (LCRC) proposed by Van der Ark, Van der Palm, and Sijtsma (2011) was adjusted such that it estimates item-score reliability, and defined as method LCRC. The existing method CA was also considered as an item-score reliability estimation method.

1.6 Outline of the Dissertation

This dissertation deals with item-score reliability and the development, evaluation, and usability of item-score reliability methods. An existing method for estimating item-score reliability was reviewed and compared to newly-developed methods to assess their performance. Simulation studies and empirical data sets were deployed, to investigate bias and precision of the different item-score reliability methods and to identify values of item-score reliability that can be expected in practice, respectively. The item-score reliability methods were compared to other item indices assessing different features of item performance that are often used in practice. Also, the usability of the item-score reliability methods was evaluated by means of a simulation study focusing on item selection in test construction.

The chapters were written as separate journal articles and can be read independently from each other. This means that some technical details, especially regarding the item-score reliability estimation methods, return in the different chapters, creating some overlap.

In Chapter 2, methods to estimate item-score reliability are explored, resulting in the development of three new methods: method MS, method λ_6, and method LCRC. All three methods fit in the same framework, based on approximating the correlation between two independent replications of the same item. The fourth method that was investigated in addition to developed methods was method CA, which was readily available, and already introduced. By means of a simulation study, the median bias, the variability (quantified as the inter-quartile range), and the percentage of outliers of the four item-score reliability methods were investigated and compared.

Chapter 3 contains an analysis of several empirical-data sets by means of the most promising item-score reliability methods identified in Chapter 2. Four other item indices assessing item features different from item-score reliability were also applied to these empirical-data sets. By means of this research, the values that can be expected in empirical data-sets were empirically identified, as well as the relationship between the item-score reliability estimation methods and the four other item indices.

For Chapter 4, the relationship between item-score reliability and the three item-score reliability methods was further investigated by means of a simulation study. In this study, the bias of the three item-score reliability methods was assessed in several realistic research conditions for a range of item-score reliability values. Also, for the same conditions and the same range, the relationship between item-score reliability and four other item indices not assessing the item-score reliability was investigated.

The usability of item-score reliability as an item selection method in test construction was investigated in Chapter 5. The goals were to use item-score reliability methods as a measure to decide which item to add to the test or to omit from the test, based on a high item-score reliability, and a low item-score reliability, respectively. The objective was to maximize the test-score reliability. Because in practice the corrected item-total correlation is already being used for item selection, this measure was also investigated in this study and used as a benchmark method to compare to the novel item-score reliability methods.

In the Epilogue, the added value of this thesis in the field of psychometrics was evaluated, and a direction for future research was discussed.

Chapter 2

Methods for Estimating Item-Score Reliability

Abstract

Reliability is usually estimated for a test score, but it can also be estimated for item scores. Item-score reliability can be useful to assess the item's contribution to the test score's reliability, for identifying unreliable scores in aberrant item-score patterns in person-fit analysis, and for selecting the most reliable item from a test to use as a single-item measure. Four methods were discussed for estimating item-score reliability: the Molenaar-Sijtsma method (method MS), Guttman's method λ_6, the latent class reliability coefficient (method LCRC), and the correction for attenuation (method CA). A simulation study was used to compare the methods with respect to median bias, variability (interquartile range [IQR]), and percentage of outliers. The simulation study consisted of six conditions: standard, polytomous items, unequal α-parameters, two-dimensional data, long test, and small sample size. Methods MS and CA were the most accurate. Method LCRC showed almost unbiased results, but large variability. Method λ_6 consistently underestimated item-score reliability, but showed a smaller IQR than the other methods.

Keywords: correction for attenuation, Guttman's method λ_6, item-score reliability, latent class reliability coefficient, method MS

Based on Zijlmans, E. A. O., Tijmstra, J., Van der Ark, L. A., and Sijtsma, K. (2018). Methods for Estimating Item-Score Reliability. *Applied Psychological Measurement, 42*, 553–570. doi: 10.1177/0146621618758290

2.1 Introduction

Reliability of measurement is often considered for test scores, but some authors have argued that it may be useful to also consider the reliability of individual items (Ginns & Barrie, 2004; Meijer & Sijtsma, 1995; Meijer et al., 1995; Wanous & Reichers, 1996; Wanous et al., 1997). Just as test-score reliability expresses the repeatability of test scores in a group of people keeping administration conditions equal (Lord & Novick, 1968, p. 65), item-score reliability expresses the repeatability of an item score. Items having low reliability are candidates for removal from the test. Item-score reliability may be useful in person-fit analysis to identify item scores that contain too little reliable information to explain person fit (Meijer & Sijtsma, 1995). Meijer, Molenaar, and Sijtsma (1994) showed that fewer items are needed for identifying misfit when item-score reliability is higher. If items are meant to be used as single-item measurement instruments, their suitability for the job envisaged requires high item-score reliability. Single-item instruments are used in work and organizational psychology for selection and assessing, for example, job satisfaction (Gonzalez-Mulé, Carter, & Mount, 2017; Harter, Schmidt, & Hayes, 2002; Nagy, 2002; Robertson & Kee, 2017; Saari & Judge, 2004; Zapf, Vogt, Seifert, Mertini, & Isic, 1999) and level of burnout (Dolan et al., 2014). Item-score reliability is also used in health research for measuring, for example, quality of life (Stewart, Hays, & Ware, 1988; Yohannes, Willgoss, Dodd, Fatoye, & Webb, 2010) and psychosocial stress (Littman, White, Satia, Bowen, & Kristal, 2006), and one-item measures have been assessed in marketing research for measuring ad and brand attitude (Bergkvist & Rossiter, 2007).

Several authors have proposed methods for estimating item-score reliability. Wanous and Reichers (1996) proposed the correction for attenuation (method CA) for estimating item-score reliability. Method CA correlates an item score and a test score both assumed to measure the same attribute. Google Scholar cited Wanous et al. (1997) 2400+ times (November 16 2017), suggesting method CA is used regularly to estimate item-score reliability. The authors proposed to use method CA for estimating item-score reliability for single-item measures that are used, for example, for measuring job satisfaction (Wanous et al., 1997). Meijer et al. (1995) advocated using the Molenaar-Sijtsma method (method MS; Molenaar & Sijtsma, 1988), which at the time was available only for dichotomous items. In this study, method MS was generalized to polytomous item scores. Two novel methods were also proposed, one based on coefficient λ_6 (Guttman, 1945) denoted as method λ_6, and the other based on the latent class reliability coefficient (Van der Ark et al., 2011), denoted as method LCRC. This study discusses methods MS, λ_6, LCRC, and CA, each suitable for polytomous item scores, and compared the methods with respect to median bias, variability expressed as interquartile range (IQR), and percentage of outliers. This study also showed that the well-known coefficients α (Cronbach, 1951) and λ_2 (Guttman, 1945) are inappropriate for being used as item-score reliability methods.

Because item-score reliability addresses the repeatability of item scores in a group of people, it provides information different from other item indices. Examples are the corrected item-total correlation (Nunnally, 1978, p. 281), which quantifies how well the item correlates with the sum score on the other items in the test; the item-factor loading (Harman, 1976, p. 15), which quantifies how well the item is associated with a factor score based on the items in the test, and thus corrects for the multidimensionality of total scores; the item scalability (Mokken, 1971, pp. 151–152), which quantifies the relationship between the item and the other items in the test, each item corrected for the influence of its marginal distribution on the relationship; and the item discrimination (e.g., see Baker & Kim, 2004, p. 4), which quantifies how well the item distinguishes people with low and high scores on a latent variable the items have in common. None of these indices addresses repeatability; hence, item-score reliability may be a useful addition to the set of item indices. A study that addresses the formal relationship between the item indices would more precisely inform us about their differences and similarities, but such a theoretical study is absent in the psychometric literature.

Following this study, which focused on the theory of item-score reliability, Zijlmans, Tijmstra, Van der Ark, and Sijtsma (2018b) estimated methods MS, λ_6, and CA from several empirical data sets to investigate the methods' practical usefulness and values that are found in practice and may be expected in other data sets. In addition, the authors estimated four item indices (corrected item total-correlation, item-factor loading, item scalability, and item discrimination) from the empirical data sets. The values of these four item indices were compared with the values of the item-score reliability methods, to establish the relationship between item-score reliability and the other four item indices.

This article is organized as follows. First, a framework for estimating item-score reliability and three of the item-score reliability methods in the context of this framework are discussed. Second, a simulation study, its results with respect to the methods' median bias, IQR, and percentage of outliers, and a real-data example are discussed. Methods to use in practical data analysis are recommended.

2.2 A Framework for Item-Score Reliability

The following classical test theory (CTT) definitions (Lord & Novick, 1968, p. 61) were used. Let X be the test score, which is defined as the sum of J item scores, indexed i $(i = 1, ..., J)$; that is, $X = \sum_{i=1}^{J} X_i$. In the population, test score X has variance σ_X^2. True score T is the expectation of an individual's test score across independent repetitions, and represents the mean of the individual's propensity distribution (Lord & Novick, 1968, pp. 29-30). The deviation of test score X from true score T is the random measurement error, E; that is, $E = X - T$. Because T and E are unobservable, their variances are also unobservable. Using these definitions, test-score reliability is defined as the proportion of observed-score variance that is true-score variance or, equivalently, one minus the proportion of observed-score variance that is error variance. Mathematically, reliability also equals the product-moment correlation between parallel tests (Lord & Novick, 1968, p. 61), denoted by $\rho_{XX'}$; that is,

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}. \tag{2.1}$$

Next to notation i, we need j to index items. Notation x and y denote realizations of item scores, and without loss of generality it is assumed that $x, y = 0, 1, \ldots, m$. Let $\pi_{x(i)} = P(X_i \geq x)$ be the marginal cumulative probability of obtaining at least score x on item i. It may be noted that $\pi_{0(i)} = 1$ by definition. Likewise, let $\pi_{x(i),y(j)} = P(X_i \geq x, X_j \geq y)$ be the joint cumulative probability of obtaining at least score x on item i and at least score y on item j.

In what follows, it is assumed that index i' indicates an independent repetition of item i. Let $\pi_{x(i),y(i')}$ denote the joint cumulative probability of obtaining at least score x and at least score y on two independent repetitions, denoted by i and i', of the same item in the same group of people. Because independent repetitions are unavailable in practice, the joint cumulative probabilities $\pi_{x(i),y(i')}$ have to be estimated from single-administration data.

Molenaar and Sijtsma (1988) showed that reliability (Equation 2.1) can be written as

$$\rho_{XX'} = \frac{\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{x=1}^{m}\sum_{y=1}^{m}\left[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}\right]}{\sigma_X^2}. \tag{2.2}$$

Equation 2.2 can be decomposed into the sum of two ratios:

$$\rho_{XX'} = \frac{\sum\sum_{i \neq j}^{J}\sum_{x=1}^{m}\sum_{y=1}^{m}\left[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}\right]}{\sigma_X^2} + \frac{\sum_{i=1}^{J}\sum_{x=1}^{m}\sum_{y=1}^{m}\left[\pi_{x(i),y(i')} - \pi_{x(i)}\pi_{y(i)}\right]}{\sigma_X^2}. \tag{2.3}$$

Except for the joint cumulative probabilities pertaining to the same item $\pi_{x(i),y(i')}$, all other terms in Equation 2.3 are observable and can be estimated from the sample. Van der Ark et al. (2011) showed that for test score X, the single-administration reliability methods α, λ_2, MS and LCRC only differ with respect to the estimation of $\pi_{x(i),y(i')}$.

To define item-score reliability, Equation 2.3 can be adapted to accommodate only one item; the first ratio and the first summation sign in the second ratio disappear, and item-score reliability $\rho_{ii'}$ is defined as

$$\rho_{ii'} = \frac{\sum_{x=1}^{m}\sum_{y=1}^{m}\left[\pi_{x(i),y(i')} - \pi_{x(i)}\pi_{y(i)}\right]}{\sigma^2_{X_i}} = \frac{\sigma^2_{T_i}}{\sigma^2_{X_i}}. \quad (2.4)$$

2.3 Methods for Approximating Item-Score Reliability

Three of the four methods that were investigated, methods MS, λ_6, and LCRC, use different approximations to the unobservable joint cumulative probability $\pi_{x(i),y(i')}$, and fit into the same reliability framework. Two other well-known methods that fit into this framework, Cronbach's α and Guttman's λ_2, cannot be used to estimate item-score reliability (see Appendix A). The fourth method, CA, uses a different approach to estimating item-score reliability and conceptually stands apart from the other three methods. All four methods estimate Equation 2.4, which contains two unknowns – in addition to $\rho_{ii'}$ bivariate proportion $\pi_{x(i),y(i')}$ (middle) and variance $\sigma^2_{T_i}$ (right) – and thus cannot be estimated directly from the data.

Method MS

Method MS uses the available marginal cumulative probabilities to approximate $\pi_{x(i),y(i')}$. The method is based on the item response model known as the double monotonicity model (Mokken, 1971, p. 118; Sijtsma & Molenaar, 2002, pp. 23-25). This model is based on the assumptions of a unidimensional latent variable; independent item scores conditional on the latent variable, which is known as local independence; response functions that are monotone nondecreasing in the latent variable; and nonintersection of the response functions of different items. The double monotonicity model implies that the observable bivariate proportions $\pi_{x(i),y(j)}$ collected in the $\mathbf{P}(++)$ matrix are nondecreasing in the rows and the columns (Sijtsma & Molenaar, 2002, pp. 104-105). The structure of the $\mathbf{P}(++)$ matrix using an artificial example is illustrated.

For four items, each having three ordered item scores, Table 2.1 shows the marginal cumulative probabilities. First, ignoring the uninformative $\pi_{0i} = 1$, we assume that probabilities can be strictly ordered, and order the eight remaining

Table 2.1: Marginal Cumulative Probabilities for Four Artificial Items with Three Ordered Item Scores

	Item			
	1	2	3	4
$\pi_{0(i)}$	1.00	1.00	1.00	1.00
$\pi_{1(i)}$	0.97	0.94	0.93	0.86
$\pi_{2(i)}$	0.53	0.32	0.85	0.72

marginal cumulative probabilities in this example from small to large:

$$\pi_{2(2)} < \pi_{2(1)} < \pi_{2(4)} < \pi_{2(3)} < \pi_{1(4)} < \pi_{1(3)} < \pi_{1(2)} < \pi_{1(1)}. \tag{2.5}$$

Van der Ark (2010) discussed the case in which Equation 2.5 contains ties. Second, the $\mathbf{P}(++)$ matrix is defined, which has order $Jm \times Jm$ and contains the joint cumulative probabilities. The rows and columns are ordered reflecting the ordering of the marginal cumulative probabilities, which are arranged from small to large along the matrix' marginals; see Table 2.2. The ordering of the marginal cumulative probabilities determines where each of the joint cumulative probabilities is located in the matrix. For example, the entry in cell (4,7) is $\pi_{2(3),1(2)}$, which equals $.81$. Mokken (1971, pp. 132-133) proved that the double monotonicity model implies that the rows and the columns in the $\mathbf{P}(++)$ matrix are nondecreasing. This is the property on which method MS rests. In Table 2.2, entry NA (i.e., not available) refers to the joint cumulative probabilities of the same item, which are unobservable. For example, in cell (5,3) the proportion $\pi_{1(4),2(4')}$ is NA and hence cannot be estimated numerically.

Method MS uses the adjacent, observable joint cumulative probabilities of different items to estimate the unobservable joint cumulative probabilities $\pi_{x(i),y(i')}$ by means of eight approximation methods (Molenaar & Sijtsma, 1988). For test scores, Molenaar and Sijtsma (1988) explained that method MS attempts to approximate the item response functions of an item and for this purpose uses adjacent items, because when item response functions do not intersect, adjacent functions are more similar to the target item response function, thus approximating repetitions of the same item, than item response functions further away. When an adjacent probability is unavailable, for example, in the first and last rows and the first and last columns in Table 2.2, only the available estimators are used. For example, $\pi_{1(1),2(1')}$ in cell (8,2) does not have lower neighbors. Hence, only the proportions $.32$, cell (8,1); $.51$, cell (7,2); and $.70$, cell (8,3) are available for approximating $\pi_{1(1),2(1')}$. For further details, see Molenaar and Sijtsma (1988) and Van der Ark (2010).

Table 2.2: $\mathbf{P}(++)$ Matrix with Joint Cumulative Probabilities $\pi_{x(i),y(j)}$ and Marginal Cumulative Probabilities $\pi_{x(i)}$

		$\pi_{2(2)}$	$\pi_{2(1)}$	$\pi_{2(4)}$	$\pi_{2(3)}$	$\pi_{1(4)}$	$\pi_{1(3)}$	$\pi_{1(2)}$	$\pi_{1(1)}$
		.32	.53	.72	.85	.86	.93	.94	.97
$\pi_{2(2)}$	.32	NA	0.20	0.27	0.29	0.30	0.31	NA	0.32
$\pi_{2(1)}$	.53	0.20	NA	0.41	0.47	0.48	0.50	0.51	NA
$\pi_{2(4)}$	.72	0.27	0.41	NA	0.64	NA	0.68	0.68	0.70
$\pi_{2(3)}$	.85	0.29	0.47	0.64	NA	0.76	NA	0.81	0.84
$\pi_{1(4)}$	.86	0.30	0.48	NA	0.76	NA	0.81	0.81	0.84
$\pi_{1(3)}$	.93	0.31	0.50	0.68	NA	0.81	NA	0.88	0.91
$\pi_{1(2)}$	.94	NA	0.51	0.68	0.81	0.81	0.88	NA	0.91
$\pi_{1(1)}$	.97	0.32	NA	0.70	0.84	0.84	0.91	0.91	NA

Note. NA = not available

Hence, following Molenaar and Sijtsma (1988), the joint cumulative probability $\pi_{x(i),y(i')}$ is approximated by the mean of at most eight approximations resulting in $\tilde{\pi}^{MS}_{x(i),y(i')}$. When the double monotonicity model does not hold, item response functions adjacent to the target item response function may intersect and not approximate the target very well, so that $\tilde{\pi}^{MS}_{x(i),y(i')}$ may be a poor approximation of $\pi_{x(i),y(i')}$. The approximation of $\pi_{x(i),y(i')}$ by method MS is used in Equation 2.4 to estimate the item-score reliability.

Method MS is equal to item-score reliability $\rho_{ii'}$ when $\sum_x \sum_y \pi_{x(i)y(i')} = \sum_x \sum_y \tilde{\pi}^{MS}_{x(i)y(i')}$. A sufficient condition is that all the entries in the $\mathbf{P}(++)$ matrix are equal; equality of entries requires item response functions that coincide. Further study of this topic is beyond the scope of this article but should be taken up in future research.

Method λ_6

An item-score reliability method based on Guttman's λ_6 (Guttman, 1945) can be derived as follows. Let ϵ_i^2 denote the variance of the estimation or residual error of the multiple regression of item score X_i on the remaining $J-1$ item scores, and determine ϵ_i^2 for each of the J items. Guttman's λ_6 is defined as

$$\lambda_6 = 1 - \frac{\sum_{i=1}^{J} \epsilon_i^2}{\sigma_X^2}. \tag{2.6}$$

It may be noted that Equation 2.6 resembles the right-hand side of Equation 2.1. Let $\mathbf{\Sigma}_{ii}$ denote the $(J-1) \times (J-1)$ inter-item variance-covariance matrix for $(J-1)$

items except item i. Let $\boldsymbol{\sigma}_i$ be a $(J-1)\times$ 1 vector containing the covariances of item i with the other $(J-1)$ items. Jackson and Agunwamba (1977) showed that the variance of the estimation error equals

$$\epsilon_i^2 = \sigma_{X_i}^2 - \boldsymbol{\sigma}_i'(\boldsymbol{\Sigma}_{ii})^{-1}\boldsymbol{\sigma}_i. \tag{2.7}$$

When estimating the reliability of an item score, Equation 2.6 can be adapted to

$$\lambda_{6_i} = 1 - \frac{\sigma_{X_i}^2 - \boldsymbol{\sigma}_i'(\boldsymbol{\Sigma}_{ii})^{-1}\boldsymbol{\sigma}_i}{\sigma_{X_i}^2} = \frac{\boldsymbol{\sigma}_i'(\boldsymbol{\Sigma}_{ii})^{-1}\boldsymbol{\sigma}_i}{\sigma_{X_i}^2}. \tag{2.8}$$

It can be shown that method λ_6 fits into the framework of Equation 2.4. Let $\tilde{\pi}_{x(i),y(i')}^{\lambda_6}$ be an approximation of $\pi_{x(i),y(i')}$ based on observable proportions, such that replacing $\pi_{x(i),y(i')}$ in the right-hand side of Equation 2.4 by $\tilde{\pi}_{x(i),y(i')}^{\lambda_6}$ results in λ_{6_i}. Hence,

$$\lambda_{6_i} = \frac{\sum_{x=1}^{m}\sum_{y=1}^{m}\left[\tilde{\pi}_{x(i),y(i')}^{\lambda_6} - \pi_{x(i)}\pi_{y(i)}\right]}{\sigma_{X_i}^2}. \tag{2.9}$$

Equating Equation 2.8 and 2.9 shows that

$$\begin{aligned} \frac{\boldsymbol{\sigma}_i'(\boldsymbol{\Sigma}_{ii})^{-1}\boldsymbol{\sigma}_i}{\sigma_{X_i}^2} &= \frac{\sum_{x=1}^{m}\sum_{y=1}^{m}\left[\tilde{\pi}_{x(i),y(i')}^{\lambda_6} - \pi_{x(i)}\pi_{y(i)}\right]}{\sigma_{X_i}^2} \iff \\ \frac{\boldsymbol{\sigma}_i'(\boldsymbol{\Sigma}_{ii})^{-1}\boldsymbol{\sigma}_i}{m^2} &= \tilde{\pi}_{x(i),y(i')}^{\lambda_6} - \pi_{x(i)}\pi_{y(i)} \iff \\ \tilde{\pi}_{x(i),y(i')}^{\lambda_6} &= \frac{\boldsymbol{\sigma}_i'(\boldsymbol{\Sigma}_{ii})^{-1}\boldsymbol{\sigma}_i}{m^2} + \pi_{x(i)}\pi_{y(i)} \end{aligned} \tag{2.10}$$

Inserting $\tilde{\pi}_{x(i),y(i')}^{\lambda_6}$ in Equation 2.4 yields method λ_6 for item-score reliability. Replacing parameters by sample statistics produces an estimate.

Preliminary computations suggest that only highly contrived conditions produce the equality $\sigma_{T_i}^2 = \boldsymbol{\sigma}_i'(\boldsymbol{\Sigma}_{ii})^{-1}\boldsymbol{\sigma}_i$ in Equation 2.8, but conditions more representative for what one may find with real data produce negative item true-score variance, also known as Heywood cases. Because this work is premature, we tentatively conjecture that in practice, method λ_6 is a strict lower bound to the item-score reliability, a result that is consistent with simulation results discussed elsewhere (e.g., Oosterwijk, Van der Ark, & Sijtsma, 2017).

Method LCRC

Method LCRC is based on the unconstrained latent class model (LCM; Hagenaars & McCutcheon, 2002; Lazarsfeld, 1950; McCutcheon, 1987). The LCM assumes local independence, meaning that item scores are independent given class membership. Two different probabilities are important, which are the latent class probabilities that provide the probability to be in a particular latent class $k(k = 1, \ldots, K)$, and the latent response probabilities that provide the probability of a particular item score given class membership. For local independence given a discrete latent variable ξ with K classes, the unconstrained LCM is defined as

$$P(X_1 = x_1, ..., X_J = x_J) = \sum_{k=1}^{K} P(\xi = k) \prod_{j=1}^{J} P(X_i = x_i \mid \xi = k). \tag{2.11}$$

The LCM (Equation 2.11) decomposes the joint probability distribution of the J item scores for the sum across K latent classes of the product of the probability to be in class k and the conditional probability of a particular item score X_i. Let $\tilde{\pi}^{LCRC}_{x(i),y(i')}$ be the approximation of $\pi_{x(i),y(i')}$ using the parameters of the unconstrained LCM at the right-hand side of Equation 2.11, such that

$$\tilde{\pi}^{LCRC}_{x(i),y(i')} = \sum_{u=x}^{m} \sum_{v=y}^{m} \sum_{k=1}^{K} P(\xi = k) P(X_i = u \mid \xi = k) P(X_i = v \mid \xi = k). \tag{2.12}$$

Approximation $\tilde{\pi}^{LCRC}_{x(i),y(i')}$ can be inserted in Equation 2.4 to obtain method LCRC. After insertion of sample statistics, an estimate of method LCRC is obtained.

Method LCRC equals $\rho_{ii'}$ if $\pi_{x(i),y(i')}$ (Equation 2.4) equals $\tilde{\pi}^{LCRC}_{x(i),y(i')}$ (Equation 2.12), hence $\pi_{x(i),y(i')} = \sum_{u=x}^{m} \sum_{v=y}^{m} \sum_{k=1}^{K} P(\xi = k) P(X_i = u \mid \xi = k) P(X_i = v \mid \xi = k)$. A sufficient condition for method LCRC to equal $\rho_{ii'}$ is that K has been correctly selected and all estimated parameters $P(\xi = k)$ and $P(X_i = x \mid \xi = k)$ equal the population parameters. This condition is unlikely to be true in practice. In samples, LCRC may either underestimate or overestimate $\rho_{ii'}$.

Method CA

The CA (Lord & Novick, 1968, pp. 69-70; Nunnally & Bernstein, 1994, p. 257; Spearman, 1904) can be used for estimating item-score reliability (Wanous & Reichers, 1996). Let Y be a random variable, which preferably measures the same attribute as item score X_i but does not include X_i. Likely candidates for Y are the rest score $R_{(i)} = X - X_i$ or the test score on another, independent test that does not include item score X_i but measures the same attribute. Let $\rho_{T_{X_i}T_Y}$ be the correlation between true scores T_{X_i} and T_Y, let ρ_{X_iY} be the correlation between X_i and Y, let $\rho_{ii'}$ be the item-score reliability of X_i, and let $\rho_{YY'}$ be the reliability of Y. Then, method CA equals

$$\rho_{T_{X_i}T_Y} = \frac{\rho_{X_iY}}{\sqrt{\rho_{ii'}} \cdot \sqrt{\rho_{YY'}}}. \tag{2.13}$$

It follows from Equation 2.13 that the item-score reliability equals

$$\rho_{ii'} = \left(\frac{\rho_{X_iY}}{\rho_{T_{X_i}T_Y}\sqrt{\rho_{YY'}}}\right)^2 = \frac{\rho^2_{X_iY}}{\rho^2_{T_{X_i}T_Y}\rho_{YY'}}. \tag{2.14}$$

Let $\tilde{\rho}^{CA}_{ii'}$ denote the item-score reliability estimated by method CA. Method CA is based on two assumptions. First, true scores T_{X_i} and T_Y correlate perfectly; that is, $\rho_{T_{X_i}T_Y} = 1$, reflecting that T_{X_i} and T_Y measure the same attribute. Second, $\rho_{YY'}$ equals the population reliability. Because many researchers use coefficient alpha (alpha$_Y$) to approximate $\rho_{YY'}$ in practice, it is assumed that alpha$_Y = \rho_{YY'}$. Using these two assumptions, Equation 2.14 reduces to

$$\tilde{\rho}^{CA}_{ii'} = \frac{\rho^2_{X_iY}}{\text{alpha}_Y}. \tag{2.15}$$

Comparing $\tilde{\rho}^{CA}_{ii'}$ and $\rho_{ii'}$, one may notice that $\tilde{\rho}^{CA}_{ii'} = \rho_{ii'}$ if the denominators in Equations 2.15 and 2.14 are equal; that is, if alpha$_Y = \rho^2_{T_{X_i}T_Y}\rho_{YY'}$. When does this happen? Assume that $Y = R_{(i)}$. Then, if the $J-1$ items on which Y is based are essentially τ-equivalent, meaning that $T_{X_i} = T_Y + b_{iY}$ (Lord & Novick, 1968, p. 50), then alpha$_Y = \rho_{YY'}$. This results in $\rho_{YY'} = \rho^2_{T_{X_i}T_Y}\rho_{YY'}$, implying that $\rho^2_{T_{X_i}T_Y} = 1$, hence $\rho_{T_{X_i}T_Y} = 1$, and this is true if T_{X_i} and T_Y are linearly related: $T_{X_i} = a_{iY}T_Y + b_{iY}$. Because it is already assumed that items are essentially τ-equivalent and because the linear relation has to be true for all J items, $b_i = 0$ for all i and $\tilde{\rho}^{CA}_{ii'} = \rho_{ii'}$ if all items are essentially τ-equivalent. Further study of the relation between $\tilde{\rho}^{CA}_{ii'}$ and $\rho_{ii'}$ is beyond the scope of this article, and is referred to future research.

2.4 Simulation Study

A simulation study was performed to compare median bias, IQR, and percentage of outliers produced by item-score reliability methods MS, λ_6, LCRC, and CA. Joint cumulative probability $\pi_{x(i),y(i')}$ was estimated using methods MS, λ_6 and LCRC. For these three methods, the estimates of the joint cumulative probabilities $\pi_{x(i),y(i')}$ were inserted in Equation 2.4 to estimate the item-score reliability. For method CA, Equation 2.15 was used.

Method

Dichotomous or polytomous item scores were generated using the multidimensional graded response model (Ayala, 1994). Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_Q)$ be the Q-dimensional latent-variable vector, which has a Q-variate standard normal distribution. Let α_{iq} be the discrimination parameter of item i relative to latent variable q, and let δ_{ix} be the location parameter for category x $(x = 1, 2, \dots, m)$ of item i.

The multidimensional graded response model (Ayala, 1994) is defined as

$$P(X_i \geq x \mid \theta) = \frac{\exp\left[\sum_{q=1}^{Q} \alpha_{iq}(\theta_q - \delta_{ix})\right]}{1 + \exp\left[\sum_{q=1}^{Q} \alpha_{iq}(\theta_q - \delta_{ix})\right]}. \tag{2.16}$$

The design for the simulation study was based on the design used by Van der Ark et al. (2011) for studying test-score reliability. A standard condition was defined for six dichotomous items ($J = 6, m + 1 = 2$), one dimension ($Q = 1$), equal discrimination parameters ($\alpha_{iq} = 1$ for all i and q) and equidistantly spaced location parameters δ_{ix} ranging from -1.5 to 1.5 (Table 2.3), and sample size $N = 1000$. The other conditions differed from the standard condition with respect to one design factor. Test length, sample size, and item-score format were considered extensions of the standard condition, and discrimination parameters and dimensionality were considered deviations, possibly affecting methods the most.

Test length (J): The test consisted of 18 items ($J = 18$). For this condition, the six items from the standard condition were copied twice.

Sample size (N): The sample size was small ($N = 200$).

Item-score format (m + 1): The J items were polytomous ($m + 1 = 5$).

Discrimination parameters (α): Discrimination parameters differed across items ($\alpha = 0.5$ or 2). This constituted a violation of the assumption of nonintersecting item response functions needed for method MS.

Dimensionality (Q): The items were two-dimensional ($Q = 2$) with latent variables correlating .5. The location parameters alternated between the two dimensions. This condition is more realistic than the condition chosen in Van der Ark et al. (2011), representing two subscale scores that are combined into an overall measure, whereas Van der Ark et al. (2011) used orthogonal dimensions.

Van der Ark et al. (2011) found that item format and sample size did not affect bias of test-score reliability, but these factors were included in this study to find out whether results for individual items were similar to results for test scores.

Data sets were generated as follows. For every replication, N latent variable vectors, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$, were randomly drawn from the $\boldsymbol{\theta}$ distribution. For each set of latent variable scores, for each item, the m cumulative response probabilities were computed using Equation 2.16. Using the m cumulative response probabilities, item scores were drawn from the multinomial distribution. In each condition, 1000 data sets were drawn.

Population item-score reliability $\rho_{ii'}$ was approximated by generating item scores for 1 million simulees (i.e., sets of item scores). For each item, the variance based on the $\boldsymbol{\theta}$s of the 1 million simulees was divided by the variance of the item

Table 2.3: Item Parameters of the Multidimensional Graded Response Model for the Simulation Design

	Design													
Item	Standard		Polytomous					Unequal α		Two Dimensions		
	α_j	δ_j	α_j	δ_{j1}	δ_{j2}	δ_{j3}	δ_{j4}	α_j	δ_j	α_{j1}	α_{j2}	δ_j
1	1	-1.5	1	-3	-2	-1	0	0.5	-1.5	1	0	-1.5
2	1	-0.9	1	-2.4	-1.4	-0.4	0.6	2	-0.9	0	1	-0.9
3	1	-0.3	1	-1.8	-0.8	0.2	1.2	0.5	-0.3	1	0	-0.3
4	1	0.3	1	-1.2	-0.2	0.8	1.8	2	0.3	0	1	0.3
5	1	0.9	1	-0.6	0.4	1.4	2.4	0.5	0.9	1	0	0.9
6	1	1.5	1	0	1	2	3	2	1.5	0	1	1.5

Note. α = item discrimination, δ = item location

score X_i to obtain the population item-score reliability. It was found that $.05 \leq \rho_{ii'} \leq .41$.

Let s_r be the estimate of $\rho_{ii'}$ in replication r $(r = 1, \ldots, R)$ by means of methods MS, λ_6, and CA. For each method, difference $(s_r - \rho_{ii'})$ is displayed in boxplots. For each item-score reliability method, median bias, IQR, and percentage of outliers were recorded. An overall measure reflecting estimation quality based on the three quantities was not available, and in cases where a qualification of a method's estimation quality was needed, we indicated how the median bias, IQR, and percentage of outliers were weighted. The computations were done using R (R Core Team, 2016). The code is available via https://osf.io/e83tp/. For the computation of method MS, the package `mokken` was used (Van der Ark, 2007, 2012). For the computation of the LCM used for estimating method LCRC, the package `poLCA` was used (Linzer & Lewis, 2011).

Results

For each condition, Figure 2.1 shows the boxplots for the difference $(s_r - \rho_{ii})$. In general, differences across items in the same experimental condition were negligible; hence, the results were aggregated not only across replications but also across the items in a condition, so that each condition contained $J \times 1000$ estimated item-score reliabilities. The bold horizontal line in each boxplot represents median bias. The dots outside the whiskers are outliers, defined as values that lie beyond 1.5 times the IQR measured from the whiskers of the first and the third quartile. For unequal αs and for $Q = 2$, results are presented separately for high and low αs and for each θ, respectively.

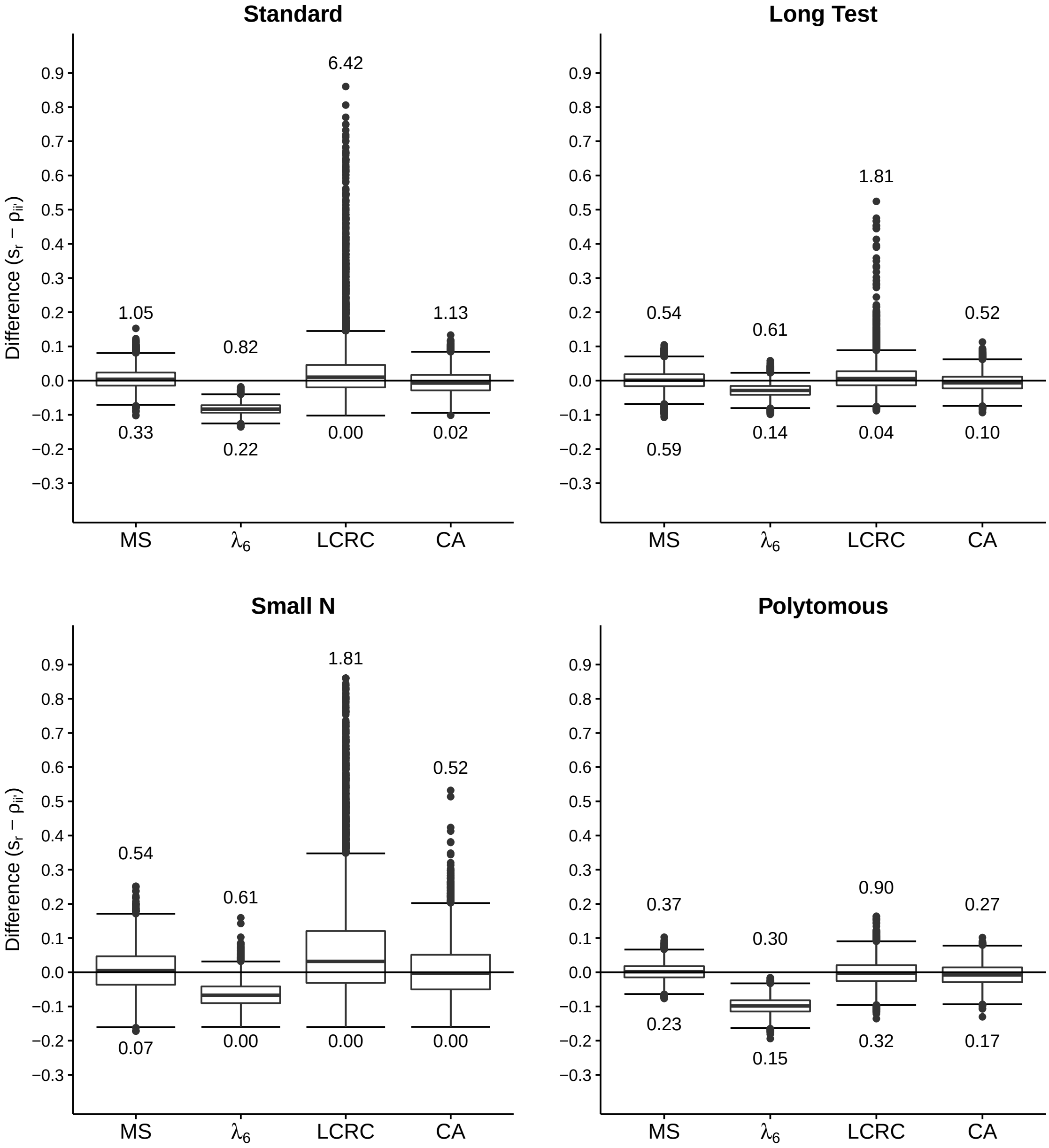

Figure 2.1: Difference $(s_r - \rho_{ii'})$, where s_r represents an estimate of methods MS, λ_6, LCRC, and CA, for six different conditions (see Table 2.3 for the specifications of the conditions).

Note. The bold horizontal line represents the median bias. The numbers in the boxplots represent the percentage outliers in that condition. MS = Molenaar-Sijtsma method; λ_6 = Guttman's method λ_6; LCRC = latent class reliability coefficient; CA = correction for attenuation.

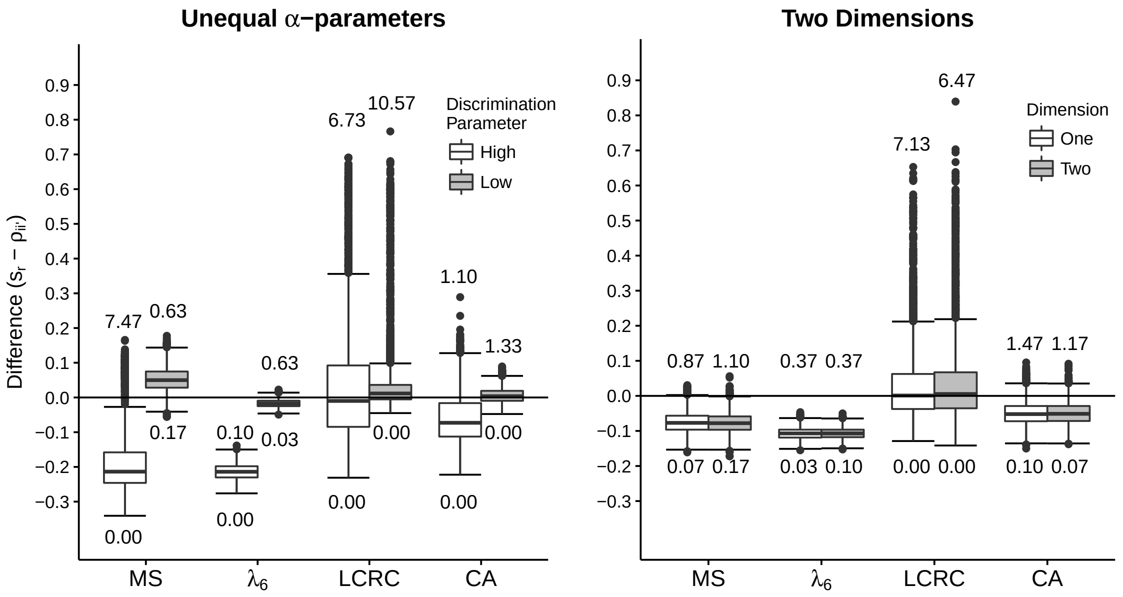

Figure 2.1, continued: Difference $(s_r - \rho_{ii'})$, where s_r represents an estimate of methods MS, λ_6, LCRC, and CA, for six different conditions (see Table 2.3 for the specifications of the conditions).

Note. The bold horizontal line represents the median bias. The numbers in the boxplots represent the percentage outliers in that condition. MS = Molenaar-Sijtsma method; λ_6 = Guttman's method λ_6; LCRC = latent class reliability coefficient; CA = correction for attenuation.

In the standard condition (Figure 2.1), median bias for methods MS, LCRC, and CA was close to 0. For method LCRC, 6.4 % of the difference $(s_r - \rho_{ii'})$ qualified as an outlier. Hence, compared with methods MS and CA, method LCRC had a large IQR. Method λ_6 consistently underestimated item-score reliability. In the long-test condition (Figure 2.1), for all methods, the IQR was smaller than in the standard condition. For the small-N condition (Figure 2.1), for all methods IQR was a little greater than in the standard condition. In the polytomous item condition (Figure 2.1), median bias and IQR results were comparable with results in the standard condition, but method LCRC showed fewer outliers (i.e, 1.2 %).

Results for high-discrimination items and low-discrimination items can be found in Figure 2.1, unequal α-parameters condition panel. Median bias was smaller for low-discrimination items. For both high and low-discimination items, method LCRC produced median bias close to 0. Compared to the standard condition, IQR was greater for high-discrimination items and the percentage of outliers was higher for both high- and low-discrimination items. For high-discrimination items, methods MS, λ_6 and CA showed greater negative median bias than for low-discrimination items. For low-discrimination items, method MS had a small positive bias and for methods λ_6 and CA, the results were similar to the standard condition. For the two-dimensional data condition (Figure 2.1), methods MS and CA produced larger median bias compared to the standard condition. Methods

LCRC and CA also produced larger IQR than in the standard condition. Method λ_6 showed smaller IQR than in the standard condition.

A simulation study performed for six items with equidistantly spaced location parameters ranging from -2.5 to 2.5, showed that the number of outliers was larger for all methods, ranging from 0 % to 9.6 % percent. This result was also found when the items having the highest and lowest discrimination parameter were omitted.

Depending on the starting values, the expectation maximization (EM) algorithm estimating the parameters of the LCM may find a local optimum rather than the global optimum of the loglikelihood. Therefore, for each item-score reliability coefficient, the LCM was estimated 25 times using different starting values. The best-fitting LCM was used to compute the item-score reliability coefficient. This produced the same results, and left the former conclusion unchanged.

2.5 Real-Data Example

A real-data set illustrated the most promising item-score reliability methods. Because method LCRC had large IQR and a high percentages of outliers and because results were better and similar for the other three methods, methods MS, λ_6, and CA were selected as the three most promising methods. The data set ($N = 425$) consisted of $0/1$ scores on 12 dichotomous items measuring transitive reasoning (Verweij, Sijtsma, & Koops, 1999). The corrected item-total correlation, the item-factor loading based on a confirmatory factor model, the item-scalability coefficient (denoted H_i; Mokken, 1971, pp. 151–152), and the item-discrimination parameter (based on a two-parameter logistic model) were also estimated. The latter four measures provide an indication of item quality from different perspectives, and use different rules of thumb for interpretation. De Groot and Van Naerssen (1969, p. 351) suggested $.3$ to $.4$ as minimally acceptable corrected item-total correlations for maximum-performance tests. For the item-factor loading, values of $.3$ to $.4$ are most commonly recommended (Gorsuch, 1983, p. 210; Nunnally, 1978, pp. 422–423; Tabachnick & Fidell, 2007, p. 649). Sijtsma and Molenaar (2002, p. 36) suggested to only accept items having $H_i \geq .3$ in a scale. Finally, Baker (2001, p. 34) recommended a lower bound of 0.65 for item discrimination.

Using these rules of thumb yielded the following results (Table 2.4). Only item 3 met the rules of thumb value for the four item indices. Item 3 also had the highest estimated item-score reliability, exceeding $.3$ for all three methods. Items 2, 4, 7, and 12 did not meet the rules of thumb of any of the item indices. These items had the lowest item-score reliability not exceeding $.3$ for any method.

Table 2.4: Estimated Item Indices for the Transitive Reasoning Data Set

Item	Item Mean	Item-Score Reliability			Item Indices			
		Method MS	Method λ_6	Method CA	Corrected Item-Total Correlation	Item-Factor Loading	Item Scalability	Item Discrimination
X_1	0.97	**0.36**	0.28	0.21	0.26	**0.85**	0.28	**2.69**
X_2	0.81	0.01	0.13	0.05	0.13	-0.04	0.08	-0.05
X_3	0.97	**0.47**	**0.30**	**0.35**	**0.33**	**0.88**	**0.40**	**3.16**
X_4	0.78	0.05	0.13	0.02	0.08	-0.10	0.05	-0.20
X_5	0.84	0.18	0.23	**0.31**	0.29	**0.73**	0.18	**1.94**
X_6	0.94	**0.32**	0.20	0.17	0.23	**0.74**	0.21	**2.04**
X_7	0.64	0.03	0.05	0.00	-0.04	-0.06	-0.03	-0.01
X_8	0.88	**0.39**	**0.30**	0.26	0.28	**0.83**	0.19	**2.54**
X_9	0.80	0.05	0.06	0.07	0.15	0.34	0.09	0.64
X_{10}	0.30	0.00	0.10	0.10	0.18	0.48	0.17	**1.03**
X_{11}	0.52	0.00	0.17	0.14	0.21	0.61	0.14	**1.36**
X_{12}	0.48	0.00	0.07	0.06	-0.17	-0.29	-0.14	-0.50

Note. Bold faced values are above the heuristic rule for that item index

2.6 Discussion

Methods MS, λ_6 and LCRC were adjusted for estimating item-score reliability. Method CA was an existing method. The simulation study showed that methods MS and CA had the smallest median bias. Method λ_6 estimated $\rho_{ii'}$ with the smallest variability, but this method underestimated item-score reliability in all conditions, probably because it is a lower bound to the reliability, rendering it highly conservative. The median bias of method LCRC across conditions was almost 0, but the method showed large variability and produced many outliers overestimating item-score reliability.

It was concluded that in the unequal α-parameters condition and in the two-dimensional condition, the methods do not estimate item-score reliability very accurately (based on median bias, IQR, and percentage of outliers). Compared with the standard condition, for unequal α-parameters, for high-discrimination items, median bias is large, variability is larger, and percentage of outliers is smaller. The same conclusion holds for the multidimensional condition. In practice, unequal α-parameters across items and multidimensionality are common, implying that $\rho_{ii'}$ is underestimated. In the other conditions, methods MS and CA produced the smallest median bias and the smallest variability, while method λ_6 produced small variability but showed larger negative median bias which rendered it conservative.

Table 2.5: *Parameters of Latent Class Models Having Two and Three Classes*

Two-class model		Three-class model	
Class weights	Response probabilities	Class weights	Response probabilities
$P(\hat{\xi}=1)=.4$	$P(X_i=1\|\hat{\xi}=1)=.5$	$P(\hat{\xi}=1)=.4$	$P(X_i=1\|\hat{\xi}=1)=.5$
$P(\hat{\xi}=2)=.6$	$P(X_i=1\|\hat{\xi}=2)=.8$	$P(\hat{\xi}=2)=.3$	$P(X_i=1\|\hat{\xi}=2)=.6$
		$P(\hat{\xi}=3)=.3$	$P(X_i=1\|\hat{\xi}=3)=1.0$

Method LCRC showed small median bias, but large variability.

We conjecture that the way the fit of the LCM is established causes the large variability, and provide some preliminary thoughts for dichotomous items. For the population probabilities $\pi_{1(i)}$ and $\pi_{1(i),1(i')}$ defined earlier, let $\hat{\pi}_{1(i)} = \sum_k P(\hat{\xi} = k)P(X_i = 1|\hat{\xi} = k)$ and $\hat{\pi}_{1(i),1(i')} = \sum_k P(\hat{\xi} = k)(P[X_i = 1|\hat{\xi} = k])^2$ be the their latent class estimates based on sample data, and let $p_{1(i)}$ denote the sample proportion of respondents that have score 1 on item i. For dichotomous items, the item-score reliability (Equation 4) reduces to

$$\rho_{ii'} = \frac{\pi_{1(i),1(i')} - \pi^2_{1(i)}}{\pi_{1(i)}(1-\pi_{1(i)})}. \tag{2.17}$$

In samples, method LCRC estimates Equation 2.17 by means of

$$\hat{\rho}_{ii'} = \frac{\hat{\pi}_{1(i),1(i')} - p^2_{1(i)}}{p_{1(i)}(1-p_{1(i)})}. \tag{2.18}$$

The fit of a LCM is based on a distance measure between $\hat{\pi}_{1(i)}$ and $p_{1(i)}$. However, the fit of the LCM is not directly relevant for Equation 2.18, because $\hat{\pi}_{1(i)}$ does not play a role in this equation. A more relevant fit measure for Equation 2.18 would be based on a distance measure between $\hat{\pi}_{1(i),1(i')}$ and an observable quantity, but such a fit measure is unavailable. The impact of $\hat{\pi}_{1(i),1(i')}$ not being considered in the model fit is illustrated by means of the following example. Table 2.5 shows the parameter estimates of LCMs with two and three classes that both produce perfect fit; that is, one can derive from the parameter estimates that for both models $\hat{\pi}_{1(i)} = p_{1(i)} = .68$. In addition, one can also derive from the parameter estimates that for the two-class model, $\hat{\pi}_{1(i),1(i')} = .484$ and $\hat{\rho}_{ii'} = .099$, whereas for the three-class model, $\hat{\pi}_{1(i),1(i')} = .508$ and $\hat{\rho}_{ii'} = .210$. This example shows that, although the two LCMs both show perfect fit, the resulting values of $\hat{\rho}_{ii'}$ vary considerably. Hence, the variability of the LCRC estimate is larger than the fit of the LCM, and this may explain the large variability of method LCRC in the simulation study.

Values for item-score reliability ranging from $.05$ to $.41$ were used. These values are small compared with values suggested in the literature. For example, Wanous and Reichers (1996) suggested a minimally acceptable item-score reliability of $.70$ in the context of overall job satisfaction, and Ginns and Barrie (2004) suggested values in excess of $.90$. It was believed that for most applications, such high values may not be realistic. In the real-data example, item-score reliability estimates ranged from $< .01$ to $.47$. Further research is required to determine realistic values of item-score reliability. In this study, the range of investigated values for $\rho_{ii'}$ was restricted. The item-score reliability methods' behavior should be investigated under different conditions for a broader range of values for $\rho_{ii'}$. This research is now under way.

Chapter 3

Item-Score Reliability in Empirical-Data Sets and Its Relationship With Other Item Indices

Abstract

Reliability is usually estimated for a total score, but it can also be estimated for item scores. Item-score reliability can be useful to assess the repeatability of an individual item score in a group. Three methods to estimate item-score reliability are discussed, known as method MS, method λ_6, and method CA. The item-score reliability methods are compared with four well-known and widely accepted item indices, which are the corrected item-total correlation, the item-factor loading, the item scalability, and the item discrimination. Realistic values for item-score reliability in empirical-data sets are monitored to obtain an impression of the values to be expected in other empirical-data sets. The relation between the three item-score reliability methods and the four well-known item indices are investigated. Tentatively, a minimum value for the item-score reliability methods to be used in item analysis is recommended.

Keywords: coefficient λ_6, corrected item-total correlation, correction for attenuation, item discrimination, item-factor loading, item scalability, item-score reliability

Based on Zijlmans, E. A. O., Van der Ark, L. A., Tijmstra, J., and Sijtsma, K. (2018). Item-score reliability in empirical-data sets and its relationship with other item indices. *Educational and Psychological Measurement, 78*, 998–1020. doi:10.1177/0013164417728358

3.1 Introduction

This article discusses the practical usefulness of item-score reliability. Usually, reliability of test scores rather than item scores is considered, because test scores and not individual item scores are used to assess an individual's ability or trait level. The test score is constructed of item scores, meaning that all the items in a test contribute to the test-score reliability. Therefore, individual item-score reliability may be relevant when constructing a test, because an item having low reliability may not contribute much to the test-score reliability and may be a candidate for removal from the test.

Item-score reliability (Wanous et al., 1997, cited 2000 + times in Google Scholar, retrieved on July 27, 2017) is used in applied psychology to assess one-item measures for job satisfaction (Gonzalez-Mulé et al., 2017; Harter et al., 2002; Nagy, 2002; Robertson & Kee, 2017; Saari & Judge, 2004; Zapf et al., 1999) and burnout level (Dolan et al., 2014). Item-score reliability is also used in health research for measuring, for example, quality of life (Stewart et al., 1988; Yohannes et al., 2010) and psychosocial stress (Littman et al., 2006), and one-item measures have been assessed in marketing research for measuring ad and brand attitude (Bergkvist & Rossiter, 2007). However, the psychometric theory of item-score reliability appears not to be well developed, and because of this and its rather widespread practical use, we think item-score reliability deserves further study.

Currently, instead of item-score reliability researchers use several other item indices to assess item quality, for example, the corrected item-total correlation (Nunnally, 1978, p. 281), also known as the item-rest correlation, the item-factor loading (Harman, 1976, p. 15), the item-scalability coefficient (Mokken, 1971, pp. 151-152), and the item-discrimination parameter (Baker & Kim, 2004, p. 4). Although useful, these indices are not specifically related to the item-score reliability. Therefore, we also investigated the relation between these item indices and item-score reliability in empirical-data sets.

Let X_i be an item score indexed i $(i = 1, \ldots, J)$, and let X be the test score, which is defined as the sum of the J item scores; that is, $X = \sum_{i=1}^{J} X_i$.

The context of our work is classical test theory. The three methods we use and briefly discuss are all based on the reliability definition proposed by Lord and Novick (1968, p. 61). To estimate item-score reliability, method MS (Molenaar & Sijtsma, 1988) uses data features related to nonparametric item response theory (IRT; Mokken, 1971, pp. 142–147), and the other two methods use estimation procedures based on multiple regression (method λ_6; Guttman, 1945) and correction for attenuation (method CA; Wanous et al., 1997; Wanous & Reichers, 1996). Consistent with classical test theory, item-score reliability for any item i, denoted by $\rho_{ii'}$, is defined as the product-moment correlation between two independent replications of the same item in the same group of people. Because independent replications are unavailable in practice, $\rho_{ii'}$ cannot be estimated directly by means of a sample correlation $r_{ii'}$. Zijlmans, Van der Ark, Tijmstra, and Sijtsma (2018) identified three promising methods for the estimation of item-score reliability, which are method MS, method λ_6, and method CA. Their simulation study results suggested that method MS and method CA had little bias. Method λ_6 produced precise estimates of $\rho_{ii'}$, but systematically underestimated $\rho_{ii'}$, suggesting the method is conservative.

Little is known about the item-score reliability values one can expect to find in empirical data and which values should be considered acceptable for an item to be included in a test. We estimated MS, λ_6, and CA values for the items in 16 empirical-data sets to gain insight into empirical-data values one may expect to find when analyzing one's data. We also estimated the corrected item-total correlation, the item-factor loading, the item scalability, and the item discrimination in these empirical-data sets, and compared their values with the values of the three item-score reliability methods.

This article is organized as follows: First, we discuss item-score reliability methods MS, λ_6, and CA, and the corrected item-total correlation, the item-factor loading, the item scalability, and the item discrimination. Second, the different sets of empirical data for which the seven item indices were estimated are discussed. Third, we discuss the results and their implications for the practical use of the three item-score reliability methods.

3.2 Method

Item-Score Reliability Methods

The following definitions (Lord & Novick, 1968, p. 61) were used. In the population, test score X has variance σ^2_X. True score T is the expectation of an individual's test score across independent replications of the same test, and represents the mean of the individual's distribution of test scores, known as his or her propensity distribution (Lord & Novick, 1968, pp. 29-30). The deviation of test score X from true score T is the random measurement error, E; that is, $E = X - T$. Because T and E are unobservable, their group variances σ^2_T and σ^2_E are also unobservable.

Furthermore, to define the test score's reliability, classical test theory uses the concept of parallel tests to formalize independent replications of the same test in the same group. Two tests with test scores X and X' are parallel (Lord & Novick, 1968, p. 61) if (a) for each person ν, true scores are equal, $T_\nu = T'_\nu$, implying at the group level that $\sigma^2_T = \sigma^2_{T'}$, and (b) for both tests, test-score variances are equal, $\sigma^2_X = \sigma^2_{X'}$. The definition implies that measurement-error variances are also equal, $\sigma^2_E = \sigma^2_{E'}$.

Using the definition of parallel tests, test-score reliability is defined as the product-moment correlation between test scores X and X', and denoted by $\rho_{XX'}$. Correlation $\rho_{XX'}$ can be shown to equal the proportion of observed-score variance that is true-score variance or, equivalently, one minus the proportion of observed-score variance that is error variance. Because variances are equal for parallel tests, the result holds for both tests. We provide the result for test score X, that is,

$$\rho_{XX'} = \frac{\sigma^2_T}{\sigma^2_X} = 1 - \frac{\sigma^2_E}{\sigma^2_X}. \tag{3.1}$$

Considering Equation 3.1 for an item score produces the item-score reliability, defined as

$$\rho_{ii'} = \frac{\sigma^2_{T_i}}{\sigma^2_{X_i}} = 1 - \frac{\sigma^2_{E_i}}{\sigma^2_{X_i}}. \tag{3.2}$$

The two terms on the right-hand side of Equation 3.2 each contain an unknown. We briefly discuss three methods to approximate item-score reliability based on one test administration. Approximations to Equation 3.1 are all lower bounds, meaning they have a negative discrepancy relative to reliability (Sijtsma & Van der Ark, 2015). For Equation 3.2 the situation is less obvious. Method λ_6 appears to be a strict lower bound, but for methods MS and CA in some situations positive bias cannot be ruled out and more research is needed (Zijlmans, Van der Ark, et al., 2018) If the item response functions coincide, method MS equals the item-score reliability (Zijlmans, Van der Ark, et al., 2018); and for method CA particular choices, not to be outlined here, lead to the conclusion that items must be essentially τ-equivalent (Lord & Novick, 1968, p. 51).

Method MS. Let π_i be the marginal proportion of the population obtaining a score of 1 on item i and $\pi_{ii'}$ the marginal proportion of the population scoring a 1 on both item i and an independent replication of item i denoted by i'. For dichotomous items, Mokken (1971, p. 143) rewrote item reliability in Equation 3.2 as (right-hand side)

$$\rho_{ii'} = 1 - \frac{\pi_i - \pi_{ii'}}{\pi_i(1-\pi_i)} = \frac{\pi_{ii'} - \pi_i^2}{\pi_i(1-\pi_i)}. \tag{3.3}$$

One estimates proportion π_i from the data as the fraction of 1 scores, but for estimating $\pi_{ii'}$ one needs an independent replication of the item next to the scores on the first administration of the same item. Because independent replications are unavailable in practice, Mokken (1971, pp. 142-147) proposed two methods for approximating $\pi_{ii'}$ by deriving information not only from item i but also from the next more-difficult item $i-1$ (which has the univariate proportion $\pi_{i-1} < \pi_i$ closest to π_i), the next easier item $i+1$ (which has the univariate proportion $\pi_{i+1} > \pi_i$ closest to π_i), or both items. Mokken (1971, pp. 146–147) assumed that items $i-1$ and $i+1$ were the two items from the test that were the most similar to item i, and thus were the most likely candidates to serve as approximate replications of item i. To gain more similarity, he also required that the items in the test were consistent with the double monotonicity model, which assumes a unidimensional latent variable θ, local independence of the item scores conditional on θ, and monotone nondecreasing and nonintersecting item response functions. Estimating $\pi_{ii'}$ uses the following principle (also see Sijtsma, 1998).

Let $P_i(\theta)$ denote the item response function of item i and let $P_{i'}(\theta)$ be the item response function of a replication of item i, and notice that by definition $P_i(\theta) = P_{i'}(\theta)$. Furthermore let $G(\theta)$ denote the cumulative distribution of the latent variable θ; then

$$\pi_{ii'} = \int_\theta P_i(\theta)P_{i'}(\theta)G(\theta). \tag{3.4}$$

Next, $P_{i'}(\theta)$ in the integrand is replaced by the linear combination

$$\tilde{P}_{i'}(\theta) = a + bP_{i-1}(\theta) + cP_{i+1}(\theta),\ a,\ b \text{ and } c \text{ are constants}. \tag{3.5}$$

We refer to Mokken (1971, pp. 142-147) for the choice of the constants a, b and c. His Method 1 uses only one neighbor item to item i and his Method 2 uses both neighbor items. Let $\tilde{\pi}_{ii'}$ be an approximation to $\pi_{ii'}$ in Equation 3.3. Inserting $\tilde{P}_{i'}(\theta)$ from Equation 3.5 in the integrand of Equation 3.4 and then integrating yields

$$\tilde{\pi}_{ii'} = a + b\pi_{i-1,i} + c\pi_{i,i+1}. \tag{3.6}$$

Equation 3.6 contains only observable quantities and can be used to approximate item-score reliability in Equation 3.3 for items that adhere to the double monotonicity model. Sijtsma and Molenaar (1987) proposed method MS as an alternative to Mokken's methods 1 and 2 to obtain statistically better estimates of test-score reliability, Molenaar and Sijtsma (1988) generalized all three methods to polytomous items and Meijer et al. (1995) proposed the item-score reliability version. The method for estimating item-score reliability of polytomous items is similar to the method for dichotomous items and hence is not discussed here. Item-score reliability based on method MS for both dichotomous and polytomous items is denoted $\rho^{MS}_{ii'}$ and estimated following a procedure discussed by Zijlmans, Van der Ark, et al. (2018).

Method λ_6. Guttman (1945) proposed test-score reliability method λ_6, which

Zijlmans, Van der Ark, et al. (2018) adapted to the item-score reliability method denoted by $\rho_{ii'}^{\lambda_6}$. For this adapted method, the residual error from the multiple regression of item i on the remaining $J-1$ item scores serves as an upper bound for error variance in the item score; hence, the resulting item-score reliability is a lower bound for true item reliability. Let $\sigma_{\epsilon_i}^2$ denote the residual error of the multiple regression of item X_i on the remaining $J-1$ item scores. Method λ_6 is defined as

$$\rho_{ii'}^{\lambda_6} = 1 - \frac{\sigma_{\epsilon_i}^2}{\sigma_{X_i}^2}. \tag{3.7}$$

Method CA. Method CA is based on the correction for attenuation (Lord & Novick, 1968, pp. 69-70; Nunnally & Bernstein, 1994, p. 257; Spearman, 1904). The method correlates an item score and a test score both allegedly measuring the same attribute (Wanous & Reichers, 1996). The item score can be obtained from the same test on which the test score was based, but the test score may also refer to another test measuring the same attribute as the item. The idea is that by correlating two variables that measure the same attribute or nearly the same attribute, one approximates parallel measures; see Equation 3.2. Let $\rho_{ii'}^{CA}$ be the item-score reliability estimate based on method CA. Let $\rho_{X_iR_i}$ be the correlation between the item score and the sum score based on the other items in the test, also known as the rest score and defined as $R_i = X - X_i$. Let α_{R_i} be the reliability of the rest score, estimated by reliability lower bound coefficient α (e.g. Cronbach, 1951). Method CA estimates the item-score reliability by means of

$$\rho_{ii'}^{CA} = \frac{\rho_{X_iR_i}^2}{\alpha_{R_i}}. \tag{3.8}$$

Item Indices Currently Used in Test Construction

Well-known item-quality indices used in test construction are (a) the corrected item-total correlation, (Lord & Novick, 1968, p. 330), (b) the loading of an item on the factor which it co-defines (Harman, 1976, p. 15), in this study called the item-factor loading; (c) the item scalability (Mokken, 1971, pp. 148–153); and (d) the item discrimination (Baker & Kim, 2004, p. 4; Hambleton & Swaminathan, 1985, p. 36). For each of these four indices, rules of thumb are available in the psychometric literature that the researcher may use to interpret the values found in empirical data and make decisions about which items to maintain in the test.

Corrected item-total correlation. The corrected item-total correlation is defined as the correlation between the item score X_i and the rest score R_i, and is denoted $\rho_{X_iR_i}$. In test construction, the corrected item-total correlation is used to define the association of the item with the total score on the other items.

Higher corrected item-total correlations within a test result in a higher coefficient α (Lord & Novick, 1968, p. 331). Rules of thumb for minimally required values of corrected item-total correlations are $.20$, $.30$ or $.40$ for maximum-performance tests (also known as cognitive tests) and higher values for typical-behavior tests (also known as noncognitive tests; De Groot & Van Naerssen, 1969, pp. 252–253; Van den Brink & Mellenbergh, 1998, p. 350). The literature does not distinguish dichotomous and polytomous items for this rule of thumb and is indecisive about the precise numerical rules of thumb for typical-behavior tests. The corrected item-total correlation is also used for the estimation of item-score reliability by means of method CA (see Equation 3.8).

Item-factor loading. To obtain the item-factor loading λ_i, a one-factor model can be estimated. Because the data consist of ordered categorical scores (including dichotomous scores), polychoric correlations are used to estimate the factor loadings (Olsson, 1979). Let ξ_i^* be a latent continuous variable measuring some attribute, υ_i the intercept of item i, η the factor-score random variable, and E_i the residual-error score for item i. The i-th observation is defined as

$$\xi_i^* = \upsilon_i + \lambda_i \eta + E_i. \tag{3.9}$$

We assume a monotone relation between X_i and ξ_i^* where thresholds are used to define the relationship between X_i and ξ_i^*. For simplicity, only integer values are assigned to X_i, see Olsson (1979) for further details. Minimum item-factor loadings of $.3$ to $.4$ are most commonly recommended (Gorsuch, 1983, p. 210; Nunnally, 1978, pp. 422–423; Tabachnick & Fidell, 2007, p. 649). For this recommendation, no distinction is made between dichotomous and polytomous items.

Item Scalability. The H_i item-scalability coefficient is defined as follows (Mokken, 1971, p. 148; Sijtsma & Molenaar, 2002, p. 57; Sijtsma & Van der Ark, 2017). Let $\text{Cov}_{\text{max}}(X_i, R_i)$ be the maximum covariance and $\rho_{\max}$ the maximum correlation between item score X_i and rest score R_i, given the marginal frequencies in the $J-1$ two-dimensional cross tables for item i and each of the other $J-1$ items in the test. The H_i coefficient is defined as

$$H_i = \frac{\text{Cov}(X_i, R_i)}{\text{Cov}_{\text{max}}(X_i, R_i)}. \tag{3.10}$$

Dividing both the numerator and denominator of the ratio in Equation 3.10 by $\sigma_{R_i}\sigma_{X_i}$ results in

$$H_i = \frac{\rho_{R_i X_i}}{\rho_{R_i X_i}^{\max}}. \tag{3.11}$$

Hence, H_i can be viewed as a normed corrected item-total correlation. The H_i coefficient can attain negative and positive values. Its maximum value equals 1 and its minimum depends on the distributions of the item scores but is of little interest in practical test and questionnaire construction. Moreover, in the context of nonparametric IRT where H_i is used mostly, given the assumptions of nonparametric IRT models, only nonnegative H_i values are allowed whereas negative values are in conflict with the nonparametric IRT models. For all practical purposes, Mokken (1971, p. 184) proposed that item-scalability coefficients should be greater than some user-specified positive constant c. Items with $H_i < c$ have relatively weak discrimination and should be removed from the test. Sijtsma and Molenaar (2002, p. 36) argue that in practice items with H_i values ranging from 0 to 0.3 are not useful because they contribute little to a reliable person ordering for all types of items. Henceforth, we call the H_i item-scalability coefficient the item scalability.

Item Discrimination. Many parametric IRT models define an item-discrimination parameter. For example, the graded response model (Samejima, 1969, 1997) contains discrimination parameter α_i (not to be confused with Cronbach's coefficient α; see Equation 3.8). In addition, let δ_{ix} be the location parameter for category x $(x = 1, 2, \ldots, m)$ of item i. The graded response model is defined as

$$P(X_i \geq x \mid \theta) = \frac{\exp\left[\alpha_i(\theta - \delta_{ix})\right]}{1 + \exp\left[\alpha_i(\theta - \delta_{ix})\right]}. \tag{3.12}$$

Equation 3.12 represents the cumulative category response function, and an item scored $0, \ldots, m$ has m such functions, for $x = 1, \ldots m$. The discrimination parameter α_i is related to the steepest slope of the item's cumulative category response function. Higher α values indicate that the item better distinguishes people with respect to latent variable θ. For dichotomous items, Baker (2001, p. 34) proposed the following heuristic guidelines for discrimination parameters: $\alpha_i < .35$, very low; $0.35 \leq \alpha_i < 0.65$ low; $0.65 \leq \alpha_i < 1.35$, moderate; $1.35 \leq \alpha_i < 1.70$, high; and $\alpha_i \geq 1.70$, very high.

Several authors (e.g., Culpepper, 2013; Gustafsson, 1977; Nicewander, 2018) proposed reliability in the context of an IRT framework. The relationship of item-score reliability versions based on these proposals to discrimination parameters in several IRT models may not be clear-cut or at least rather complex. Lord (1980) argued that the relationship between item discrimination and IRT-based item-score reliability is far from simple and differs for most IRT models.

Empirical-Data Sets

We selected 16 empirical-data sets collected by means of different tests and questionnaires and representing a wide variety of attributes. In each data set, for each item we estimated item-score reliability by means of each of the three item-score reliability methods. The two goals were to compare the values of the different methods to find differences and similarities, and to derive guidelines for reasonable values to be expected in the analysis of empirical data. We also compared the values for the three item-score reliability methods with the corrected item-total correlation, the item-factor loading, the item scalability, and the item discrimination. The goal was to investigate whether the item-score reliability and the other four item indices identified the same items as weak or strong relative to the other items in a scale.

Five data sets came from tests measuring maximum performance and 11 data sets came from questionnaires measuring typical behavior. A detailed overview of the data sets can be found in Table 3.1. Table 3.2 provides a classification of the tests and questionnaires by maximum performance and typical behavior, and also by number of items and number of item scores. It was impossible for the authors to get a hold on a typical data set for each cell in Table 3.2, basically because several combinations of test properties are rare in practice. For example, maximum performance is usually measured using tests containing more than 10 dichotomously scored items, but not by means of shorter tests and rarely by means of tests containing polytomously scored items or the combination of both properties. Hence, for the maximum-performance category we were unable to find data sets with fewer than 10 items or containing polytomous item scores. For the typical-behavior category, we were unable to obtain dichotomous-item data sets with fewer than 20 items. Such data sets are expected to be rare in practice, and because they are rare we did not consider their absence damaging to the conclusions of this study. Tests and questionnaires for which we were able to obtain data sets differed with respect to number of items, number of answer categories (and number of item scores), and sample size. The adjective checklist (ACL; Gough & Heilbrun Jr., 1980) and the HEXACO personality inventory (abbreviated HEX; Ashton & Lee, 2001, 2007) contained scores from 22 and 24 subscales, respectively. We considered the ACL and the HEX different data clusters and within each cluster we analyzed the subscale data separately. The other 14 data sets all referred to a single scale, and were considered a third data cluster, denoted the various-data cluster.

Analysis

The three item-score reliability methods and the four accepted item indices were estimated for each data set. Listwise deletion was used to accommodate missing values. Within the three data clusters scatter plots were generated for each combination of the seven item indices, showing the relationship between all possible pairs of item indices.

Table 3.1: Overview of the Empirical-Data Sets

Data Set	Attribute	N	J	$m+1$	*Percentage missingness*	*Recode items*	*Reference*
1 VER	Verbal intelligence by means of verbal analogies	990	32	2	0	-	Meijer, Sijtsma, and Smid (1990)
2 BAL	Intelligence by balance scale problem-solving	484	25	2	0		Van Maanen, Been, and Sijtsma (1989)
3 CRY	Tendency to cry	705	23	2	0	-	Vingerhoets and Cornelius (2001)
4 IND	Inductive reasoning	484	43	2	1.24	-	De Koning, Sijtsma, and Hamers (2003)
5 RAK	Word comprehension	1641	60	2	0	-	Bleichrodt, Drenth, Zaal, and Resing (1985)
6 TRA	Transitive reasoning	425	12	2	0	-	Verweij et al. (1999)
7 COP	Strategies for coping with industrial malodor	828	17	4	0	-	Cavalini (1992)
8 WIL	Willingness to participate in labor union action	496	24	5	0	-	Van der Veen (1992)
9 SEN	Sensation seeking tendency	441	13	7	0	-	Van den Berg (1992)
10 DS14	Type D personality	541	14	5	0.13	1, 3	Denollet (2005)
11 TMA	Taylor Manifext Anxiety Scale	5410	50	2	0.97	1 - 3 - 4 9 - 12 - 18 - 20 - 29 32 - 38 - 50	Taylor (1953)
12 LON	Loneliness	7440	11	3	0.58	1 - 4 - 7 - 8 - 11	De Jong Gierveld and Van Tilburg (1999)
13 SAT	Satisfaction with life	7423	4	5	0.43	-	Diener, Emmons, Larsen, and Griffin (1985) Pavot and Diener (1993)
14 SES	Rosenberg Self-Esteem Scale	47974	10	4	0.43	3 - 5 - 8 - 9 - 10	Rosenberg (1965)
15 ACL	Personality traits	433	218	6	0	-	Gough and Heilbrun Jr. (1980)
16 HEX	HEXACO Personality Inventory	22786	240	8	< 0.01 (129 cases)	-	Ashton and Lee (2001) Ashton and Lee (2007)

Table 3.2: Overview of the Empirical-Data Sets Arranged by Number of Items and Number of Item Scores

	Maximum performance		Typical Behavior	
No. of items	No. of answer categories		No. of answer categories	
	2	> 2	2	> 2
≤ 10				SAT SES ACL HEX
$10 < J < 20$	TRA			COP SEN DS14 LON
≥ 20	VER BAL IND RAK		CRY TMA	WIL

Note. See Table 3.1 for the descriptions of the data sets.

The three item-score reliability methods use different approaches, but are all intended to approximate true item-score reliability in Equation 3.2. Hence, we were interested to know the degree to which the three methods produced the same numerical values. Numerical identity was expressed by means of the coefficient of identity (Zegers & Ten Berge, 1985), which runs from -1 to 1, with higher positive values meaning that the values of the two indices studied are more alike, and the value 1 meaning that they are numerically identical. The product-moment correlation provides identity up to a linear transformation, thus it does not provide the exact information we were interested in but it was also given because it is well known and provides approximately, albeit not precisely, the information required. When assessing the relationship between an item-score reliability method and each of the other four item indices or among the latter four indices, one needs to realize that indices in each pair estimate a different parameter. Hence, in considering the degree to which two different indices suggest item quality is in the same direction, an ordinal association measure is sufficient. We used Kendall's τ to express this association, and even though it was not quite optimal for our purposes, we provided the product-moment correlation for completeness.

To investigate what values can be expected for the item-score reliability methods at the cutoff values for the other item indices, we regressed each of the three item-score reliability methods on each of the four item indices, thus, producing 12 bivariate regression equations. This enabled us to estimate the item-score reliability at the cutoff value of the item index ($.3$ for corrected item-total correlation, $.3$ for item-factor loading, $.3$ for item scalability, and $.7$ for item discrimination), for every combination of item-score reliability method and item index giving an indication of what a good cutoff value would be for the values estimated by the item-score reliability methods.

For estimating the item-score reliability methods, R code (R Core Team, 2016) was used, which was also employed by Zijlmans, Van der Ark, et al. (2018).

The package lavaan (Rosseel, 2012) was used for estimating the item-factor loadings, the package mokken was used for estimating the H_i coefficient (Van der Ark, 2007, 2012), and the package ltm was used for estimating the discrimination parameters (Rizopoulos, 2006) using the two-parameter logistic model for dichotomous data and the graded response model for polytomous data.

3.3 Results

For method MS, the values of the item-score reliability estimates ranged from $.00$ to $.70$ (mean $.29$), for method λ_6, values ranged from $.03$ to $.81$ (mean $.34$), and for method CA, values ranged from $.00$ to $.90$ (mean $.30$). For the three data clusters, Figure 3.1 shows the scatter plots for pairs of item-score reliability methods. The identity coefficient for all pairs of item-score reliability methods exceeded $.9$. The plots show more scatter for the various-data cluster. For the ACL and HEX data clusters, the scatter shows stronger association. In all three data sets, in many cases method λ_6 had higher values than methods MS and CA. Product-moment correlations between item-score reliability methods were higher than $.70$ for all combinations and all data clusters. In the HEX data cluster correlations exceeded $.80$.

Figure 3.2 shows the scatterplots comparing corrected item-total correlation with the three item-score reliability methods. Method CA produced positive values when corrected item-total correlations were negative. The positive values resulted from squaring the corrected item-total correlation, see Equation 3.8. Kendall's τ exceeded $.87$ for corrected item-total correlation and method CA in all three data clusters, while the other two item-score reliability methods showed lower values for Kendall's τ, with a maximum of $.75$. Corrected item-total correlations correlated highly with item-score reliability values in the ACL and HEX data clusters, but lower in the various-data cluster.

Figure 3.3 shows the relationship between the item-factor loadings and the three item-score reliability methods. Because most of the scatter lies above the 45-degree line, in many cases the item-factor loading was higher than the three item-score reliability estimates. In the ACL and HEX data clusters, Kendall's τ was highest between item-factor loading and method λ_6 (> 0.78). In the various-data cluster, Kendall's τ was highest, equaling $.63$, between the item-factor loading and method CA. In the HEX data cluster, the correlation between item-factor loading and item-score reliability methods was highest, followed by the ACL data cluster. The various-data cluster showed the lowest correlations between item-factor loading and item-score reliability methods.

Figure 3.4 shows the relationship between item scalability H_i and the three

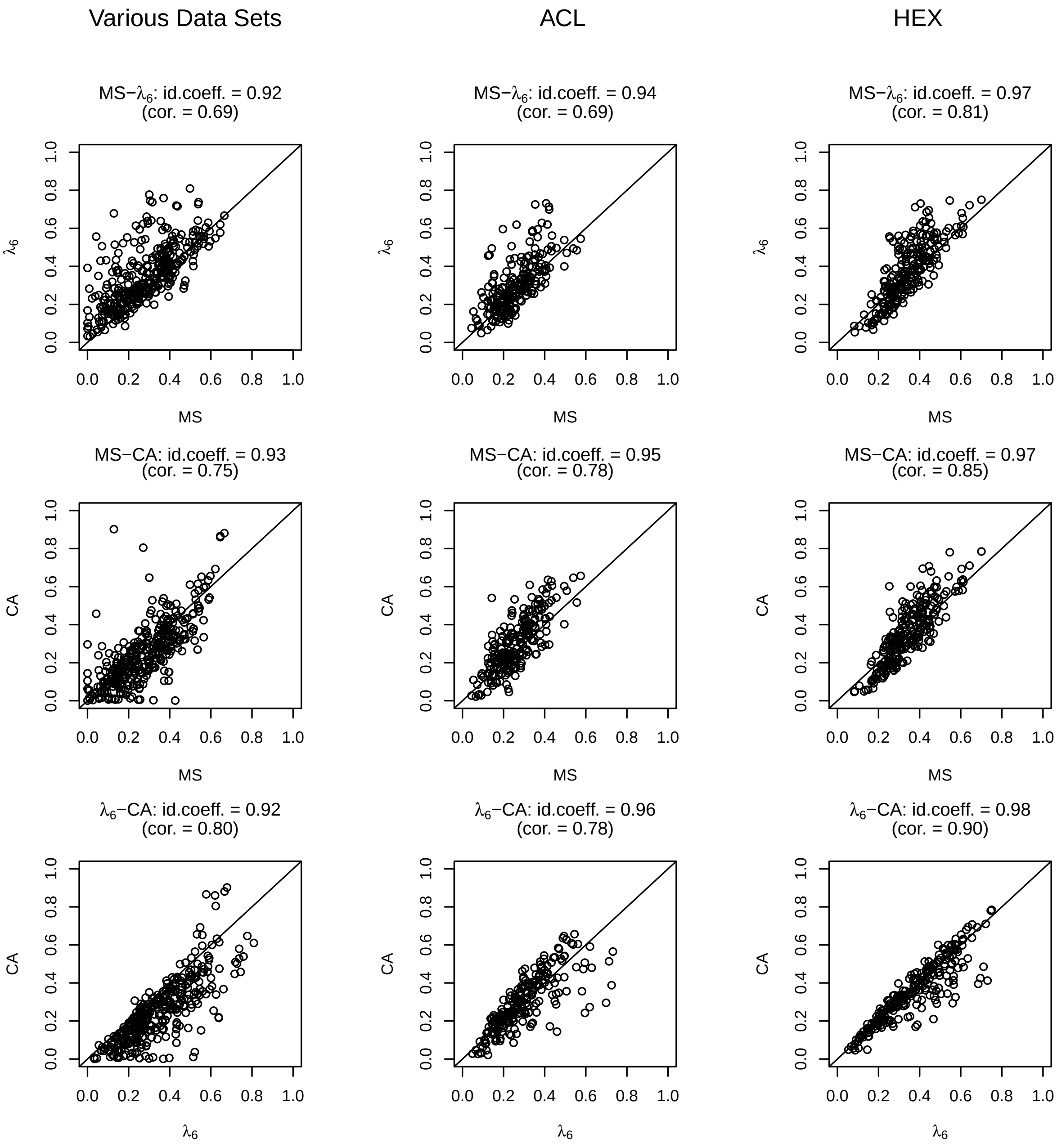

Figure 3.1: Scatter plots for the three data clusters, comparing the item-score reliability estimates for methods MS, λ_6, and CA.

Note: id. coeff. = identity coefficient, cor = correlation between two method estimates. See Table 3.1 for a description of the data sets.

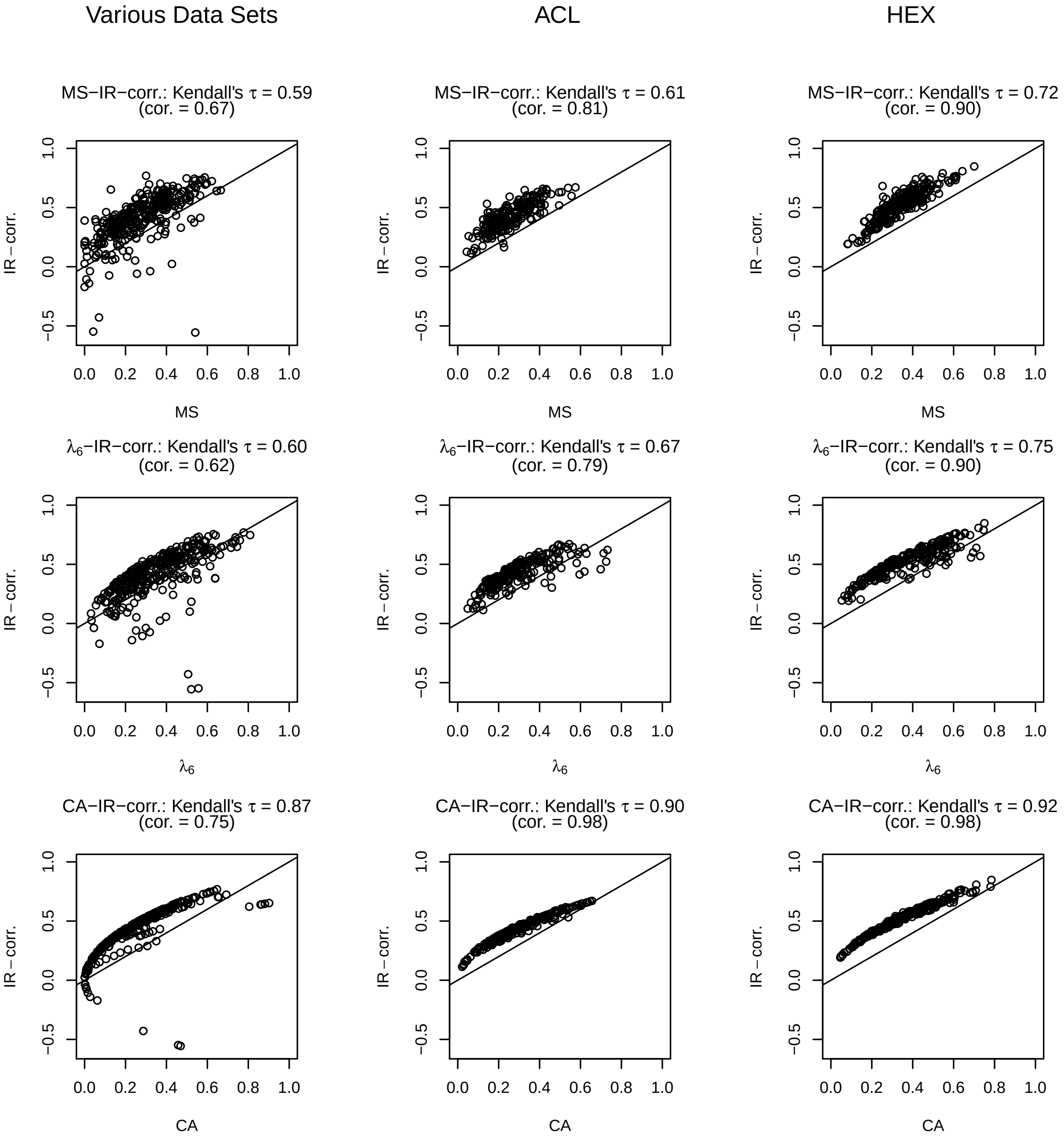

Figure 3.2: Scatter plots for the three data clusters comparing the item-score reliability methods with the corrected item-total correlation (IR-corr.).

Note: cor = correlation between two method estimates. See Table 3.1 for a description of the data sets.

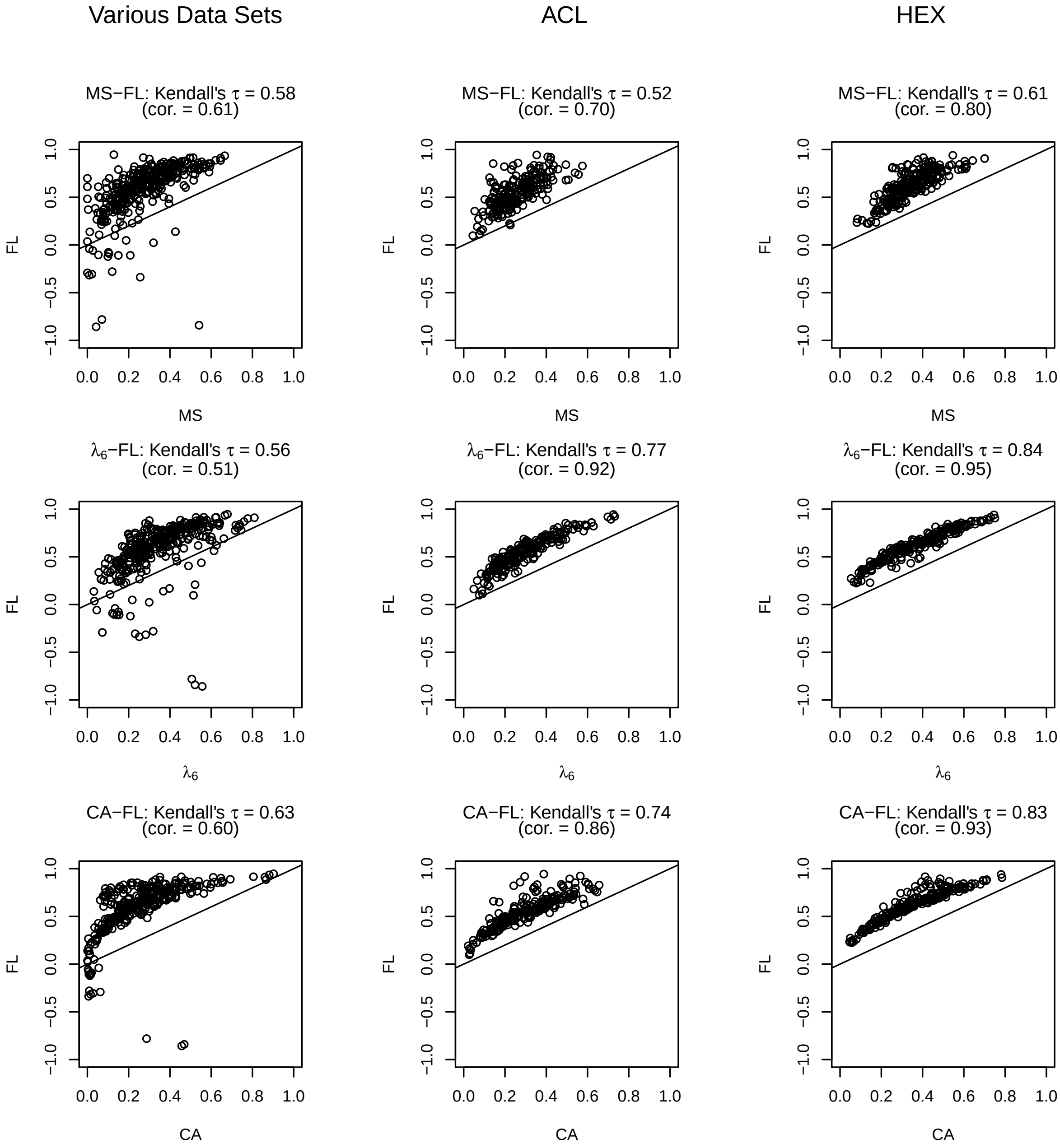

Figure 3.3: Scatter plots for the three data clusters comparing the item-score reliability methods with the item-factor loading (FL).

Note: cor = correlation between two method estimates. See Table 3.1 for a description of the data sets.

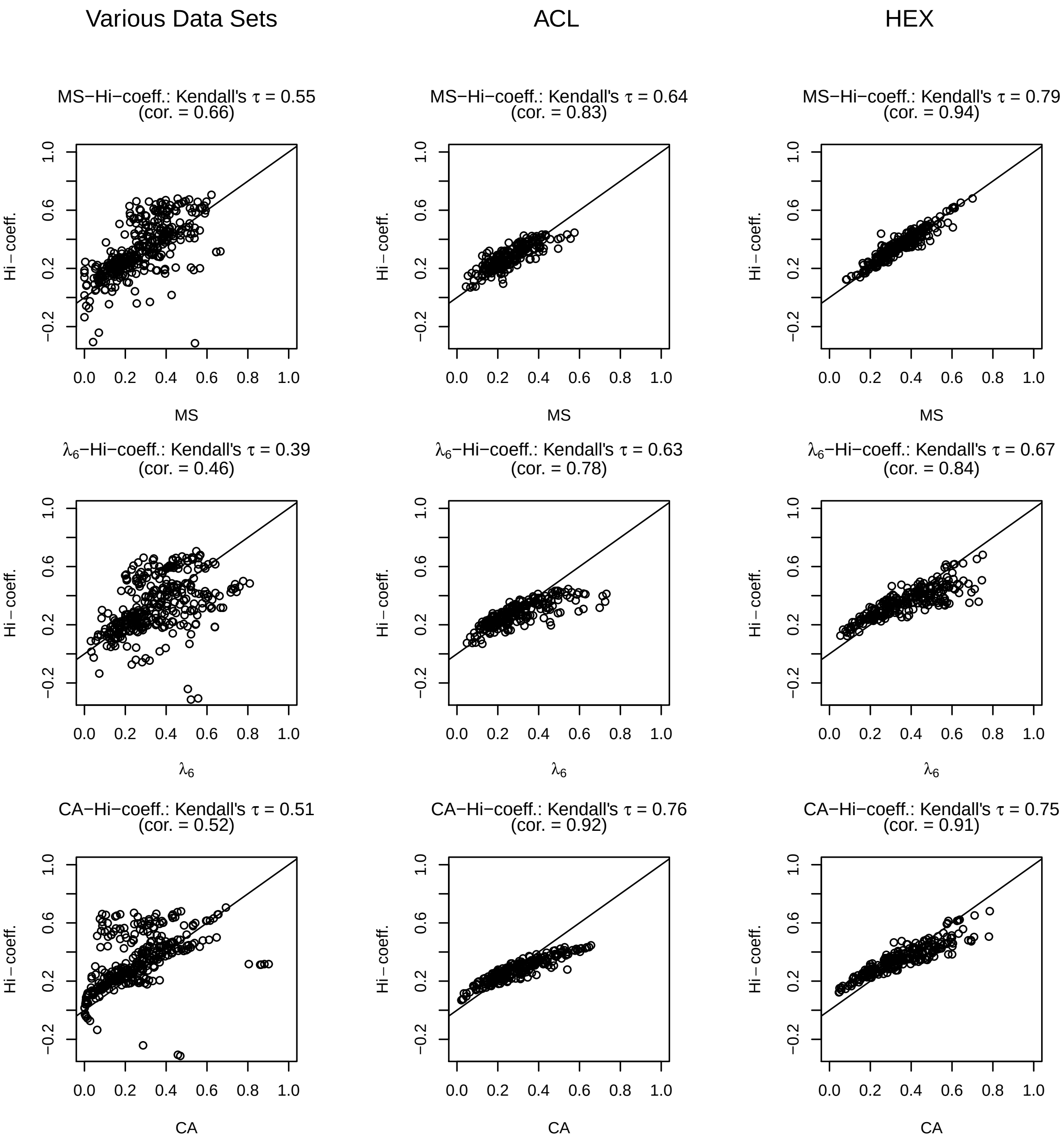

Figure 3.4: Scatter plots for the three data clusters comparing the item-score reliability methods with the H_i-coefficient (H_i-coeff.).
Note: cor = correlation between two method estimates. See Table 3.1 for a description of the data sets.

item-score reliability methods. Negative H_i values corresponded with positive CA values, resulting in scatter similar to Figure 3.2. In the various-data cluster, Kendall's τ was lower and the scatter showed more spread than in the ACL and HEX data clusters, where Kendall's τ showed higher values in excess of 0.63. In the various-data cluster, correlations between H_i values and the three reliability methods were relatively low, ranging from $.46$ to $.66$. In the ACL and HEX data clusters correlations were higher, ranging from $.78$ to $.94$.

Table 3.3: Estimates of the Three Item-Score Reliability Methods by the Four Other Item Indices using a Bivariate Regression Analysis

	Method MS	Method λ_6	Method CA
Corrected item-total correlation	0.20	0.24	0.17
Item-factor loading	0.18	0.20	0.15
H_i-coefficient	0.28	0.33	0.28
Item discrimination	0.22	0.25	0.20

Figure 3.5 shows the relationship between item discrimination and the three item-score reliability methods. A discrimination value equal to 10.77 in data set RAK was assessed to be an outlier and was removed from the scatter plot. The next largest discrimination value in this data cluster was 5.7 and the mean estimated discrimination was 1.5. Kendall's τ between discrimination and CA values was highest for the ACL and HEX data clusters. Kendall's τ between item discrimination and MS values was lowest, with values of $.53$, $.51$ and $.59$ for the various-data cluster, the ACL data cluster, and the HEX data cluster, respectively. The correlation between item discrimination and item-score reliability was lower in the various-data cluster than in the ACL and HEX data clusters. In the various-data cluster, correlations ranged from $.49$ to $.60$, and in the ACL and HEX data clusters correlations ranged from $.67$ to $.90$.

Figure 3.6 shows the relationship between corrected item-total correlation, item-factor loading, item scalability, and item discrimination. Kendall's τ was highest between item discrimination and item-factor loading in the ACL and HEX data clusters. In these data clusters, correlations were high for the four accepted item indices. Corrected item-total correlation and item-factor loading correlated higher than $.9$ in all three clusters. In the ACL and HEX data clusters, corrected item-total correlation and item scalability also correlated higher than $.9$.

Table 3.3 provides the results for the bivariate regression estimating the three item-score reliability coefficients by the cutoff values of four other item indices. The item-factor loading estimated the lowest item-score reliability values: $.18$ for method MS, $.20$ for method λ_6, and $.15$ for method CA. The H_i coefficient estimated the highest item-score reliability values: $.28$ for method MS, $.33$ for method λ_6, and $.28$ for method CA.

3.4 Discussion

We estimated item-score reliability methods MS, λ_6, and CA in various empirical-data sets, and investigated which values the researcher may expect to find in his

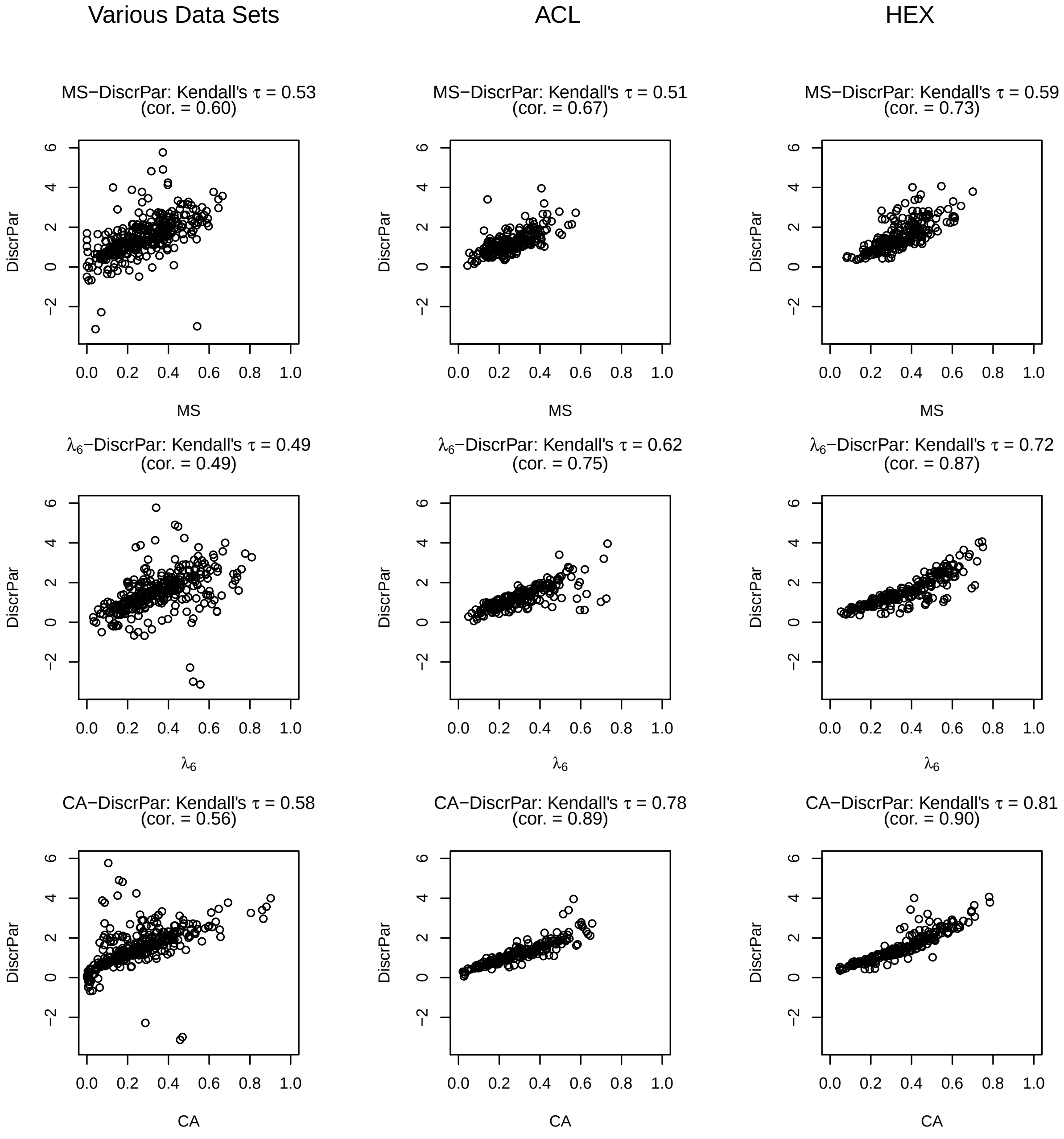

Figure 3.5: Scatter plots for the three data clusters comparing the item-score reliability methods with the discrimination parameter (DiscrPar).

Note: cor = correlation between two method estimates. See Table 3.1 for a description of the data sets.

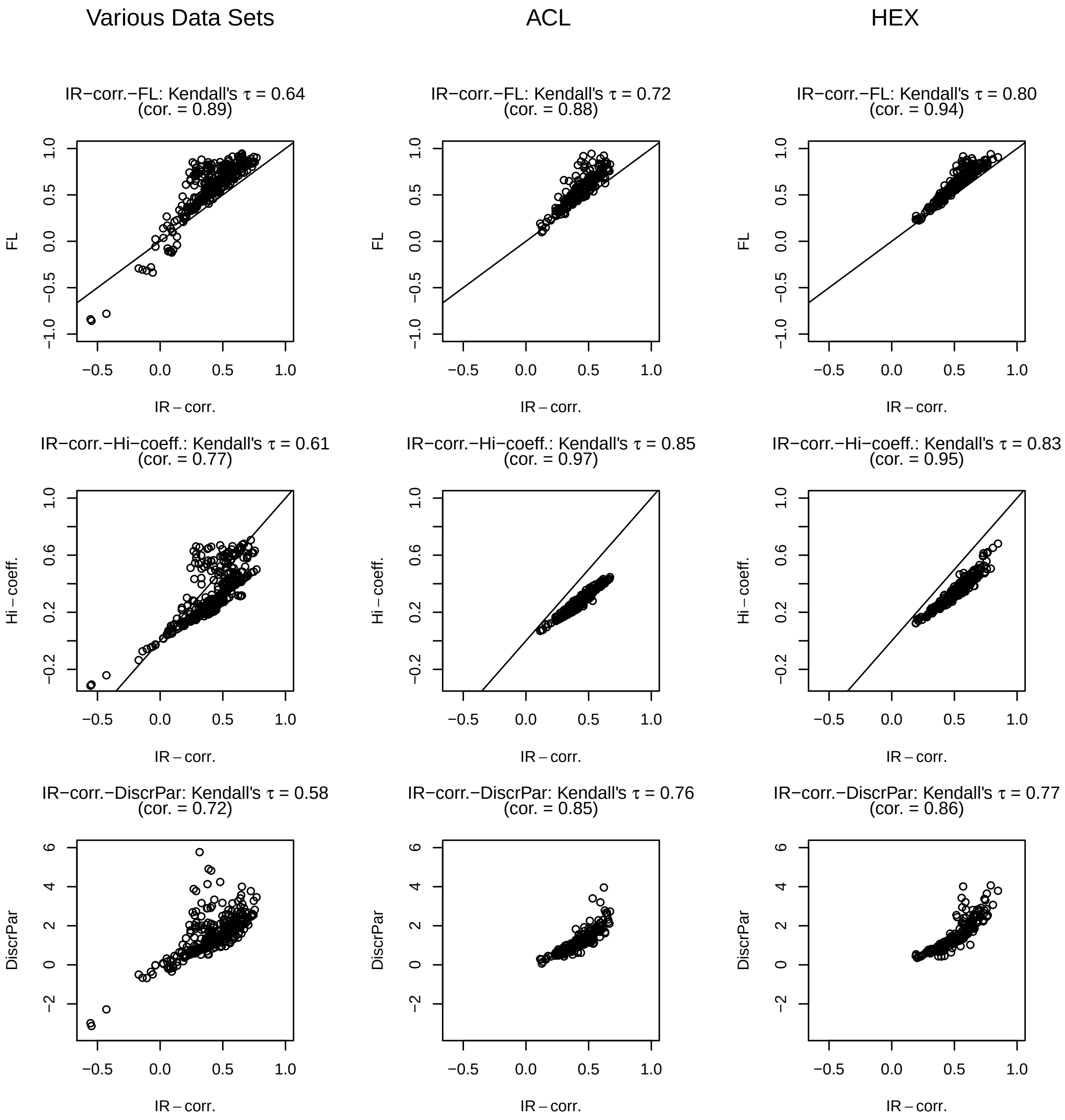

Figure 3.6: Scatter plots for three data clusters comparing corrected item-total correlation, item-factor loading, the H_i coefficient, and the discrimination parameter. See Table 3.1 for description of data sets.
Note: cor = correlation between two method estimates.

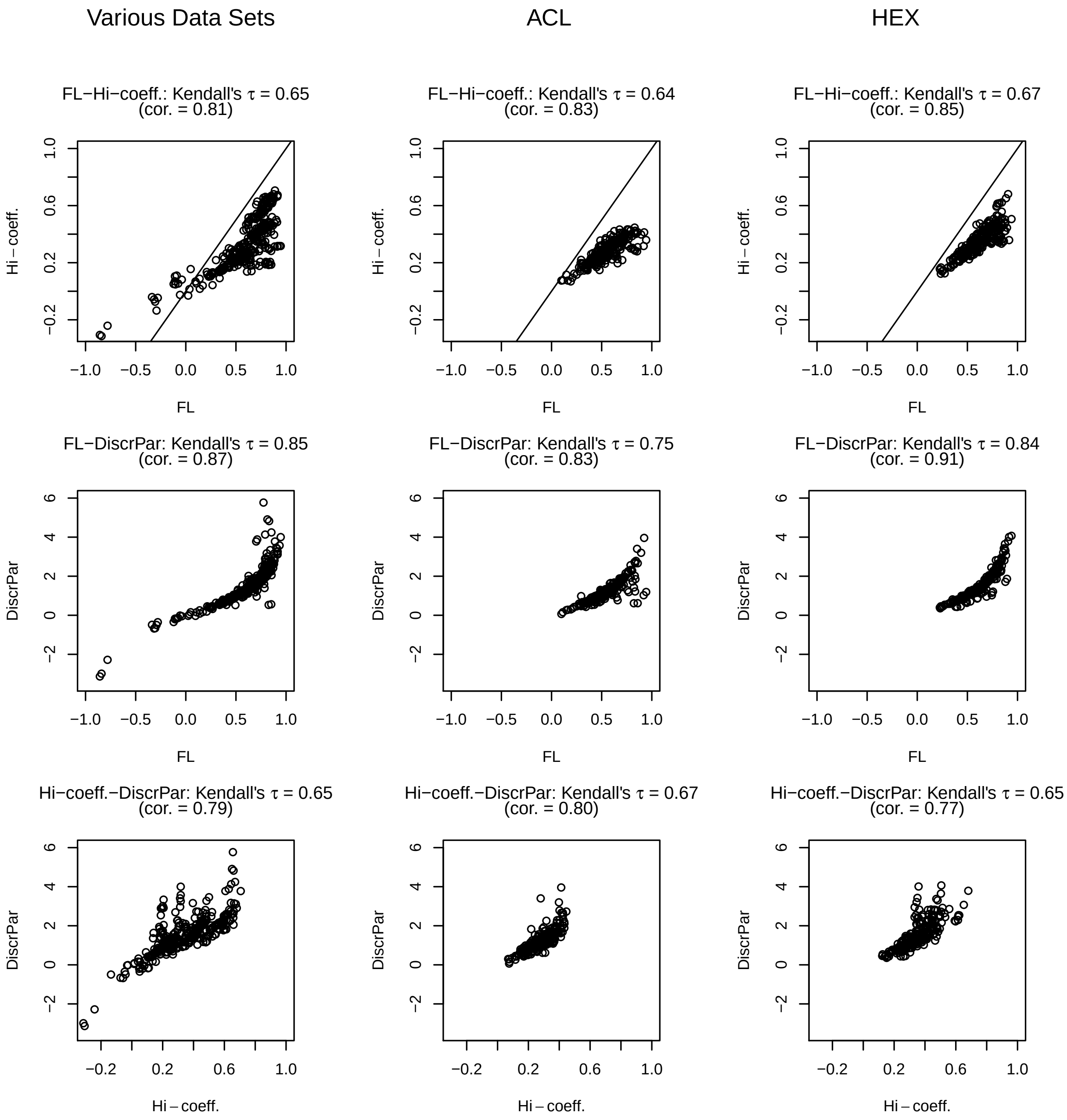

Figure 3.6, continued: Scatter plots for three data clusters comparing corrected item-total correlation, item-factor loading, the H_i coefficient, and the discrimination parameter. See Table 3.1 for description of data sets.

Note: cor = correlation between two method estimates.

empirical-data set. The identity-coefficient values between the three item-score reliability methods were all higher than $.9$. The product-moment correlations between the three item-score reliability methods yielded values in excess of $.7$. Identity values in excess of $.9$ suggest that the uniformed versions of the three item-score reliability methods yielded nearly identical values, suggesting a high degree of interchangeability of methods for item selection. We conclude that in practice the three item-score reliability methods can be used interchangeably. The three item-score reliability methods have the same computing time, but methods λ_6 and CA are much simpler to program.

The relationships between the three item-score reliability methods and the four accepted item indices showed there is a strong association between the corrected item-total correlation and the item-score reliability methods, especially method CA. This result can be explained by the relation between method CA and the corrected item-total correlation (Equation 3.8). The other associations between the item-score reliability methods and the other item indices are weaker. For the other four item indices, the researcher can use available rules of thumb to decide when an item is a candidate for revision or for elimination from a test. Based on investigating a polytomous single-item measure with five response categories, Wanous et al. (1997) suggested using a lower bound of $.7$ for the item-score reliability. Given the values that were obtained for the items in the empirical-data sets we selected, and given the results from the bivariate linear regression, we conjecture that this requirement may be too stringent in practice: Instead, a value of $.3$ would be a realistic lower bound for item-score reliability.

We found that λ_6 values often exceeded MS and CA values. In a simulation study, Zijlmans, Van der Ark, et al. (2018) found that for many conditions in the experimental design, method λ_6 underestimated the true item-score reliability whereas methods MS and CA were almost unbiased, which seems to contradict the results of the present study. An explanation may be that our data sets do not fit in any of the experimental conditions Zijlmans, Van der Ark, et al. (2018) investigated, making a comparison between the two studies awkward. Our data sets were multidimensional, with relatively large numbers of items that had a considerable variation in discrimination. Zijlmans, Van der Ark, et al. (2018) studied the factors dimensionality, variation in discrimination within a test, and test length separately, and found that for the multidimensional data, for unequal discrimination, and for many items, the differences between methods MS, λ_6, and CA were either absent or less clear than in other experimental conditions. Hence, a combination of these factors may have caused the relatively high λ_6 values in the present study. In future research, these conditions, which are realistic for most data sets, should be studied further in a fully crossed simulation design.

Values we found for accepted item indices in empirical data could serve as a starting point for a simulation study that further investigates the relationship between item-score reliability and accepted item indices. Furthermore, little knowledge about the relation between item-score reliability and test-score reliability is available, rendering the investigation of this relationship urgent. Also, the effect of omitting items with low item-score reliability on the total-score reliability should be investigated.

Chapter 4

Investigating the Relationship between Item-Score Reliability, its Estimation Methods, and Other Item Indices

Abstract

Reliability is usually estimated for a test score, but it can also be estimated for item scores. A higher item-score reliability (denoted $\rho_{ii'}$) indicates a higher degree of repeatability of the item score and is therefore important when evaluating the quality of an item. Based on earlier research, in the current study three methods were further investigated for estimating item-score reliability: method MS, Guttman's method λ_6, and the correction for attenuation (method CA). The goal of the research was to further investigate, under various conditions (1) the relationship between $\rho_{ii'}$ and the three item-score reliability methods, (2) the relationship between $\rho_{ii'}$ and four other item indices, and (3) the feasability of a lower bound for item-score reliability estimates of $.3$. This was done by means of a simulation study where the item's difficulty parameter, variance of the other items' location parameters, and number of items in the test are varied. All methods showed increasing bias for higher values of $\rho_{ii'}$. Method CA showed good results for items with a non-deviant location parameter. There seems to be a one-to-one relationship between the item-factor loading and $\rho_{ii'}$. A lower bound of $.3$ seems to be too stringent in practice.

Keywords: coefficient λ_6, corrected item total-correlation, correction for attenuation, item discrimination, item-factor loading, item scalability, item-score reliability

This chapter has been submitted for publication as Zijlmans, E. A. O., Tijmstra, J., Van der Ark, L. A., and Sijtsma, K. (2018). Investigating the Relationship between Item-Score Reliability, its Estimation Methods, and Other Item Indices.

4.1 Introduction

This article investigates the relationship between item-score reliability, methods to approximate item-score reliability, and four other item-level indices. Item-score reliability expresses the repeatability of an individual item score. A higher item-score reliability indicates a higher degree of repeatability of the item score and is therefore important when evaluating the quality of an item. Just as test-score reliability is important for evaluating the quality of a test, item-score reliabilty is important for evaluating the quality of an item. Reliability for individual item scores is used in applied psychology to assess one-item measures for job satisfaction (Gonzalez-Mulé et al., 2017; Harter et al., 2002; Nagy, 2002; Robertson & Kee, 2017; Saari & Judge, 2004; Zapf et al., 1999) and burnout level (Dolan et al., 2014). Item-score reliability is also used in health research for measuring, for example, quality of life (Stewart et al., 1988; Yohannes et al., 2010) and psychosocial stress (Littman et al., 2006), and one-item measures have been assessed in marketing research for measuring ad and brand attitude (Bergkvist & Rossiter, 2007).

Items with a low item-score reliability might not contribute much to the test-score reliability, and could therefore be candidates for revision or for elimination from the test. Item-score reliability is denoted $\rho_{ii'}$ and defined similarly to the test-score reliability as the correlation between two independent replications of the same item in the same group of people (Lord & Novick, 1968, pp. 47–50, 61). These replications are unavailable in practice, because respondents answering the same item a second time will remember what they answered the first time, and this will influence their answer on the second occasion. Therefore, item-score reliability has to be estimated from the data collected in one test administration.

Zijlmans, Van der Ark, et al. (2018) identified three promising methods to estimate item-score reliability: method MS, method λ_6, and method CA. Each of these three methods attempts to approximate the item-score reliability based on data from a single test-administration. As will be elaborated in the method section, these different methods make use of different assumptions and thus can result in different estimates of the item-score reliability $\rho_{ii'}$. Zijlmans, Van der Ark, et al. (2018) investigated the three methods under realistic simulation conditions, and found that generally all three methods underestimated item-score reliability. However, in these conditions, parameter $\rho_{ii'}$ was not manipulated, meaning that the relationship between item-score reliability and the item-score reliability methods was not systematically investigated for a wide range of values for $\rho_{ii'}$.

In a second study, Zijlmans, Tijmstra, et al. (2018b) investigated which estimates can be expected of the three item-score reliability methods in empirical-data sets, and what the observed relationship is between the three methods and four more commonly used item indices: corrected item total-correlation, item-factor loading, item scalability, and item discrimination. The relationship between $\rho_{ii'}$ and the four other item indices was not investigated, because $\rho_{ii'}$ was unknown for the empirical-data sets. Based on the reliability estimates based on methods MS, λ_6, and CA found in empirical-data sets and cutoff scores for the four other item indices described in the literature, Zijlmans, Tijmstra, et al. (2018b) tentatively proposed a lower bound of $.3$ for the three item-score reliability methods. However, the authors did not investigate the values of $\rho_{ii'}$ one may expect when a cutoff score of $.3$ is used for the item-score reliability methods. Also, it was not investigated which values the item-score reliability methods estimate at $\rho_{ii'} = .3$. Therefore, we investigated whether the cutoff score of $.3$ for the three methods indicates a sufficiently high item-score reliability for practical use of the item.

In this study, three research questions were investigated: (1) What is the relationship between $\rho_{ii'}$ and the three item-score reliability methods MS, λ_6, and CA under various testing conditions and for a range of $\rho_{ii'}$ values? (2) What is the relationship between $\rho_{ii'}$ and the four item indices under various testing conditions and for a range of $\rho_{ii'}$ values? (3) When following the suggested lower bound of $.3$ for methods MS, λ_6, and CA, what are corresponding $\rho_{ii'}$ values? Also, if $\rho_{ii'} = .3$, what are the values of the methods MS, λ_6, and CA?

This article is organized as follows. First, we discuss item-score reliability, methods MS, λ_6, and CA to estimate item-score reliability, and corrected item total-correlation, item-factor loading, item scalability, and item discrimination. Second, we discuss the various testing conditions for which the data were generated, and we explain the data-generating process. Third, we discuss the results and their implications for estimating item-score reliability.

4.2 Method

Reliability and Item-Score Reliability Methods

The following definitions (Lord & Novick, 1968, p. 61) were used. Let X_i be the item score indexed i $(i = 1 \ldots, J)$, and let X be the test score, defined as the sum of the J item scores; that is, $X = \sum_{i=1}^{J} X_i$. Item-score reliability $\rho_{ii'}$ is defined as the correlation between two replications of item score X_i, and can be shown to equal the proportion of observed-score variance ($\sigma^2_{X_i}$) that is true-score variance ($\sigma^2_{T_i}$) or, equivalently, one minus the proportion of observed item-score variance that is error variance ($\sigma^2_{E_i}$); that is,

$$\rho_{ii'} = \frac{\sigma^2_{T_i}}{\sigma^2_{X_i}} = 1 - \frac{\sigma^2_{E_i}}{\sigma^2_{X_i}}. \tag{4.1}$$

Three methods to approximate item-score reliability were investigated: method MS, method λ_6, and method CA. These methods are briefly discussed here; see Zijlmans, Van der Ark, et al. (2018) for details. Next to the three item-score reliability methods, four item indices are briefly discussed: corrected item total-correlation, item-factor loading, item scalability, and item discrimination.

Method MS is based on the Molenaar-Sijtsma reliability method (Sijtsma & Molenaar, 1987) which estimates test-score reliability. The method relies on the double monotonicity model for dichotomous items (Mokken, 1971), which assumes a unidimensional latent variable θ, local independence of the item scores conditional on θ, and monotone nondecreasing and nonintersecting item-response functions. Because independent replications of item scores are unavailable in practical research, Mokken (1971, pp. 142-147) proposed two methods for approximating independent replications by deriving information not only from item i but also from the next more-difficult item $i-1$, the next-easier item $i+1$, or both items. He assumed that, because they were adjacent to item i, items $i-1$ and $i+1$ were the two items from the test that were the most similar to item i, and thus were the most likely candidates to serve as approximate replications of item i. In this study, we denote the item-score reliability based on method MS by $\rho^{MS}_{ii'}$, and estimate $\rho^{MS}_{ii'}$ using the method outlined in Van der Ark (2010).

Zijlmans, Van der Ark, et al. (2018) adapted test-score reliability method λ_6 from Guttman (1945) to the item-score reliability method $\rho^{\lambda_6}_{ii'}$. This adapted method estimates item-score reliability by subtracting the ratio of the residual error from the multiple regression of item i on the remaining $J-1$ item scores, denoted $\sigma^2_{\epsilon_i}$, and the item variance from unity. Method λ_6 estimates item-score reliability by means of

$$\rho^{\lambda_6}_{ii'} = 1 - \frac{\sigma^2_{\epsilon_i}}{\sigma^2_{X_i}}. \tag{4.2}$$

Method CA is based on the correction for attenuation (Lord & Novick, 1968, pp. 69-70; Nunnally & Bernstein, 1994, p. 257; Spearman, 1904). The method correlates an item score and a test score both assumed to measure the same attribute (Wanous & Reichers, 1996). The item score can be obtained from the same test on which the test score was based, but the test score may also refer to another test allegedly measuring the same attribute as the item. Let $\rho^{CA}_{ii'}$ be the item-score reliability estimate based on method CA. Let $\rho_{X_iR_{(i)}}$ be the correlation between the item score and the sum score based on the other $J-1$ items in the test, also known as the rest score and defined as $R_{(i)} = X - X_i$. Let $\alpha_{R_{(i)}}$ be the reliability of the rest score, estimated by reliability lower bound coefficient α (e.g., Cronbach, 1951).

Method CA is defined as

$$\rho_{ii'}^{CA} = \frac{\rho^2_{X_i R_{(i)}}}{\alpha_{R_{(i)}}}. \tag{4.3}$$

Other Item Indices

Well-known item-functioning indices used in test construction are (1) the corrected item total-correlation (Lord & Novick, 1968, p. 330), (2) the loading of an item on the factor which it co-defines (Harman, 1976, p. 15), in this study called the item-factor loading; (3) the item scalability (Mokken, 1971, pp. 148-153); and (4) the item discrimination (Baker & Kim, 2004, p. 4; Hambleton & Swaminathan, 1985, p. 36). The four item indices are discussed briefly.

The corrected item total-correlation is defined as the correlation between item score X_i and rest score $R_{(i)}$, and is denoted $\rho_{X_i R_{(i)}}$. In test construction, the corrected item total-correlation is used to capture the association of the item with the other items. Higher corrected item total-correlations in a test result in a higher value of coefficient α (Lord & Novick, 1968, p. 331).

To obtain the item-factor loading λ_i, a one-factor model can be estimated. Let υ_i be the intercept of item i, η the factor-score random variable, and E_i the residual-error score for item i. The i-th item score is defined as

$$X_i = \upsilon_i + \lambda_i \eta + E_i. \tag{4.4}$$

When the data consist of ordered categorical scores, polychoric correlations are used to estimate the item-factor loading; see Olsson (1979) for further details.

The H_i item-scalability coefficient is defined as follows (Mokken, 1971; Sijtsma & Molenaar, 2002, p. 57; Sijtsma & Van der Ark, 2017). Let $\text{Cov}_{\text{max}}(X_i, R_i)$ be the maximum possible covariance between item score X_i and rest score R_i, given the marginal frequencies in the $J-1$ cross tables for item i and each of the other $J-1$ items in the test. The H_i coefficient is defined as

$$H_i = \frac{\text{Cov}(X_i, R_i)}{\text{Cov}_{\text{max}}(X_i, R_i)}. \tag{4.5}$$

The H_i coefficient can attain negative and positive values with a maximum equal to 1. The minimum H_i value is negative and depends on the distributions of the item scores, but is of little interest in practical test and questionnaire construction. Moreover, in the context of nonparametric IRT where H_i is used, given the assumptions of nonparametric IRT models only nonnegative H_i values are allowed whereas negative values are in conflict with the nonparametric IRT models. Henceforth, we call the H_i item-scalability coefficient the item scalability.

Many parametric IRT models define an item-discrimination parameter. For example, the two-parameter logistic model (Birnbaum, 1968) contains discrimination parameter a_i. In addition, let b_i be the location parameter of item i. The two-parameter logistic model is defined as

$$P(X_i = 1 \mid \theta) = \frac{\exp\left[a_i(\theta - b_i)\right]}{1 + \exp\left[a_i(\theta - b_i)\right]}. \tag{4.6}$$

The discrimination parameter a_i is associated with the steepest slope of the item response function, located at b_i. Higher a_i values indicate that the item better distinguishes people with respect to latent variable θ relative to b_i.

Simulation Study

We investigated whether the following three item and test characteristics influence the relationship between $\rho_{ii'}$ on the one hand, and the three item-score reliability methods and the four item indices on the other hand: (1) the difficulty of item i, (2) the variance of the discrimination parameters of the $J - 1$ items in the test other than the item of interest, and (3) the test length. Item difficulty was considered, because the bias of the item-score reliability methods with respect to $\rho_{ii'}$ may be influenced by the location of the item, dependent on its location with respect to the other items in the test. Because all methods use the other items in the test to approximate $\rho_{ii'}$, the item's location relevant to the other items' locations may influence the estimation. Variance of the discrimination parameters of the items in the test other than the item of interest may play a role for method MS, because in logistic models that we use for generating item scores non-zero variance violates the assumption of nonintersecting item response functions. We expect this violation to influence the estimation of the independent replication method MS tries to approximate. Test length can be expected to influence the relationship between methods λ_6 and CA on the one hand and $\rho_{ii'}$ on the other hand, because methods λ_6 and CA use all $J - 1$ items other than the item of interest to estimate $\rho_{ii'}^{\lambda_6}$ and $\rho_{ii'}^{CA}$, respectively.

For each of the three design factors two levels were considered: item difficulty of either 0 or 1.5 for the item of interest; equal or unequal discrimination parameters for the items in the test; 5 or 25 items other than the item of interest. This resulted in a $2 \times 2 \times 2$ full-factorial design.

For each condition, for a specific value of $\rho_{ii'}$, we investigated the estimated values of methods MS, λ_6, and CA, and corrected item total-correlation, item-factor loading, item scalability, and item discrimination. To cover the full range of realistic $\rho_{ii'}$ values, 41 different values of $\rho_{ii'}$ were used to generate data ($\rho_{ii'} = 0, 0.02, \ldots, 0.80$), while keeping all other item and test characteristics fixed. This way, the relationship between $\rho_{ii'}$, the three item-score reliability methods and the four item-functioning indices could be investigated for each of the eight conditions separately. We used the two-parameter logistic model (Equation 4.6) to generate for all J items data.

For the items in the test other than the item of interest, the following parameter values were used: $b_i = -1, -0.5, 0, 0.5, 1$ and $a_i = 1$ for the equal-discrimination condition and $a_i = 1.5, .5, 1.5, .5, 1.5$ for the unequal-discrimination condition. For the long-test condition ($J = 25$), five copies of the item sets including these parameters were used.

We generated data based on a specific value of $\rho_{ii'}$ for the item of interest. Because generating item scores based on $\rho_{ii'}$ is infeasible, we had to find values for the parameters in Equation 4.6 that would result in the desired value of $\rho_{ii'}$. We chose a standard normal distribution for θ and a starting value for a_i. For $\rho_{ii'} = 0$, we used $a_i = 0$ and after trying different values, we chose $a_i = .18$ as a starting value for the next item-score reliability value, $\rho_{ii'} = .02$. The value for b_i was either 0 or 1.5, depending on the simulation condition. Using these parameter values, the following procedure was followed: Step (1) We generated two sets of data, each having a sample size of one million data records, and calculated the population item-score reliability as the correlation between the scores for item i in the two datasets, denoted $r_{ii'}$. Step (2) The value of $r_{ii'}$ resulting from Step 1 was compared to the desired value for $\rho_{ii'}$. If the desired value was too low, we increased a_i with $.05$ and repeated Step 1. If the $r_{ii'}$ value was equal to the desired value of $\rho_{ii'}$ up to the third decimal place, the procedure continued to Step 3. Step (3) The value of a_i resulting in the desired value of $\rho_{ii'}$ was saved and the procedure re-started at Step 1 for the next value of $\rho_{ii'}$. This way, the relevant values for a_i were determined for each value of $\rho_{ii'}$ in both the condition where $b_i = 0$ and where $b_i = 1.5$. Using the saved values for a_i in Step 3, the item scores for the item of interest were generated.

For each of the eight conditions and for each of the 41 values of $\rho_{ii'}$, once they were determined, 1000 datasets of size $N = 1000$ each were generated by drawing θs from a standard normal distribution. For each design cell, the mean value of each of the seven item indices was determined across 1000 replicated datasets.

4.3 Results

For the eight different conditions, for each $\rho_{ii'}$ value, Figure 4.1 shows the mean of 1000 replications for the item-score reliability methods MS, λ_6, and CA. The 45-degree solid line indicates unbiased estimation, and deviation from the 45-degree solid line indicates bias of the item-score reliability method with respect to

$\rho_{ii'}$. Next, we discuss the results for each of the seven methods separately.

Method MS. When $\rho_{ii'} = 0$, method MS also equaled 0, but when $0 < \rho_{ii'} < .2$, method MS exceeded $\rho_{ii'}$. As of $\rho_{ii'} \approx .15$ (lower bundle of 5 curves: 4 items having equal as, 1 unequal as; call "equal as" or a_e) and $\rho_{ii'} \approx .25$ (upper bundle: 3 items having unequal as; call "unequal" as or a_u) until $\rho_{ii'} = .8$, method MS values leveled off quickly until $.32$ (equal as) and $.42$ (unequal as). Thus, between roughly $\rho_{ii'} = .2$ and $\rho_{ii'} = .8$, underestimation by method MS increased from 0 to approximately $.4$. This result renders method MS a weak estimator of $\rho_{ii'}$, but most important, the user can have confidence that MS values above the desired threshold for $\rho_{ii'}$ indicate sufficient reliability, even if they are underestimates. If sufficient reliability is arbitrarily chosen to be $\rho_{ii'} = .3$, then for items having equal as, the results showed that in the conditions in this study MS values should be at least $.23$, and for items having unequal as, MS values should be at least $.28$. These MS values are approximations. The result that for somewhat higher $\rho_{ii'}$ values, MS values soon begin to grossly underestimate reliability, and do this at an increasing rate dependent on the discrimination pattern of the items, renders the use of method MS problematic but not unfeasible. Because of the gross $\rho_{ii'}$ underestimation together with people's tendency to focus on high reliability values, if one wants to avoid the risk of deleting items that have sufficiently high reliability insufficiently shown by their MS values, an MS cutoff value equal to, say, $.25$, may be adequate for practical use. Thus, $\rho_{ii'}^{MS} \geq .25$ indicates sufficiently high item-score reliability.

Method λ_6. The graph for method λ_6 shows three bundles of curves, and upon closer inspection, the middle bundle consists of two bundles each containing a pair of curves where one bundle crosses the other. The two middle bundles are characterized by "unequal as" (gradually sloped curves) and "average b_i & short test" (steeply sloped curves), and together they constitute a picture comparable albeit not identical to the graph for method MS. The lower bundle is characterized by "high b_i & short test" (relatively flat curves) and the upper bundle by "average b_i & long test" (steeper curves). The latter two lower and upper bundles are different from the method MS graph, and together the constellation of bundles renders the results difficult to interpret. Unlike the results for method MS, the results for method λ_6 provide underestimates for almost all $\rho_{ii'}$ values including the lowest $\rho_{ii'}$ values, and similarly to method MS, method λ_6 estimates level off as $\rho_{ii'}$ increases but variation of negative bias is much greater across design conditions than with method MS. An additional simulation study, not reported here in detail, showed that the overestimation by method λ_6 for lower values of $\rho_{ii'}$ in the long-test conditions disappears when the data consist of continuous instead of discrete item scores. Given item and test properties, a useful cutoff value for method λ_6 could be $.2$. However, this does not hold for items in a short test that have a b_i parameter that is an outlier with respect to the other b_i values in the test. In these conditions, method λ_6 did not differentiate sufficiently between different levels of $\rho_{ii'}$ and showed low values along the entire range of $\rho_{ii'}$ values. For items where the difficulty is close to the mean of the other items' difficulties and that are in a long test, the results showed that method λ_6 approximates $\rho_{ii'}$ quite well.

Method CA. Considering the bundles of the curves, results for method CA look like results for method λ_6 with the middle bundles left out. The lower bundle consists of the four "high b_i" conditions, where the item's location parameter is located 1.5 standard deviation from the middle of the θ distribution. Because method CA has the squared corrected item-total correlation in the numerator, and because for the "high b_i" conditions the item scores are skew relative to the rest-score distribution, corrected item-total correlations are suppressed (Nunnally, 1978, p. 145), which also suppresses method CA values. Hence, the gross underestimation of $\rho_{ii'}$ for "high b_i". The upper bundle consists of the four "average b_i" conditions, and given the item's location relative to the θ distribution, correlations are not suppressed, and method CA gives good estimates. The results for method CA showed that a cutoff value of $.3$ can be used for "average b_i" conditions, but for "high b_i" conditions, method CA provides results that are hardly useful.

To summarize, all three item-score reliability methods showed increasing negative bias as $\rho_{ii'}$ increased. This indicates that in many cases, methods MS, λ_6, and CA understimate the population item-score reliability. At $\rho_{ii'} = .3$, method MS estimated values between $.25$ and $.30$. The highest MS values for $\rho_{ii'} = .8$ were approximately $.4$. Method λ_6 values at $\rho_{ii'} = .3$ ranged from $.10$ to $.28$. For $\rho_{ii'} = .8$, the highest values were approximately $.58$. Method CA values at $\rho_{ii'} = .3$ ranged from $.2$ to $.3$, but for items with "high b_i", a CA value as high as $.3$ was never estimated. For items with "average b_i", the highest CA values equaled $.65$. These results suggest that for values from $\rho_{ii'} = .3$ onwards all methods underestimate $\rho_{ii'}$ which means that the lower bound of $.3$ suggested by Zijlmans, Tijmstra, et al. (2018b) might be too stringent in practice.

For the eight different conditions and for a range of $\rho_{ii'}$ values, Figure 4.2 shows the mean of 1000 replications for each of the four item-functioning indices. The horizontal line shows the cut-off value for each index, suggested in the literature, and indicating a sufficient level for that item-functioning index in terms of item quality.

Corrected item total-correlation. The relationship between $\rho_{ii'}$ and the corrected item total-correlation was characterized by two different patterns. One pattern consisted of conditions referring to "average b_i" and the other pattern consisted of conditions referring to "high b_i". As $\rho_{ii'}$ increased, the corrected item

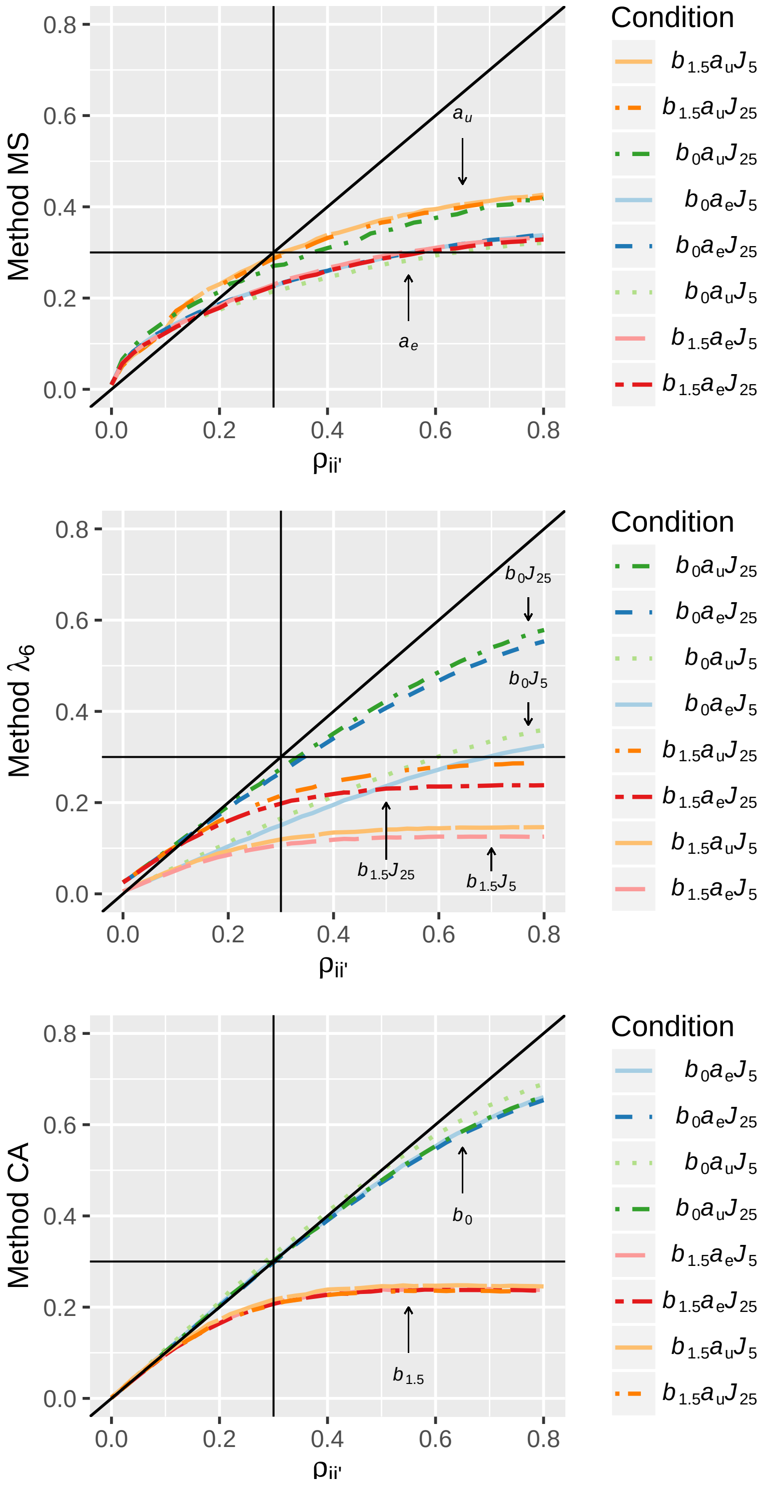

Figure 4.1: Mean bias of the three item-score reliability methods in the eight different conditions. The 45-degree solid line indicates no bias.

Note: $b_0 = b_i = 0$, $b_{1.5} = b_i = 1.5$, a_e = equal discrimination parameters, a_u = unequal discrimination parameters, $J_5 = 5$ other items in the test, $J_{25} = 25$ other items in the test. A combination of these codes corresponds to a design cell.

total-correlation increased, but from $\rho_{ii'} = .2$ onwards, for "high b_i" the estimated corrected item total-correlations stabilized at approximately $.33$. For "average b_i", the corrected item total-correlation further increased for higher values of $\rho_{ii'}$. Compared to shorter tests, for the longer tests with "average b_i", higher values were estimated for the corrected item total-correlation. The same result was found for the conditions referring to "average b_i". Varying a parameters of the other items in the test did not seem to influence the relationship between the corrected item total-correlation and $\rho_{ii'}$. The cut-off value of $.3$ for the corrected item total-correlation was estimated between $.1 < \rho_{ii'} < .25$.

Item-factor loading. Estimating the item-factor loading resulted in non-converged solutions for 577 and 546 replications of the whole range of $\rho_{ii'}$ values in the "high b_i" and "long test" conditions for equal and unequal as respectively. The number of non-converged models increased as $\rho_{ii'}$ increased. This resulted in non available values for at most 8.8 per cent per value of $\rho_{ii'}$. The relationship between $\rho_{ii'}$ and the item-factor loading was characterized by one bundle containing all curves, meaning that the design factors do not differentially influence the relationship. The small and non-linear difference between the "average b_i" and "high b_i" curves is negligible. As $\rho_{ii'}$ increased, the value for the item-factor loading increased. The relationship between $\rho_{ii'}$ and the item-factor loading may be transformed to become linear, but this was not a goal we pursued. The cutoff value of $.3$ was located at $\rho_{ii'} = .05$.

Item scalability. The relationship between $\rho_{ii'}$ and item scalability was characterized by two different bundles, one for conditions with "average b_i" and one for conditions with "high b_i". For larger values of $\rho_{ii'}$, both bundles showed increasing values for item scalability. Compared to corrected item total-correlation, for item scalability the bundles were reversed. An explanation could be that item scalability is a normed item-rest covariance that corrects for the maximum possible covariance between two variables. Because the maximum possible covariance between an item having an "high b_i" and the rest score can become very small as the distribution of the dichotomous variable is skewer (Nunnally, 1978, p. 145), and because the item scalability corrects for this effect, the results are reversed between the corrected item total-correlation and the item scalability. The cutoff value of $.3$ for item scalability was found at $\rho_{ii'} = .2$ for $b_i = 1.5$ and at $\rho_{ii'} = .4$ for $b_i = 0$.

Item discrimination. The relationship between $\rho_{ii'}$ and item discrimination did not show a clear structure or pattern. The eight conditions showed little difference, which was to be expected since for all conditions with the same b_i the same item discrimination was used. This resulted in two bundles characterized by the same pattern, determined by the value for b_i. The curves increased slowly for

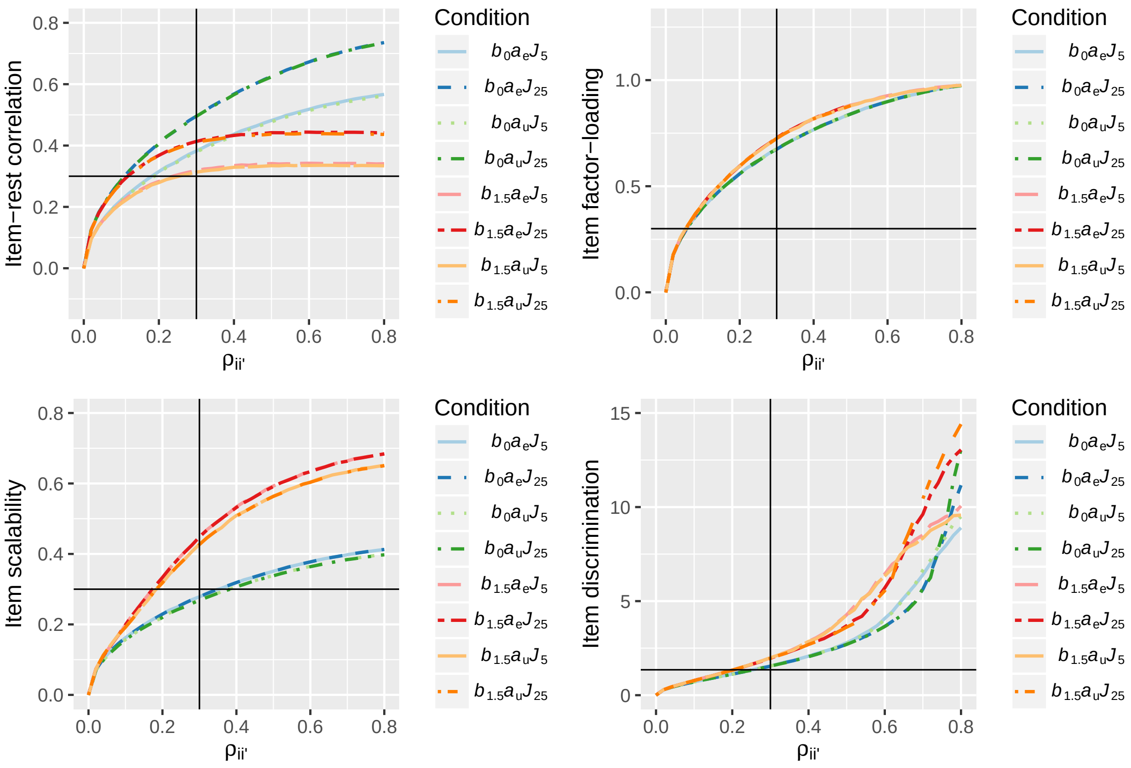

Figure 4.2: Relationship between $\rho_{ii'}$ and the four item indices in the eight different conditions. The cutoff values described in the literature are shown by a horizontal line. *Note*: $b_0 = b_i = 0$, $b_{1.5} = b_i = 1.5$, a_e = equal discrimination parameters, a_u = unequal discrimination parameters, $J_5 = 5$ other items in the test, $J_{25} = 25$ other items in the test. A combination of these codes corresponds to a design cell.

lower values of $\rho_{ii'}$, but showed a steeper increase from $\rho_{ii'} \;\; = .6$ onwards. The cut-off value of 1.35 was at $\rho_{ii'} \;\; = .2$.

We investigated the estimates of the four other item-functioning indices for different values of $\rho_{ii'}$ and compared them to their cutoff values. For long tests, the cut-off score for the corrected item total-correlation was at $\rho_{ii'} \;\; = .2$, and for short tests the cutoff score was at $\rho_{ii'} \;\; = .2$. The cutoff score for item scalability for items with $b_i = 1.5$ was just below $\rho_{ii'} \;\; = .2$, and for items with $b_i = 0$, the cut-off score was at $\rho_{ii'} \;\; = .4$. The cut-off score of the item-factor loading was located at $\rho_{ii'} \;\; = .05$. The cutoff-value for item discrimination was estimated at $\rho_{ii'} \;\; = .2$ for all conditions. The lower bound of $\rho_{ii'} \;\; = .3$ Zijlmans, Tijmstra, et al. (2018b) proposed seems to be too high, given that the cutoff scores for other item-functioning indices are located at a lower level of $\rho_{ii'}$.

4.4 Discussion

This study discusses the relationship between item-score reliability $\rho_{ii'}$ and three methods to estimate $\rho_{ii'}$, which are methods MS, λ_6, and CA. Also, the relationship between $\rho_{ii'}$ and four item-functioning indices was investigated. Finally, the proposed lower bound for item-score reliability of $.3$ was investigated to evaluate whether this lower bound matches cutoff values for other item-functioning indices.

All methods showed increasing bias for increasing values of $\rho_{ii'}$. Method MS showed small bias until $\rho_{ii'} = .2$, but for higher values for $\rho_{ii'}$ the bias increased rapidly. Method λ_6 showed small bias in conditions with a longer test and person-centered location parameter ($b_i = 0$). Method CA shows almost unbiased values for conditions with a person-centered location parameter, but larger bias for conditions with a high location parameter ($b_i = 1.5$). The increasing negative bias all methods showed for increasing values of $\rho_{ii'}$ may be caused by the increasing contrast between the item of interest, for which the item-score reliability was computed, and the other items in the test. Because methods MS, λ_6, and CA use the other items in the test to approximate $\rho_{ii'}$, the more the item differs from the other items in the test, the harder it seems for the methods to give a good approximation of $\rho_{ii'}$. While the characteristics for the $J-1$ items are kept constant, the item of interest starts to differ more from those items for increasing values of $\rho_{ii'}$. Because the discrimination parameter was used to manipulate the level of $\rho_{ii'}$, for higher levels of $\rho_{ii'}$ the item of interest has a higher discrimination parameter. For $\rho_{ii'} = .8$, the discrimination parameter was 7.787 for a person-centered location parameter and 10.281 for a high location parameter. These values differ so much from the discrimination parameters of the other items in the test that ranged from -1 to 1, that this discrepancy might have produced great underestimation of $\rho_{ii'}$ for the item of interest. The reason for this underestimation is that all three methods use the other items in the test to approximate an independent replication of the item score. If the other items in the test deviate much from the item of interest, this approximation seems to get worse.

Methods MS, λ_6, and CA underestimated $\rho_{ii'}$ in some conditions at the $\rho_{ii'} = .3$ level. Hence, the conclusion Zijlmans, Tijmstra, et al. (2018b) drew to use a lower bound of $.3$ for sufficient item-score reliability might be too strict in practice. Also, because methods MS, λ_6, and CA grossly underestimated $\rho_{ii'}$ for higher values of $\rho_{ii'}$, the estimated values are inconsistent with true population item-score reliability. The analysis of several empirical-data sets in Zijlmans, Tijmstra, et al. (2018b) showed maximum values of $.70$, $.81$, and $.90$ for respectively methods MS, λ_6, and CA. However, these values are located at the end of the distribution, with values of $.38$, $.45$, and $.34$ for the third quartile for methods MS, λ_6, and CA, respectively. This suggests that in practice high values are rare.

Except for the item scalability in the $b_i = 0$ condition, the cutoff values for the four other item-functioning indices were between $.05 < \rho_{ii'} \leq .2$. We conclude that sufficient item-functioning is within this range. Especially for the item-factor loading, the cutoff value of $.3$ was at a very low value of $\rho_{ii'}$, that is $\rho_{ii'} = .05$. This indicates that the item-factor loading selects items at a level where $\rho_{ii'}$ is low. The item-score reliability showed a near one-to-one relationship with the item-factor loading. These results suggest that the item-factor loading might be a good predictor for item-score reliability, but the relationship might also be a consequence of the data-generating model, which is unidimensional, resulting in good model fit for the one-factor model.

We conclude that the proposed lower bound of $.3$ for the item-score reliability methods may be too stringent in practice, because methods MS, λ_6, and CA underestimate $\rho_{ii'}$, especially for higher values of $\rho_{ii'}$. Also, there is a relationship between the four other item-functioning indices and $\rho_{ii'}$, but the corrected item total-correlation, item scalability and item discrimination seem to measure different aspects of item functioning. The item-factor loading could be an interesting measure to approximate item-score reliability and deserves further investigation to determine its usability in the context of item-score reliability.

Chapter 5

Item-Score Reliability as a Selection Tool in Test Construction

Abstract

This study investigates the usefulness of item-score reliability as a criterion for item selection in test construction. Methods MS, λ_6, and CA were investigated as item-assessment methods in item selection and compared to the corrected item-total correlation, which was used as a benchmark. An ideal ordering to add items to the test (bottom-up procedure) or omit items from the test (top-down procedure) was defined based on the population test-score reliability. The orderings the four item-assessment methods produced in samples were compared to the ideal ordering, and the degree of resemblance was expressed by means of Kendall's τ. To investigate the concordance of the orderings across 1000 replicated samples, Kendall's W was computed for each item-assessment method. The results showed that for both the bottom-up and the top-down procedure, item-assessment method CA and the corrected item-total correlation most closely resembled the ideal ordering. Generally, all item assessment methods resembled the ideal ordering better, and concordance of the orderings was greater, for larger sample sizes and greater variance of the item discrimination parameters.

Keywords: corrected item-total correlation, correction for attenuation, item-score reliability, item selection in test construction, method CA, method λ_6, method MS

Based on Zijlmans, E. A. O., Tijmstra, J., Van der Ark, L. A., and Sijtsma, K. (2019). Item-Score Reliability as a Selection Tool in Test Construction. *Frontiers in Psychology, 9*, 2298. doi: 10.3389/fpsyg.2018.02298

5.1 Introduction

When adapting an existing test, the test constructor may wish to increase or decrease the number of items for various reasons. On the one hand, the existing test may be too short, resulting in test-score reliability that is too low. In this case, adding items to the test may increase test-score reliability. On the other hand, the existing test may be too long to complete in due time. A solution could be to decrease the number of items, but after removal of a number of items, the test score based on the remaining items must be sufficiently reliable. Test constructors could use the reliability of individual items to make decisions about the items to add to the test or to remove from the test. This article investigates the usefulness of item-score reliability methods for making informed decisions about items to add or remove when adapting a test.

Several approaches to item selection in test construction have been investigated. Raubenheimer (2004) investigated an item selection procedure that maximizes coefficient alpha of each subscale within a multi-dimensional test, and simultaneously maximizes both the convergent and discriminant validity using exploratory factor analysis. Raykov (2007, 2008) discussed the use of procedure “alpha if item deleted” to omit items from a test and concluded that maximizing coefficient alpha results in loss of criterion validity. Erhart et al. (2010) studied item reduction by either maximizing coefficient alpha or the item fit of the partial credit model (Masters, 1982). They concluded that both item reduction approaches should be accompanied by additional analyses. Because the quality of a test depends on more than only the reliability of its test score, taking additional information in consideration obviously is a wise strategy. However, in this study we preferred to focus on optimizing test-score reliability in the process of adding items to the test or omitting items from the test. This enabled us to assess the value of particular item selection procedures and item-assessment methods, in particular, item-score reliability methods. The usability of item-score reliability in item selection procedures was investigated in detail.

Zijlmans, Van der Ark, et al. (2018) investigated four methods to estimate item-score reliability. The three most promising methods from this study were methods MS, λ_6, and CA. Zijlmans, Tijmstra, et al. (2018b) applied these methods to empirical-data sets and investigated which values of item-score reliability can be expected in practice and how these values relate to four other item indices that did not assess the item-score reliability in particular, item discrimination, item loadings, item scalability, and corrected item-total correlations. In a third study (Zijlmans, Tijmstra, Van der Ark, & Sijtsma, 2018a), the relationship between the item-score reliability methods and the four other item indices was further investigated by means of a simulation study. The use of the three item-score reliability methods for maximizing test-score reliability has not been investigated yet. Therefore, in this study the usefulness of item-score reliability methods MS, λ_6, and CA for constructing reliable tests was investigated.

The three research questions we addressed are the following. First, are item-score reliability methods useful for adding items to a test or omitting from a test, when the goal is to maximize test-score reliability of the resulting test? Second, to what extent do the orderings in which the three item-score reliability methods select items resemble the theoretically optimal ordering in which items are selected or removed when maximizing population test-score reliability? Third, do the orderings produced by each of the three item-score reliability methods bear more resemblance to the theoretically optimal ordering than the ordering the corrected item-total correlation produced? These questions were addressed by means of a simulation study.

This article is organized as follows. First, we discuss bottom-up and top-down procedures for constructing a test. Second, we discuss the item-assessment methods we used, which are item-score reliability methods MS, λ_6, and CA, and the corrected item-total correlation. Third, we discuss the design for the simulation study and the data-generating process. Finally, the results and their implications for test construction are discussed.

5.2 Item Selection in Test Construction

In this study, we focus on two procedures for test construction. For both procedures, the test constructor has to make an informed decision about the balance between the desired length of the test and the desired minimum test-score reliability. Hence, we focus entirely on selection or omission of items based on formal assessment methods. The first procedure selects items from the pool of available items, and adds the selected items one by one to the preliminary test. We refer to this procedure as the bottom-up procedure. The second procedure uses the complete pool of available items as the initial test, and selects items one by one for elimination from this test. We refer to this procedure as the top-down procedure.

Bottom-Up Procedure

The bottom-up procedure starts by defining an initial test consisting of two items from the pool of available items. In general, and apart from the present study, different criteria to select the initial two-item test can be used. For example, the test constructor may consider the two items he or she starts with the substantive kernel of the test, or he or she may choose the two items that have proven to be of excellent quality in the past. In both examples, the researcher includes the item in the test. In our study, the item pair having the highest test-score reliability was selected. The selected item pair constituted the initial test and both items were removed from the pool of available items. In the next step, the third item was added to the two-item test that maximizes the test-score reliability $\rho_{XX'}$ for a three-item test, based on all available choices; then, the fourth item was added to the three-item test following the same logic, and so on. In practice, it is impossible to estimate $\rho_{XX'}$, because parallel test scores X and X' are usually unavailable (Lord & Novick, 1968, p. 106). In this study, four item-assessment methods were investigated that can be used to add the items to the test. The test constructor may use one of these four item-assessment methods to continue the bottom-up procedure until the test has the desired length or a sufficiently high test-score reliability, or both. Because in practice, different test constructors may entertain different requirements for test length and minimum reliability, adding items to the preliminary test may stop at different stages of the procedure. Hence, for the sake of completeness, we described the complete ordering based on adding each of the available items to the test until all items were selected.

Computational Example Bottom-Up Procedure

We discuss a computational example. We started with 20 equally difficult items for which the item discrimination parameters were ordered from smallest to largest. To select the initial 2-item test, we considered the theoretical $\rho_{XX'}$ values for all possible 2-item tests. Test-score reliability was defined theoretically based on available item parameters of the 2-parameter logistic model (2PLM), assuming a standard normal distribution of the latent variable (see Appendix B). Items 19 and 20 had the highest test-score reliability, so this pair constituted the initial test; see Table 5.1. In Step 1, a pool of 18 items was available from which to add an item to the preliminary test version. Consider the columns in Table 5.1 headed by Step 1. The $\rho_{XX'}$ column shows the $\rho_{XX'}$ values for each 3-item test including one of the available items, so that one can evaluate the test-score reliability of each 3-item test. Item 18 resulted in the highest test-score reliability and was added in Step 2. The procedure continued until the pool of available items was empty. In the penultimate step, which is Step 18, item 2 was added to the preliminary test. Item 1 was added in the last step. From this procedure, we derived the ordering in which the items were added to the preliminary test versions, when the goal was to maximize the test-score reliability in each step.

Table 5.1: Example Item-Selection Procedure Following the Bottom-Up Method Based on the Test-Score Reliability $\rho_{XX'}$.

Step 1			Step 2			···	Step 17			Step 18		
Items in Test	Items in Pool	$\rho_{XX'}$ if item added	Items in Test	Items in Pool	$\rho_{XX'}$ if item added	···	Items in Test	Items in Pool	$\rho_{XX'}$ if item added	Items in Test	Items in Pool	$\rho_{XX'}$ if item added
20	1	0.462	20	1	0.556	···	20	1	0.807	20	1	0.809
19	2	0.467	19	2	0.560	···	19	2	0.808	19	**2**	**0.810**
	3	0.473	18	3	0.564	···	18	**3**	**0.808**	18		
	4	0.479		4	0.568	···	17			17		
	5	0.484		5	0.573	···	16			16		
	6	0.491		6	0.577	···	15			15		
	7	0.497		7	0.582	···	14			14		
	8	0.503		8	0.586	···	13			13		
	9	0.510		9	0.591	···	12			12		
	10	0.516		10	0.595	···	11			11		
	11	0.523		11	0.600	···	10			10		
	12	0.530		12	0.605	···	9			9		
	13	0.536		13	0.610	···	8			8		
	14	0.543		14	0.615	···	7			7		
	15	0.550		15	0.620	···	6			6		
	16	0.557		16	0.625	···	5			5		
	17	0.564		**17**	**0.630**	···	4			4		
	18	**0.571**				···				3		

Note. The example is based on the condition with small variance of discrimination parameters. The $\rho_{XX'}$ column indicates the possible test-score reliability, if in the next step, that item would be added to the test. For each step the selected item is indicated in boldface.

Top-Down Procedure

For the top-down procedure, the complete pool of available items constitutes the initial test, and the items are deleted one by one until two items remain. Ideally, the test-score reliability of all twenty 19-item tests is computed, determining which item should be omitted, so that the test consisting of the remaining 19 items had the highest test-score reliability of all 19-item tests. The first item that is omitted either increases the test-score reliability the most or decreases the test-score reliability the least. This procedure is repeated until the test consisted of only two items. Two items constitute the minimum, because an item-assessment method cannot be applied to a single item. This procedure results in an ordering in which items were omitted from the test. A test constructor usually does not continue until the test consists of only two items but rather stops when the resulting test has the desired length or the desired test-score reliability, or both. In our study, for the sake of completeness, we continued the item selection until the test consisted of only two items so that the results for the complete top-down procedure were visible.

Table 5.2: Example Item-Selection Procedure Following the Top-Down Method Based on the Test-Score Reliability $\rho_{XX'}$.

	Step 1			Step 2		···		Step 17			Step 18	
Items Omitted	Items in Test	$\rho_{XX'}$ if item omitted	Items Omitted	Items in Test	$\rho_{XX'}$ item omitted	···	Items Omitted	Items in Test	$\rho_{XX'}$ item omitted	Items Omitted	Items in Test	$\rho_{XX'}$ item omitted
	1	**0.810**	1	**2**	**0.808**	···	1	**17**	0.571	1	**18**	**0.481**
	2	0.809		3	0.807	···	2	18	0.564	2	19	0.47
	3	0.809		4	0.807	···	3	19	0.557	3	20	0.46
	4	0.808		5	0.806	···	4	20	0.550	4		
	5	0.808		6	0.806	···	5			5		
	6	0.807		7	0.805	···	6			6		
	7	0.806		8	0.804	···	7			7		
	8	0.806		9	0.804	···	8			8		
	9	0.805		10	0.803	···	9			9		
	10	0.804		11	0.802	···	10			10		
	11	0.804		12	0.801	···	11			11		
	12	0.803		13	0.800	···	12			12		
	13	0.802		14	0.799	···	13			13		
	14	0.801		15	0.799	···	14			14		
	15	0.800		16	0.798	···	15			15		
	16	0.800		17	0.797	···	16			16		
	17	0.799		18	0.796	···				17		
	18	0.798		19	0.795	···						
	19	0.797		20	0.794	···						
	20	0.796				···						

Note. The example is based on the condition with small variance of discrimination parameters. The $\rho_{XX'}$ column indicates the possible test-score reliability, if that item would be omitted from the test in the next step. For each step the selected item is indicated in boldface.

Computational Example Top-Down Procedure

We employed the same 20 items we used in the computational example for the bottom-up procedure, and included all items in the initial test. In Step 1, $\rho_{XX'}$ was computed when a particular item was omitted from the test (using the procedure outlined in Appendix B; see Table 5.2, column $\rho_{XX'}$ if item omitted for the $\rho_{XX'}$ values). In the example, omitting item 1 from the test produced the highest $\rho_{XX'}$ value. Thus, item 1 was omitted and we continued with the 19-item test. In Step 2, omitting item 2 resulted in the highest test-score reliability for the remaining 18 items. This procedure was repeated until the test consisted of only items 19 and 20.

Item-Assessment Methods

We used the three item-score reliability methods MS, λ_6, and CA, and the corrected item-total correlation to add items to the preliminary test or to omit items from the preliminary test. The corrected item-total correlation was included to compare the item-score reliability methods to a method that has been used for a long time in test construction research. Both the bottom-up and the top-down procedures were applied using the four item-assessment methods instead of test-score reliability $\rho_{XX'}$. The eight orderings that resulted from combining the two item selection procedures with the four item-assessment methods were compared to the ordering based on the theoretical test-score reliability, to infer which item-assessment method resembled the ordering that maximizes the test-score reliability best.

Item-score reliability methods

The following definitions were used (Lord & Novick, 1968, p. 61). Test score X is defined as the sum of the J item scores. Let X_i be the item score, indexed i $(i = 1, \ldots, J)$; $X = \sum_{i=1}^{J} X_i$. Item-score reliability is defined as the ratio of the true-score variance, denoted $\sigma^2_{T_i}$, and the observed-score variance, denoted $\sigma^2_{X_i}$. The observed-score variance can be split in true-score variance and error variance denoted $\sigma^2_{E_i}$, which means item-score reliability can also be defined as 1 minus the proportion of observed-score variance that is error variance; that is,

$$\rho_{ii'} = \frac{\sigma^2_{T_i}}{\sigma^2_{X_i}} = 1 - \frac{\sigma^2_{E_i}}{\sigma^2_{X_i}}. \tag{5.1}$$

Three methods to approximate item-score reliability were used to decide which item will be added to the test or omitted from the test: method MS, method λ_6, and method CA. These methods are briefly discussed here; see Zijlmans, Van der Ark, et al. (2018) for details.

Method MS

Method MS is based on the Molenaar-Sijtsma test-score reliability method (Molenaar & Sijtsma, 1988; Sijtsma & Molenaar, 1987). This method uses the double monotonicity model for dichotomous items proposed by Mokken (1971) which assumes a unidimensional latent variable θ, locally independent item scores, and monotone, nondecreasing, and nonintersecting item-response functions. The items are ordered from most difficult to easiest and this ordering is used to obtain an approximation of an independent replication of the item of interest, denoted i. Mokken (1971, pp. 142-147) proposed to approximate independent replications of the item scores by using information from the item of interest, the next-easier item $i + 1$, the next more-difficult item $i - 1$, or both neighbor items. The idea is that items that are close to the item of interest in terms of location provide a good approximation of an independent replication of the target item. We denote method MS for estimating item-score reliability $\rho^{MS}_{ii'}$ and estimate the independent replication approximated in $\rho^{MS}_{ii'}$ using the procedure as explained by Van der Ark (2010).

Method λ_6

Test-score reliability method λ_6 (Guttman, 1945) was adjusted by Zijlmans, Van der Ark, et al. (2018), such that it approximates the reliability of an item score. Let ϵ_i^2 be the residual error variance from the multiple regression of the score on item i on the remaining $J-1$ item scores. The ratio of ϵ_i^2 and the observed item variance $\sigma^2_{X_i}$ is subtracted from unity to obtain the item-score reliability estimate by means of method λ_6, denoted $\rho^{\lambda_6}_{ii'}$; that is,

$$\rho^{\lambda_6}_{ii'} = 1 - \frac{\epsilon_i^2}{\sigma^2_{X_i}}. \tag{5.2}$$

Method CA

Method CA is based on the correction for attenuation (Lord & Novick, 1968, pp. 69-70; Nunnally & Bernstein, 1994, p. 257; Spearman, 1904) and correlates the item score with a test score, which is assumed to measure the same attribute as the item score (Wanous & Reichers, 1996). This test score can be obtained from the remaining $J-1$ items in the test or the items in a different test, which is assumed to measure the same attribute as the target item. We denote item-score reliability approximated by method CA as $\rho^{CA}_{ii'}$. The corrected item-total correlation is defined as $\rho_{X_iR_{(i)}}$, and correlates the item score with the rest score, defined as $R_{(i)} = X - X_i$. Coefficient $\alpha_{R_{(i)}}$ is a lower bound to the reliability of the rest score, estimated by reliability lower bound coefficient α (e.g., Cronbach, 1951). Method CA is defined as

$$\rho^{CA}_{ii'} = \frac{\rho^2_{X_iR_{(i)}}}{\alpha_{R_{(i)}}}. \tag{5.3}$$

Earlier research (Zijlmans, Van der Ark, et al., 2018) showed that methods MS and CA had little bias. Method λ_6 produced precise results, but underestimated $\rho_{ii'}$, suggesting it is a conservative method. These results showed that the three methods are promising for estimating $\rho_{ii'}$.

Corrected Item-Total Correlation

The corrected item-total correlation $\rho_{X_iR_{(i)}}$ was defined earlier. Higher corrected item-total correlations in a test result in a higher value of coefficient α (Lord & Novick, 1968, p. 331). In test construction, the corrected item-total correlation is used to define the association of the item with the total score on the other items. The corrected item-total correlation is also used by method CA (see Equation 5.3).

5.3 Simulation Study

By means of a simulation study it was investigated whether the four item-assessment methods added items to a test (bottom-up procedure) or omitted items from a test (top-down procedure) in the same ordering that would result from adding items or omitting items, such that the theoretical $\rho_{XX'}$ was maximized in each item selection step.

Method

For the bottom-up procedure, for each item-assessment method, we investigated the ordering in which items were added. In each selection step, the item was selected that had the greatest estimated item-score reliability or the greatest corrected item-total correlation based on its inclusion in the preliminary test. For the top-down procedure, for each item-assessment method, we investigated the ordering in which items were omitted, this time in each step omitting the item that had the smallest item-score reliability or the smallest corrected item-total correlation. The ordering in which items were added or omitted was compared to the ideal ordering if theoretical test-score reliability was used. The degree to which the orderings produced by an item-assessment method resembled the ideal ordering, was expressed by Kendall's τ. The concordance of orderings produced by each item-assessment method over samples was expressed by means of Kendall's W. Next, we discuss the details of the simulation study.

Dichotomous scores for 20 items were generated using the two-parameter logistic model (Birnbaum, 1968). Let θ be the latent variable representing a person's attribute, α_i the discrimination parameter of item i, and β_i the location parameter of item i. The two-parameter logistic model is defined as

$$P(X_i = 1 \mid \theta) = \frac{\exp\left[\alpha_i(\theta - \beta_i)\right]}{1 + \exp\left[\alpha_i(\theta - \beta_i)\right]}. \tag{5.4}$$

The variance of the discrimination parameters was varied. The discrimination parameter of an item conceptually resembles its item-score reliability (Tucker, 1946). All sets of discrimination parameters had the same median value. We used sets of values that had the same mean on the log scale, which guaranteed that all discrimination parameters were positive and that for each condition the median discrimination parameter was 1, and considered values equidistantly spaced ranging from -0.5 to 0.5, -1 to 1, or -2 to 2 on the log scale. The variance of the discrimination parameters is referred to as either *small*, ranging from 0.61 to 1.65 on the original scale, *average*, ranging from 0.37 to 2.72, or *large*, ranging from 0.14 to 7.39. For all items, the location parameter β_i had a value of 0. We did not vary β_i,

Table 5.3: Item Parameters used to Generate the Item Scores

Item No.	Small Variance of α	Average Variance of α	Large Variance of α	β
Item 1	0.61	0.37	0.14	0
Item 2	0.64	0.41	0.17	0
Item 3	0.67	0.45	0.21	0
Item 4	0.71	0.50	0.25	0
Item 5	0.75	0.56	0.31	0
Item 6	0.79	0.62	0.39	0
Item 7	0.83	0.69	0.48	0
Item 8	0.88	0.77	0.59	0
Item 9	0.92	0.85	0.73	0
Item 10	0.97	0.95	0.90	0
Item 11	1.03	1.05	1.11	0
Item 12	1.08	1.17	1.37	0
Item 13	1.14	1.30	1.69	0
Item 14	1.20	1.45	2.09	0
Item 15	1.27	1.61	2.58	0
Item 16	1.34	1.78	3.18	0
Item 17	1.41	1.98	3.93	0
Item 18	1.48	2.20	4.85	0
Item 19	1.56	2.45	5.99	0
Item 20	1.65	2.72	7.39	0

Note: $\alpha =$ discrimination parameter, $\beta =$ location parameter

The sets of discrimination parameters had the same mean, and contain values equidistantly spaced, ranging from -0.5 to 0.5, -1 to 1, and -2 to 2 on the log scale, respectively.

because this would complicate the simulation design, rendering the effect of item discrimination, approximating item-score reliability, on the item selection process harder to interpret. Table 5.3 shows the item parameters that were used to generate the item scores. Next to the bottom-up and top-down procedures, we varied the sample size N. We generated item scores for either a small sample ($N = 200$) or a large sample ($N = 1000$). These choices resulted in 2 (sample sizes) $\times$ 3 (variances of discrimination parameters) $=$ 6 design cells. In each design cell, 1000 data sets were generated. The 1000 data sets in each cell were analyzed by the two item selection procedures, each using the four item-assessment methods.

From the parameters of the data generating model, the ideal ordering for both the bottom-up procedure and the top-down procedure was determined using test-score reliability $\rho_{XX'}$ (see Appendix B for the procedure). The goal was to maximize $\rho_{XX'}$ in each step of the two procedures. For the bottom-up and the top-down procedures, we determined the ideal order to add items to the test or omit items from the test, based on maximizing the theoretical $\rho_{XX'}$ in every step. The two item selection procedures and the three sets of discrimination parameters resulted in six ideal orderings. Because discrimination parameters increased going from item 1 to item 20, and the item ordering did not differ over the sets of discrimination parameters, only two ideal item orderings were different. Consequently, for the item selection we have two ideal orderings, one for the bottom-up procedure (consecutively adding items $18, 17, \ldots, 1$) and one for the top-down procedure (consecutively omitting items $1, 2, \ldots, 18$).

The agreement between the ordering determined by each of the item-assessment methods and the ideal ordering determined by $\rho_{XX'}$ was expressed in each data set by means of Kendall's τ. Kendall's τ ranges from -1 to 1, a large negative value indicating that the orderings are dissimilar and a large positive value indicating that the orderings are similar. The item ranks produced by the item-assessment methods can be displayed as a vector, and so can the ideal rank defined at the population level. When the ranks for both elements in a pair agreed this was defined as a concordant pair (C), otherwise the pair was discordant (D). The total number of pairs equals $n(n-1)/2$, where n is the length of the vectors. Kendall's τ is defined as

$$\tau = \frac{\mathrm{C} - \mathrm{D}}{n(n-1)/2}. \tag{5.5}$$

In our study, $n = 18$, based on 18 item selection steps for both item selection methods. We computed the mean for the 1000 Kendall's τ values obtained in each simulation condition, for every combination of item selection procedure and item-assessment method. The mean quantified the resemblance between the ordering each of the item-assessment methods produced and the ideal ordering.

To investigate how much the orderings the item-assessment methods produced differ over 1000 data sets, we computed Kendall's coefficient of concordance, W. Kendall's W expresses the level of agreement between multiple orderings, and W ranges from 0 to 1, a higher value indicating that the orderings an item-assessment method produced are more consistent, resulting in smaller variation. Suppose that item i is given rank r_{ij} in data set j, where there are in total n ranks and m data sets. Then the total rank of object i is $R_i = \sum\limits_{j=1}^{m} r_{ij}$ and the mean value of these ranks is $\bar{R} = \frac{1}{n}\sum\limits_{i=1}^{n} R_i$. The sum of squared deviations, S, is defined as $S = \sum\limits_{i=1}^{n}(R_i - \bar{R})^2$. Kendall's W is defined as

$$W = \frac{12S}{m^2(n^3 - n)}. \tag{5.6}$$

The number of objects n in our study was the number of item selection steps, which was 18. The number of data sets J equaled 1000. For every simulation condition and every item selection method, Kendall's W expressed the agreement among orderings produced by each of the item-assessment methods.

Results

For the bottom-up procedure (upper part of Table 5.4) and the top-down procedure (lower part of Table 5.4), for the six design conditions of sample size and variance of discrimination parameters, Table 5.4 shows the mean Kendall's τ between the ideal ordering and the ordering produced by each of the four item-assessment methods. In the condition with a small sample size and small variance of the discrimination parameters, mean Kendall's τ ranged from $.44$ for method MS to $.59$ for method CA and the corrected item-total correlation. For both procedures, a larger sample size resulted in a higher mean τ value. Mean Kendall's τ increased as variance of discrimination parameters increased for both item selection procedures. For a large sample size and large variance of discrimination parameters, mean τ values ranged from $.80$ for method MS to $.96$ for method CA and the corrected item-total correlation. For both procedures, method CA and the corrected item-total correlation showed the highest mean τ values, meaning that these item-assessment methods resembled the ordering based on $\rho_{XX'}$ best. These two item assessment-methods showed numerically equal mean τ values, which resulted from the nearly identical orderings method CA and the corrected item-total correlation produced. For the bottom-up procedure, four out of six conditions showed exactly the same ordering for each replication using either method CA or the corrected item-total correlation. For the top-down procedure, this result was found in two out of six conditions. Overall, method MS performed worst of all item-assessment methods, where the difference in τ values with the other item-assessment methods was smaller for a larger sample size and increasing variance of the discrimination parameters.

For the bottom-up item selection method (upper part) and the top-down item selection method (lower part), for each of the four item-assessment methods in the six different conditions, Table 5.5 shows Kendall's W values. For both item selection methods, larger sample size showed an increase of W, indicating that the ordering was more alike across replications as sample size increased. This result was also found for increasing variance of discrimination parameters.

For both procedures, method CA and the corrected item-total correlation

Table 5.4: Mean Kendall's τ for 1000 Replications Between the Ordering Based on the Population Test-Score Reliability and the Ordering Produced by the Three Item-Score Reliablility Methods and the Corrected Item-Total Correlation (CITC), for the Bottom-Up and the Top-Down Procedure in the Six Different Conditions.

	Bottom-Up Procedure					
	$N = 200$			$N = 1000$		
	Small Variance of α	Average Variance of α	Large Variance of α	Small Variance of α	Average Variance of α	Large Variance of α
Method MS	0.44	0.67	0.80	0.69	0.83	0.89
Method λ_6	0.55	0.75	0.83	0.80	0.91	0.94
Method CA	0.58	0.78	0.87	0.81	0.92	0.96
CITC	0.58	0.78	0.87	0.81	0.92	0.96
	Top-Down Procedure					
	$N = 200$			$N = 1000$		
	Small Variance of α	Average Variance of α	Large Variance of α	Small Variance of α	Average Variance of α	Large Variance of α
Method MS	0.46	0.64	0.75	0.61	0.73	0.80
Method λ_6	0.55	0.75	0.83	0.81	0.91	0.94
Method CA	0.59	0.78	0.87	0.81	0.92	0.96
CITC	0.59	0.78	0.87	0.81	0.92	0.96

showed the highest W values, suggesting the smallest variance of the orderings over data sets for these item-assessment methods. For the average and large variance of discrimination parameters, W values were all greater than $.96$, meaning that methods λ_6, CA, and the corrected item-total correlation showed almost no variation over replications. For a large sample size, methods λ_6, CA, and the corrected item-total correlation showed similar Kendall's W values.

For each combination of item selection procedure and item-assessment method, the orderings produced by the item-assessment method were used to compute the $\rho_{XX'}$ values in every step of this ordering. This meant that for the items selected at a particular step, we used the item parameters and the distribution of θ to compute $\rho_{XX'}$, and we repeated the computation at each selection step in each of the 1000 samples. For the top-down item selection procedure and item-assessment method CA, Figure 5.1 shows for each step the range of $\rho_{XX'}$ values between the 2.5 and 97.5 percentiles of 1000 values. This combination of item selection procedure and item-assessment method produced the intervals that were narrowest. Intervals became wider as the test grew shorter. For the top-down procedure and item-assessment method MS, Figure 5.2 shows the widest intervals. In both Figure 5.1 and Figure 5.2, intervals grew wider as the test grew shorter.

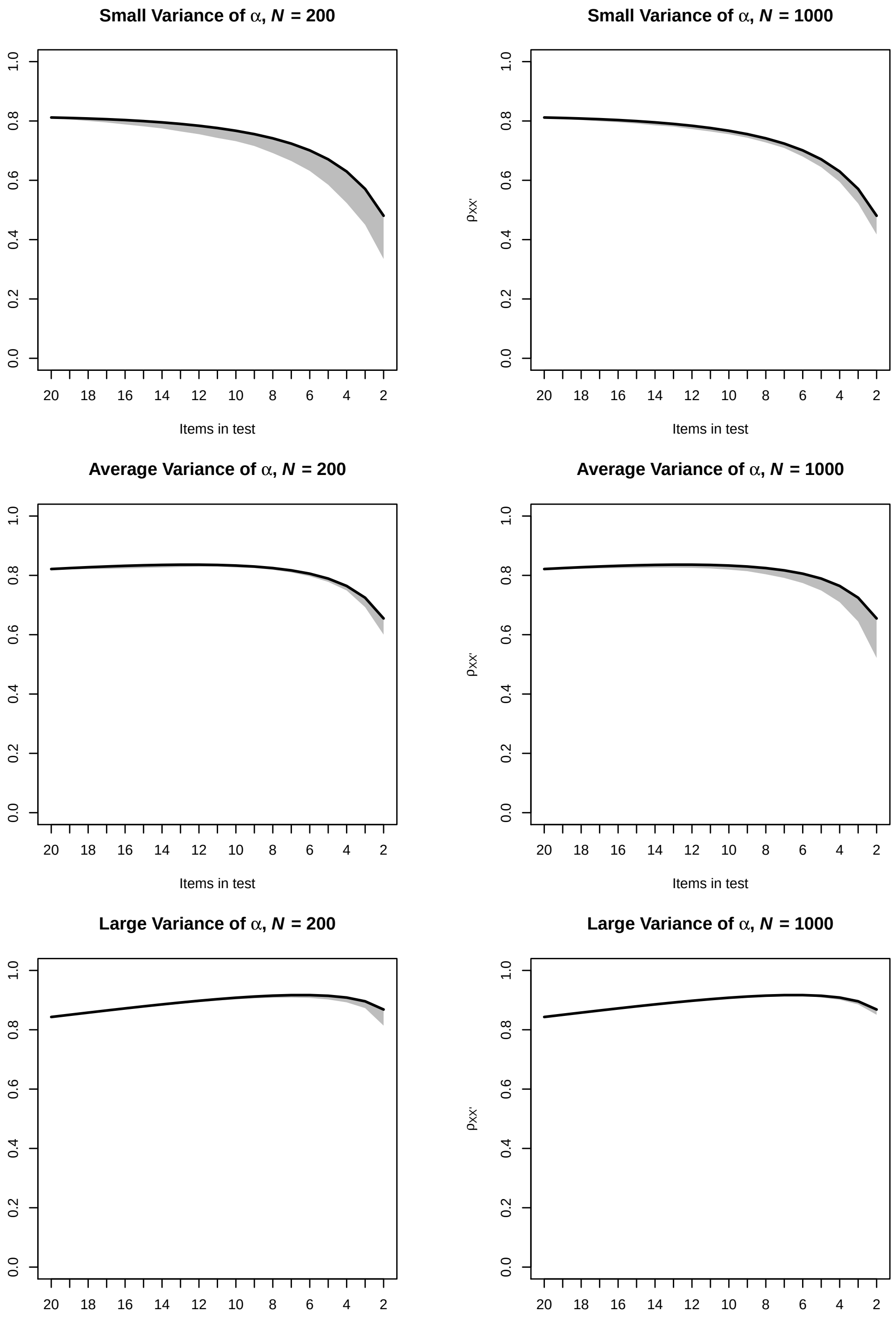

Figure 5.1: Range of $\rho_{XX'}$ values between the 2.5 and 97.5 percentiles of 1000 values produced by method CA in the six conditions for the top-down procedure. The black line indicates the $\rho_{XX'}$ value for the ideal ordering.

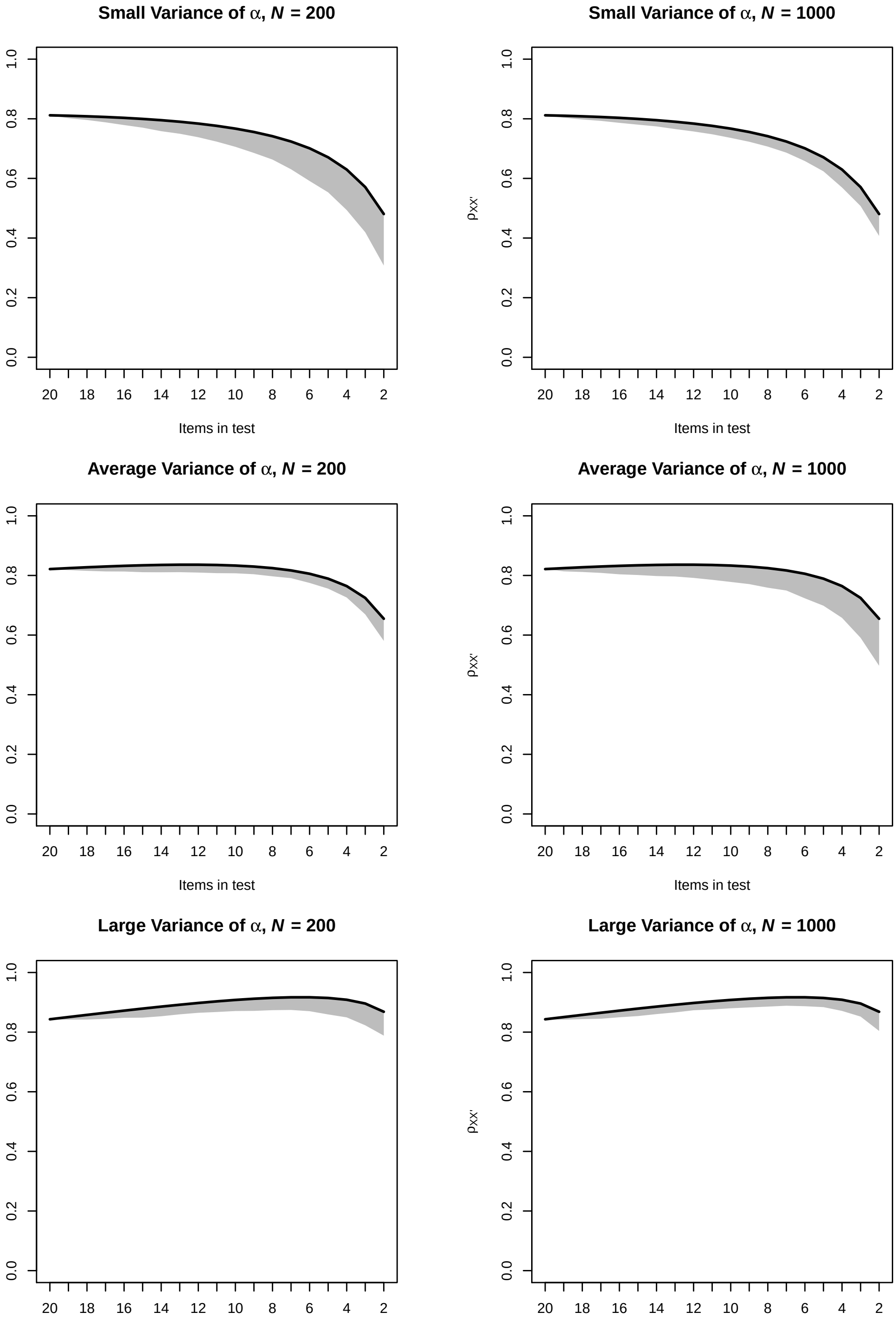

Figure 5.2: Range of $\rho_{XX'}$ values between the 2.5 and 97.5 percentiles of 1000 values produced by method MS in the six conditions for the top-down procedure. The black line indicates the $\rho_{XX'}$ value for the ideal ordering.

Table 5.5: Kendall's W for 1000 Replications Between the Ordering Based on the Population Test-Score Reliability and the Ordering Produced by the Three Item-Score Reliablility Methods and the Corrected Item-Total Correlation (CITC), for the Bottom-Up and the Top-Down Procedure in the Six Different Conditions.

	Bottom-Up Procedure					
	$N = 200$			$N = 1000$		
	Small Variance of α	Average Variance of α	Large Variance of α	Small Variance of α	Average Variance of α	Large Variance of α
Method MS	0.53	0.79	0.90	0.80	0.92	0.96
Method λ_6	0.65	0.87	0.92	0.91	0.97	0.98
Method CA	0.69	0.89	0.95	0.91	0.97	0.99
CITC	0.69	0.89	0.95	0.91	0.97	0.99
	Top-Down Procedure					
	$N = 200$			$N = 1000$		
	Small Variance of α	Average Variance of α	Large Variance of α	Small Variance of α	Average Variance of α	Large Variance of α
Method MS	0.37	0.65	0.80	0.61	0.77	0.85
Method λ_6	0.53	0.82	0.89	0.88	0.96	0.98
Method CA	0.59	0.85	0.93	0.88	0.96	0.98
CITC	0.59	0.85	0.93	0.88	0.96	0.98

5.4 Discussion

This study investigated the usefulness of item-score reliability methods to select items with the aim to produce either a longer or a shorter test. In practice, a test constructor may aim at a particular minimally acceptable test-score reliability or a maximally acceptable number of items. The results showed that the benchmark corrected item-total correlation was the best item-assessment method in both the bottom-up and top-down item selection procedure. This means that the frequently employed and simpler corrected item-total correlation is, next to method CA, one of the best item-assessment methods when constructing tests.

Because method CA computes the item-score reliability using the corrected item-total correlation (Equation 5.3), it is not surprising that these two item-assessment methods showed nearly identical results. Given these identical results, using the corrected item-total correlation seems more obvious in practice than using method CA, because the corrected item-total correlation is readily available in most statistical programs and using method CA would merely introduce a more elaborate method. However, this does not mean that item-score reliability does not contribute to the test construction process. Method CA is an estimation method to approximate item-score reliability, while the corrected item-total correlation measures the correlation between an item and the other items in the test. Method CA was developed to estimate item-score reliability, and to this means uses the corrected item-total correlation. The corrected item-total correlation is used to measure the coherence between an item and the other items in a test. This means that these two measures were developed and are used with a different purpose in mind.

In our study, once an item was selected for addition to the test or removal from the test, the selection result was irreversible. An alternative stepwise procedure might facilitate adding items to the test with the possibility of removing them again later in the procedure or removing items from the test with the possibility of adding them again later in the procedure. An alternative stepwise item selection procedure in combination with the item-assessment methods may produce an ordering closer to the ideal order than the bottom-up or top-down procedures. Such procedures are the topic of future research. Also, the frequently used assessment method "coefficient alpha if item deleted" was not considered in this study. This assessment method would be easily applicable in the top-down item selection procedure, but for the bottom-up item selection procedure we would have to come up with something like "coefficient alpha if item added". However, because the scope of this study was to investigate the construction of tests using assessment methods at the item level, and because coefficient alpha is on the test-level, we did not consider this assessment method. Also, we only studied one-dimensional data, because a test is assumed to measure one attribute. Deviations from one-dimensional data, which are unavoidable in practice because measurement in psychology is prone to systematic error, are the topic of future research.

Even though item-score reliability did not turn out to be a better item assessment method than the frequently employed corrected item-total correlation, it still has many useful applications. For example, when selecting a single item from a pool of items for constructing a single-item measure, item-score reliability can be used to ensure that the selected item has high item-score reliability. Single-item measures are often used in work and organizational psychology to asses job satisfaction (Gonzalez-Mulé et al., 2017; Harter et al., 2002; Nagy, 2002; Robertson & Kee, 2017; Saari & Judge, 2004; Zapf et al., 1999) or level of burnout (Dolan et al., 2014). Single-item measures have also been assessed in marketing research for measuring ad and brand attitude (Bergkvist & Rossiter, 2007) and in health research for measuring, for example, quality of life (Stewart et al., 1988; Yohannes et al., 2010) and psychosocial stress (Littman et al., 2006). Also, in person-fit analysis item-score reliability can be applied to identify items that contain too little reliable information to explain person fit (Meijer et al., 1995). This leaves many useful applications for item-score reliability.

5

Chapter 6

Epilogue

Main findings, consequences, and future research

This dissertation dealt with several aspects of item-score reliability. The literature was reviewed in Chapter 1 to create an overview of what had been investigated with regard to item-score reliability before. The review showed that in the past some attention was given to item-score reliability, but no thorough investigation addressing different methods and their statistical properties had been executed. Also, the main part of the literature concentrated on using the methods from the much cited studies by Wanous and Reichers (1996), Wanous et al. (1997), and Wanous and Hudy (2001) for the construction of single-item measures, because their method was developed for estimating the reliability of a single-item measure. Their methods and studies are used primarily to investigate whether single-item measures have sufficiently high item-score reliability.

In Chapter 2 of this dissertation, available methods in the literature to estimate item-score reliability were reviewed. One useful method resulting from this review was a method based on the correction for attenuation formula (Lord & Novick, 1968; Nunnally & Bernstein, 1994; Spearman, 1904), first employed by Wanous and Reichers (1996). We called it method CA. Three new methods were developed in Chapter 2, based on existing methods to estimate test-score reliability: method MS, method λ_6, and method LCRC. A simulation study showed that method MS and method CA had the smallest bias and the greatest precision. Method λ_6 showed negative bias, but great precision. Finally, method LCRC also showed small bias, but estimates were not very precise, rendering it an unreliable method. From this study, we concluded that methods MS, λ_6, and CA were promising for future research. Questions that arose from this research were what values could be expected of item-score reliability estimations in practice, as well as what a realistic cutoff value would be for an item having sufficient reliability. Another question was what the relationship between item-score reliability and other available item indices not assessing the item-score reliability is, in particular, the corrected item-total correlation, the item-factor loading, the item scalability, and the item discrimination. This led to the study in Chapter 3.

The study in Chapter 3 used several empirical-data sets to investigate values that could be expected in practice for three item-score reliability estimation methods, which were methods MS, λ_6, and CA. Next to the item-score reliability, four other item indices addressing item features other than reliability were estimated: the corrected item-total correlation, the item-factor loading, the item scalability, and the item discrimination. Scatterplots of all possible combinations of item-score reliability methods and item indices showed a strong association between the corrected-item total correlation and the item-score reliability methods, especially method CA. The latter is not suprising, because method CA uses the corrected item-total correlation to estimate item-score reliability. Based on the cutoff values for the other item indices described in the literature, a cutoff value for item-score reliability estimates could be determined. We concluded that an estimated item-score reliability of $.3$ indicated that an item score was sufficiently reliable. But, because in this study empirical-data sets are used, we did not know the true value of population item-score reliability $\rho_{ii'}$. This raised questions about what value of $\rho_{ii'}$ could be expected for an estimated item-score reliability of $.3$ by means of the approximation methods MS, λ_6, and CA. Another question was what exactly the relationship is between $\rho_{ii'}$ and the four item indices. To answer these questions an additional simulation study was needed. The study in Chapter 4 of this dissertation further investigated the relationship between $\rho_{ii'}$ and the three item-score reliability methods MS, λ_6, and CA. Also, the relationship between $\rho_{ii'}$ and the corrected item-total correlation, the item-factor loading, the item scalability, and the item discrimination was investigated. Finally, the feasibility of a cutoff value of $.3$ indicating sufficient reliability for an item score was further examined. For a range of $\rho_{ii'}$ values and for various conditions, estimates of the three item-score reliability approximation methods were investigated to see what values for $\rho_{ii'}$ one can expect for the previously determined cutoff value of $.3$. All item-score reliability estimation methods showed increasing bias for increasing values of $\rho_{ii'}$. Also, the results showed a one-to-one relationship between the item factor-loading and $\rho_{ii'}$. The lower bound of $.3$ seemed to be too stringent in practice.

The relationship between test-score reliability and item-score reliability was further investigated in the study in Chapter 5. The scope of the research was item selection in test construction based on item-score reliability. The goal was to construct a test that was as reliable as possible, based on including or excluding items that are reliable or unreliable, respectively. The items were selected based on their item-score reliability, and the order in which they were selected was compared to the ordering that would ideally be followed if the test-score reliability was known. Next to the estimates based on the three item-score reliability approximation methods, the corrected item-total correlation was used as an item assessment method. The corrected item-total correlation was added as a benchmark method, because it is often employed in practice. The results showed that the ideal order was closer approximated by the item assessment methods when the sample size was larger and when variance of the item discrimination parameters was greater. Method CA and the corrected item-total correlation turned out to be the best item assessment methods for constructing reliable tests; they showed the closest resemblance to the ideal ordering.

From this thesis several lessons can be learned about item-score reliability. Several studies have touched upon the subject of item-score reliability, but its estimation methods and their statistical properties, such as bias and precision, had never been investigated before. One important aspect of estimating item-score reliability is that it is impossible to estimate the reliability of a single-item measure using data from only the target item. All methods discussed in this dissertation use the other items in the test to estimate item-score reliability. As was shown in the literature review in Chapter 1, until now, item-score reliability has mainly been discussed in the context of single-item measures. Even though most authors mention problems that could arise when using a single-item measure, a frequently used counter argument is that short scales may sometimes just be as reliable as long instruments (Burisch, 1984). Especially when the attribute the single-item measure taps into is clearly defined, homogeneous, and theoretically deduced, Loo and Kelts (1998) argue that it is an appropriate measure. Therefore, using a single-item measure for a broad and heterogeneous attribute is not advised by Loo (2002). Gardner, Cummings, Dunham, and Pierce (1998) compared multi-item measures of psychological constructs with single-item measures of the same psychological constructs and concluded that neither type of measure came out of the comparison as the clearly better option. They recommend that researchers use the type of measure that is suitable for the research question at hand. We nevertheless wish to argue that single-item measures are inferior to multi-item measures, and even though researchers often use single-item measures, this does not seem an uncontroversial idea in practice. As noted before, (Spector, 1992, p. 4) calls “yes” or “no” single-item measures *notoriously* unreliable, because responses are not consistent over time. Next to that, when adding more (parallel) items in a test, the true-score variance will increase quadratically, while the error-score variance will increase linearly (Lord & Novick, 1968, p. 86). This means that the ratio of error-score variance and true-score variance will decrease, and thereby the test-score reliability will increase. We therefore argue that important decisions should not be made based on just a single measurement, because making decisions based on one item score seems to be too much prone to error. We conclude that item-score reliability is interesting, but should always be interpreted in a context of multi-item measures.

The literature review in Chapter 1 showed that method CA was used in several studies. The results in Chapter 2 and 4 showed that method CA performed well with respect to bias and precision. This result means that method CA is a viable method for estimating item-score reliability. We noticed that the item-score reliability values found by studies using the method developed by Wanous and Reichers (1996) and Wanous et al. (1997), were remarkably higher than the values we found with simulated data. For example, Ginns and Barrie (2004) found item-score reliability values for their single-item scale measuring College Teaching Effectiveness of $.96$, and Postmes et al. (2012) found an item-score reliability value of $.76$. Based on the studies carried out in chapters 2, 3, and 4, we deem a value indicating sufficient item-score reliability of $.3$ realistic. Explanations for this difference could be that the other items in the test, used to estimate the reliability of a single item score, correlate highly with the item of interest, resulting in a high estimate of item-score reliability by method CA. This would mean that this condition was not explicitly defined as such in our simulation design, and therefore the high values found in empirical data by other studies is an interesting topic for future research.

During the execution of the research in this dissertation, an interesting question was how item-score reliability methods related to other, frequently used item indices that address item features other than reliability, albeit related to reliability. The research in chapters 3 and 4 addressed the relationship between item-score reliability and the item indices corrected item-total correlation, item factor-loading, item scalability, and item discrimination. Relationships between item-score reliability and the item indices were observed, specifically between method CA and the corrected item-total correlation, and between $\rho_{ii'}$ and the item-factor loading. The first relationship can be explained by the fact that method CA uses the corrected item-total correlation to approximate item-score reliability (see Equation 4.3). The second relationship, shown in Figure 4.2, was explained by the unidimensional data-generating model, which resulted in good model fit for the one-factor model. In future studies, more elaborate simulation studies investigating different research conditions, such as multidimensional data, might reveal more about the relationship between $\rho_{ii'}$ and the item-factor loading.

Another direction for future research suggested by this dissertation is the usability of item-score reliability. From the comparison between item-score reliability and other item indices addressing item features other than reliability (see chapters 3, 4, and 5) several relationships emerged. For example, in Chapter 5, the results showed that when selecting items to construct a test that is as reliable as possible, using the corrected item-total correlation gave the same results as using method CA. One can argue that item-score reliability therefore does not contribute much in addition to existing item indices. However, we argue that there actually are differences and there is a future for item-score reliability. The main difference between item-score reliability and the item indices addressing item features other than reliability is the goal with which they were developed. Item-score reliability methods approximate the repeatability of an item score, while for example the corrected item-total correlation measures the association between an item score and the sum score of the other items in the test. This means that other item quantities are assessed by the different item indices, and depending on the interest of the researcher, one does not make the other superfluous.

The results in this dissertation showed that we cannot expect the same values for item-score reliability as for test-score reliability. In Chapter 3 a lower bound of $.3$ was established, and in Chapter 4 this lower bound turned out to be maybe even too high. Because an item score contains much less information than a test score, it seems logical that its reliability is lower. Future research might investigate values to be expected for item-score reliability by means of investigating tests that have proven to be reliable in the past. What item-score reliability values are found for the items in such a test that has been proven to be reliable? This kind of research would shed a new light on values that are acceptable for item-score reliability.

Even though item-score reliability seems to be mainly used in the context of single-item measures, we believe there are more situations in which item-score reliability could be useful. Item-score reliability assesses the repeatability of an item score, which is an important quality feature of an item. Using individually reliable items in a test will increase the quality of the measurement of the comprehensive test performance as a whole. When the goal is to measure an attribute to make decisions about the respondent, it is of greatest importance that these measures are reliable. Investigating the item-score reliability of the different items in a test in addition to other item and test properties, is therefore interesting when a researcher wants to assess the quality of the measurements obtained by using a certain item. Based on the estimated item-score reliability values, researchers can assess how to value their obtained measurements. This might lead to omitting item scores, gathering more data, or caution when interpreting the results.

Finally, validity is also an important aspect of measurement. Reliability is a necessary condition for obtaining validity. However, when an item has high reliability, this does not necessarily mean it is also valid. Both construct and predictive validity are important aspects when making decisions based on measurements, and it is questionable whether the validity of a single-item measure is sufficient for basing decisions on a single measurement. Covering an entire attribute by means of a single-item measure seems impossible. To ensure the entire attribute is covered when measuring or predicting this attribute, multi-item measures are the better option.

Appendix A

Item-Score Reliability Methods Based on Cronbach's α and Guttman's λ_2

In this Appendix it is shown what coefficient α and coefficient λ_2 would look like in the context of a single item score.

Coefficient α

An item-score reliability coefficient based on coefficient α can be constructed as follows. Let $\tilde{\pi}^{\alpha}_{x(i),y(i')}$ be an approximation of $\pi_{x(i),y(i')}$ based on observable probabilities, such that replacing $\pi_{x(i),y(i')}$ in the right-hand side of Equation 3 by $\tilde{\pi}^{\alpha}_{x(i),y(i')}$ results in coefficient α; that is,

$$\alpha = \frac{\sum\limits_{i \neq j}\sum\limits_{x}\sum\limits_{y}\left[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}\right]}{\sigma^2_X} + \frac{\sum\limits_{i}\sum\limits_{x}\sum\limits_{y}\left[\tilde{\pi}^{\alpha}_{x(i),y(i')} - \pi_{x(i)}\pi_{y(i)}\right]}{\sigma^2_X}. \tag{A.1}$$

Van der Ark et. al (2011) showed that the numerator of the ratio on the right-hand side equals

$$\sum_{i}\sum_{x}\sum_{y}\left[\tilde{\pi}^{\alpha}_{x(i),y(i')} - \pi_{x(i)}\pi_{y(i)}\right] = Jm^2\bar{\pi}, \tag{A.2}$$

where $\bar{\pi}$ is the mean of the $J(J-1)m^2$ observable terms in the numerator of the first ratio in Equation 3,

$$\bar{\pi} = \frac{\sum\limits_{i \neq j}\sum\limits_{x}\sum\limits_{y}\left[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}\right]}{J(J-1)m^2}. \tag{A.3}$$

Hence, coefficient α equals

$$\alpha = \frac{\sum\limits_{i \neq j}\sum\limits_{x}\sum\limits_{y}\left[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}\right]}{\sigma^2_X} + \frac{Jm^2\bar{\pi}}{\sigma^2_X}. \tag{A.4}$$

Let w_i be an arbitrary weight with $w_i \geq 0$ and $\sum_i w_i = 1$. Coefficient α in Equation A.4 can also be written as

$$\alpha = \frac{\sum\limits_{i \neq j} \sum\limits_{x} \sum\limits_{y} \left[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}\right]}{\sigma_X^2} + \frac{\sum\limits_{i} w_i J m^2 \bar{\pi}}{\sigma_X^2}. \tag{A.5}$$

The aim of including w_i in the definition of α is to demonstrate identifiability problems in α for item scores. Consistent with Equation 4, for an item score i, Equation A.5 may be reduced to

$$\alpha_i = \frac{w_i \bar{\pi}}{\sigma_{X_i}^2}. \tag{A.6}$$

Because w_i is arbitrary, coefficient α for item scores is unidentifiable, which makes this item-score reliability coefficient unsuited for estimating item-score reliability. Note that a natural choice would be to have $w_i = 1$ for all i. In that case, the numerator of Equation A.6 is a constant and coefficient α for item scores is completely determined by the variance of the item.

Coefficient λ_2

A line of reasoning similar to that for coefficient α can be applied to coefficient λ_2. Let $\tilde{\pi}^{\lambda_2}_{x(i),y(i')}$ be an approximation of $\pi_{x(i),y(i')}$ based on observable probabilities, such that replacing $\pi_{x(i),y(i')}$ in Equation 3 by $\tilde{\pi}^{\lambda_2}_{x(i),y(i')}$ results in coefficient λ_2; that is,

$$\lambda_2 = \frac{\sum\limits_{i \neq j} \sum\limits_{x} \sum\limits_{y} \left[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}\right]}{\sigma_X^2} + \frac{\sum\limits_{i} \sum\limits_{x} \sum\limits_{y} \left[\tilde{\pi}^{\lambda_2}_{x(i),y(i')} - \pi_{x(i)}\pi_{y(i)}\right]}{\sigma_X^2}. \tag{A.7}$$

Van der Ark et. al (2011) showed that

$$\sum_{i} \sum_{x} \sum_{y} \left[\tilde{\pi}^{\lambda_2}_{x(i),y(i')} - \pi_{x(i)}\pi_{y(i)}\right] = \sqrt{\frac{J}{J-1} \sum\sum_{i \neq j} \left\{\sum_{x} \sum_{y} \left[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}\right]\right\}^2} = \gamma. \tag{A.8}$$

Hence, coefficient λ_2 equals

$$\lambda_2 = \frac{\sum\limits_{i \neq j} \sum\limits_{x} \sum\limits_{y} \left[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}\right]}{\sigma_X^2} + \frac{\gamma}{\sigma_X^2}. \tag{A.9}$$

Let w_{ixy} be an arbitrary weight with $w_{ixy} \geq 0$ and $\sum\limits_i \sum\limits_x \sum\limits_y w_{ixy} = m^2 J$. Using weights w_i, coefficient λ_2 in Equation A.9 can also be written as

$$\lambda_2 = \frac{\sum\limits_{i \neq j} \sum \sum\limits_x \sum\limits_y \left[\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}\right]}{\sigma_X^2} + \frac{\sum\limits_i w_i \gamma}{\sigma_X^2}, \tag{A.10}$$

Consistent with Equation 4, for an item score i, based on Equation A.10 we can consider

$$\lambda_{2_i} = \frac{w_i \gamma}{\sigma_{X_i}^2}. \tag{A.11}$$

Similar to the item version of coefficient α, the item version of coefficient λ_2 is unidentifiable because w_i can have multiple values, which renders this version of coefficient λ_2 not a candidate to estimate $\rho_{ii'}$. Setting w_i to 1 results in a coefficient that depends on the item variance, making it unsuited as a coefficient for item-score reliability.

Appendix B

Deriving the Item-Score Reliability and Test-Score Reliability from the Two-Parameter Logistic Model

Let a test consist of J items each with $m+1$ categories $0, 1, \ldots, m$; and let X_i ($i = 1, \ldots, J$) denote the score on item i. Let θ denote a latent trait with known distribution $G(\theta)$ with mean μ_θ and variance σ^2_θ. The two-parameter logistic model (2PLM) models dichotomous items; hence $m = 1$. Let α_i denote the discrimination parameter and let β_i denote the location parameter. In the 2PLM, $P(X_i = 1|\theta) \equiv P_{i\theta}$ is modeled as

$$P_{i\theta} = \frac{\exp\left[\alpha_i(\theta - \beta_i)\right]}{1 + \exp\left[\alpha_i(\theta - \beta_i)\right]}. \tag{B.1}$$

The first partial derivative of $P_{i\theta}$ with respect to θ equals

$$P'_{i\theta} = \frac{\partial P_{i\theta}}{\partial \theta} = \alpha_i P_{i\theta}(1 - P_{i\theta}) \tag{B.2}$$

(e.g., Baker, 1992, p. 81). Latent trait θ and true score T are related. Let $T_{i\theta}$ denote the item true score given a latent trait value. Under the classical test theory model, by definition, $T_{i\theta} = E(X_i|\theta)$ (Lord & Novick, 1968, p. 34). Furthermore, from straightforward algebra, it follows that $E(X_i|\theta) = \sum_{x=1}^{m} P(X_i \geq x|\theta)$, which reduces to $E(X_i|\theta) = P_{i\theta}$ for $m = 1$. Hence, in the 2PLM, $T_{i\theta} = P_{i\theta}$ (Lord, 1980, p. 46). Let $\sigma^2_{T_i}$ denote the variance of T_i. Following the delta method (e.g., Agresti, 2002, pp. 577–581),

$$\sigma^2_{T_i} \approx (P'_{i\mu_\theta})^2 \sigma^2_\theta. \tag{B.3}$$

Inserting the right-hand side of Equation B.1, in which θ has been replaced by μ_θ, into Equation B.2; and subsequently, inserting the right-hand side of Equation B.2 into Equation B.3 yields

$$\sigma^2_{T_i} \approx \alpha^2 \left(\frac{\exp(\alpha_i(\mu_\theta - \beta_i))}{1 + \exp(\alpha_i(\mu_\theta - \beta_i))}\right)^2 \left(1 - \frac{\exp(\alpha_i(\mu_\theta - \beta_i))}{1 + \exp(\alpha_i(\mu_\theta - \beta_i))}\right)^2 \sigma^2_\theta. \tag{B.4}$$

Let $P_i \equiv P(X_i = 1) = \int_\theta P_{i\theta} dG(\theta)$. Let $\sigma^2_{X_i}$ denote the variance of X_i. Because X_i is dichotomous,

$$\sigma^2_{X_i} = P_i(1 - P_i) = \int_\theta P_{i\theta} dG(\theta) \left(1 - \int_\theta P_{i\theta} dG(\theta)\right). \tag{B.5}$$

Let $P_{ij} \equiv P(X_i = 1, X_j = 1)$. Due to the local independence assumption in the 2PLM, $P_{ij} = \int_\theta P_{i\theta} P_{j\theta} dG(\theta)$. Let σ_{X_i,X_j} denote the covariance between X_i and X_j. In the classical test theory $\sigma_{X_i,X_j} = \sigma_{T_i,T_j}$ for $i \neq j$. Because X_i is dichotomous,

$$\sigma_{X_i,X_j} = \sigma_{T_i,T_j} = P_{ij} - P_i P_j = \int_\theta P_{i\theta} P_{j\theta} dG(\theta) - \int_\theta P_{i\theta} dG(\theta) \int_\theta P_{j\theta} dG(\theta). \tag{B.6}$$

Let

$$\sigma^2_T = \sum_{i=1}^{J} \sigma^2_{T_i} + \sum_{\substack{i=1 \\ i \neq j}}^{J} \sum_{j=1}^{J} \sigma_{T_i,T_j} \tag{B.7}$$

and

$$\sigma^2_X = \sum_{i=1}^{J} \sigma^2_{X_i} + \sum_{\substack{i=1 \\ i \neq j}}^{J} \sum_{j=1}^{J} \sigma_{X_i,X_j} \tag{B.8}$$

denote the true-score variance and test-score variance, respectively; where the item variances and covariances can be derived from the 2PLM using Equations B.4, B.5, and B.6.

Item-score reliability $\rho_{ii'} = \frac{\sigma^2_{T_i}}{\sigma^2_{X_i}}$ can be obtained from Equation B.4 and Equation B.5. Test-score reliability $\rho_{XX'} = \frac{\sigma^2_T}{\sigma^2_X}$ can be obtained from Equation B.7 and Equation B.8.

References

Agresti, A. (2002). *Categorical data analysis*. New York, NY: Wiley. doi: 10.1002/0471249688

Arvey, R. D., Landon, T. E., Nutting, S. M., & Maxwell, S. E. (1992). Development of physical ability tests for police officers: A construct validation approach. *Journal of Applied Psychology*, *77*, 996–1009. doi: 10.1037/0021-9010.77.6.996

Ashton, M. C., & Lee, K. (2001). A theoretical basis for the major dimensions of personality. *European Journal of Personality*, *15*, 327–353. doi: 10.1002/per.417

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*, 150–166. doi: 10.1177/1088868306294907

Ayala, R. D. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, *18*(2), 155–170. doi: 10.1177/014662169401800205

Baker, F. B. (1992). *Item response theory: parameter estimation techniques*. New York, NY: Marcel Dekker.

Baker, F. B. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assesment and Evaluation. Retrieved from `http://files.eric.ed.gov/fulltext/ED458219.pdf`

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: CRC Press. doi: 10.1201/9781482276725

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, *44*, 175–184. doi: 10.1509/jmkr.44.2.175

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical Theories of Mental Test Scores* (pp. 397-424). Reading, MA: Addison-Wesley.

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1985). *Revisie Amsterdamse Kinderintelligentie Test (RAKIT) [Revision of the Amsterdam Child Intelligence Test.]*. Lisse, The Netherlands: Swets & Zeitlinger.

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist*, *39*, 214–227. doi: 10.1037/0003-066x.39.3.214

Cavalini, P. M. (1992). *It's an ill wind that brings no good: Studies on odour annoyance and the dispersion of odour concentrations from industries* (Unpublished doctoral dissertation). University of Groningen, The Netherlands.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16*, 64–73. doi: 10.2307/3150876

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi: 10.1007/bf02310555

Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement*, *37*, 201–225. doi: 10.1177/0146621612470210

De Groot, A., & Van Naerssen, R. (1969). *Studietoetsen: Construeren, afnemen, analyseren. [Educational testing: Construction, administration, analysis.]*. The Hague, The Netherlands: Mouton.

De Jong Gierveld, J., & Van Tilburg, T. G. (1999). Manual of the loneliness scale. *Vrije Universiteit Amsterdam, Department of Social Research Methodology*. Retrieved from `https://research.vu.nl/ws/portalfiles/portal/1092113`

De Koning, E., Sijtsma, K., & Hamers, J. H. M. (2003). Construction and validation of a test for inductive reasoning. *European Journal of Psychological Assessment*, *19*, 24–39. doi: 10.1027//1015-5759.19.1.24

Denollet, J. (2005). DS14: Standard assessment of negative affectivity, social inhibition, and type D personality. *Psychosomatic Medicine*, *67*, 89–97. doi: 10.1097/01.psy.0000149256.81953.49

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*, 71–75. doi: 10.1207/s15327752jpa4901_13

Dolan, E. D., Mohr, D., Lempa, M., Joos, S., Fihn, S. D., Nelson, K. M., & Helfrich, C. D. (2014). Using a single item to measure burnout in primary care staff: A psychometric evaluation. *Journal of General Internal Medicine*, *30*, 582–587. doi: 10.1007/s11606-014-3112-6

Dolbier, C. L., Webster, J. A., McCalister, K. T., Mallon, M. W., & Steinhardt, M. A. (2005). Reliability and validity of a single-item measure of job satisfaction. *American Journal of Health Promotion*, *19*, 194–198. doi: 10.4278/0890-1171-19.3.194

Drolet, A. L., & Morrison, D. G. (2001). Rejoinder to grapentine. *Journal of Service

Research, *4*, 159–160. doi: 10.1177/109467050142008

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum. doi: 10.4324/9781410605269

Erhart, M., Hagquist, C., Auquier, P., Rajmil, L., Power, M., Ravens-Sieberer, U., & Group, E. K. (2010). A comparison of rasch item-fit and Cronbach's alpha item reduction analysis for the development of a quality of life scale for children and adolescents. *Child: Care, Health and Development*, *36*, 473–484. doi: 10.1111/j.1365-2214.2009.00998.x

Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft*, *69*(2), 195.

Gardner, D. G., Cummings, L. L., Dunham, R. B., & Pierce, J. L. (1998). Single-item versus multiple-item measurement scales: An empirical comparison. *Educational and Psychological Measurement*, *58*, 898–915. doi: 10.1177/0013164498058006003

Gignac, G. E., & Wong, K. K. (2018). A psychometric examination of the anagram persistence task: More than two unsolvable anagrams may not be better. *Assessment*. doi: 10.1177/1073191118789260

Ginns, P., & Barrie, S. (2004). Reliability of single-item ratings of quality in higher education: A replication. *Psychological Reports*, *95*, 1023–1030. doi: 10.2466/pr0.95.3.1023-1030

Gonzalez-Mulé, E., Carter, K. M., & Mount, M. K. (2017). Are smarter people happier? Meta-analyses of the relationships between general mental ability and job and life satisfaction. *Journal of Vocational Behavior*, *99*, 146–164. doi: 10.1016/j.jvb.2017.01.003

Gorsuch, R. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum. doi: 10.4324/9780203781098

Gough, H. G., & Heilbrun Jr., A. B. (1980). *The Adjective Check List, manual 1980 edition*. Palo Alto, CA: Consulting Psychologists Press.

Gustafsson, J.-E. (1977). *The Rasch model for dichotomous items: Theory, applications and a computer program*. Unpublished Report. Retrieved from `https://eric.ed.gov/?id=ED154018`

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255–282. doi: 10.1007/bf02288892

Guttman, L. (1946). The test-retest reliability of qualitative data. *Psychometrika*, *11*(2), 81–95. doi: 10.1007/bf02288925

Hagenaars, J. A. P., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press. doi: 10.1017/cbo9780511499531.001

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. New York, NY: Springer. doi: 10.1007/978-94-017-1988-9

Harman, H. H. (1976). *Modern factor analysis*. Chicago, IL: University of Chicago Press.

Harter, J. K., Schmidt, F. L., & Hayes, T. L. (2002). Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: A meta-analysis. *Journal of Applied Psychology*, *87*, 268–279. doi: 10.1037/0021-9010.87.2.268

Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review*, *34*, 93. doi: 10.2307/2092790

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika*, *42*, 567–578. doi: 10.1007/BF02295979

Knapp, T. R. (1977). The reliability of a dichotomous test-item: A "correlationless" approach. *Journal of Educational Measurement*, *14*, 237–252. doi: 10.1111/j.1745-3984.1977.tb00041.x

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Studies in social psychology in World War II. vol. IV: Measurement and prediction* (pp. 362–412). Princeton, NJ: Princeton University Press.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*, 1–29. doi: 10.18637/jss.v042.i10

Littman, A. J., White, E., Satia, J. A., Bowen, D. J., & Kristal, A. R. (2006). Reliability and validity of 2 single-item measures of psychosocial stress. *Epidemiology*, *17*, 398–403. doi: 10.1097/01.ede.0000219721.89552.51

Loo, R. (2002). A caveat on using single-item versus multiple-item scales. *Journal of Managerial Psychology*, *17*, 68–75. doi: 10.1108/02683940210415933

Loo, R., & Kelts, P. (1998). A caveat on using single-item measures. *Employee Assistance Quarterly*, *14*, 75–80. doi: 10.1300/j022v14n02_06

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum. doi: 10.4324/9780203056615

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi: 10.1007/bf02296272

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA: Sage. doi: 10.4135/9781412984713

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18*, 111–120. doi: 10.1177/014662169401800202

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, *8*, 261–272. doi: 10.1207/s15324818ame0803_5

Meijer, R. R., Sijtsma, K., & Molenaar, I. W. (1995). Reliability estimation for single dichotomous items based on Mokken's IRT model. *Applied Psychological Measurement*, *19*, 323–335. doi: 10.1177/014662169501900402

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, *14*, 283–298. doi: 10.1177/014662169001400306

Melián-González, S., Bulchand-Gidumal, J., & López-Valcárcel, B. G. (2015). New evidence of the relationship between employee satisfaction and firm economic performance. *Personnel Review*, *44*, 906–929. doi: 10.1108/pr-01-2014-0023

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. Berlin, Germany: Walter de Gruyter. doi: 10.1515/9783110813203

Molenaar, I., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve methoden*, *9*, 115–126.

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology*, *75*, 77–86. doi: 10.1348/096317902167658

Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods*, *23*, 351–362. doi: 10.1037/met0000132

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, *14*, 485–500. doi: 10.1207/s15327906mbr1404_7

Oosterwijk, P. R., Van der Ark, L. A., & Sijtsma, K. (2017). Overestimation of reliability by Guttman's λ_4, λ_5, and λ_6 and the Greatest Lower Bound. In *Springer proceedings in mathematics & statistics* (pp. 159–172). Springer International Publishing. doi: 10.1007/978-3-319-56294-0_15

Pavot, W., & Diener, E. (1993). Review of the satisfaction with life scale. *Psychological Assessment*, *5*, 164–172. doi: 10.1037//1040-3590.5.2.164

Postmes, T., Haslam, S. A., & Jans, L. (2012). A single-item measure of social identification: Reliability, validity, and utility. *British Journal of Social Psychology*, *52*, 597–617. doi: 10.1111/bjso.12006

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Raubenheimer, J. (2004). An item selection procedure to maximize scale reliability and validity. *SA Journal of Industrial Psychology*, *30*, 59–64. doi: 10.4102/sajip.v30i4.168

Raykov, T. (2007). Reliability if deleted, not 'alpha if deleted': Evaluation of scale reliability following component deletion. *British Journal of Mathematical and Statistical Psychology*, *60*, 201–216. doi: 10.1348/000711006x115954

Raykov, T. (2008). Alpha if item deleted: A note on loss of criterion validity in scale development if maximizing coefficient alpha. *British Journal of Mathematical and Statistical Psychology*, *61*, 275–285. doi: 10.1348/000711007x188520

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25. doi: 10.18637/jss.v017.i05

Robertson, B. W., & Kee, K. F. (2017). Social media at work: The roles of job satisfaction, employment status, and facebook use with co-workers. *Computers in Human Behavior*, *70*, 191–196. doi: 10.1016/j.chb.2016.12.080

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the rosenberg self-esteem scale. *Personality and Social Psychology Bulletin*, *27*, 151–161. doi: 10.1177/0146167201272002

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press. doi: 10.1515/9781400876136

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. doi: 10.18637/jss.v048.i02

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International journal of research in marketing*, *19*, 305–335. doi: 10.1016/s0167-8116(02)00097-6

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*, 493–502. doi: 10.1037/0022-3514.57.3.493

Saari, L. M., & Judge, T. A. (2004). Employee attitudes and job satisfaction. *Human Resource Management*, *43*, 395–407. doi: 10.1002/hrm.20032

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85–100). Springer.

Shamir, B., & Kark, R. (2004). A single-item graphic scale for the measurement of organizational identification. *Journal of Occupational and Organizational Psychology, 77*, 115–123. doi: 10.1348/0963179043229l5946

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22*, 3–31. doi: 10.1177/01466216980221001

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika, 74*, 107–120. doi: 10.1007/s11336-008-9101-0

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52*, 79–97. doi: 10.1007/bf02293957

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage. doi: 10.4135/9781412984676

Sijtsma, K., & Van der Ark, L. A. (2015). Conceptions of reliability revisited and practical recommendations. *Nursing Research, 64*, 128–136. doi: 10.1097/nnr.0000000000000077

Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*, 137–158. doi: 10.1111/bmsp.12078

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology, 15*, 72–101.

Spector, P. E. (1992). *Quantitative applications in the social sciences: Summated rating scale construction*. Newbury Park, CA: Sage. doi: 10.4135/9781412986038

Stewart, A. L., Hays, R. D., & Ware, J. E. (1988). The MOS short-form general health survey. *Medical Care, 26*, 724–735. doi: 10.1097/00005650-198807000-00007

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Pearson.

Taylor, J. A. (1953). A personality scale of manifest anxiety. *The Journal of Abnormal and Social Psychology, 48*, 285–290.

Tucker, L. R (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1–13. doi: 10.1007/bf02288894

Van den Berg, P. T. (1992). Persoonlijkheid en werkbeleving: De validiteit van persoonlijkheidsvragenlijsten, in het bijzonder die van een spanningsbehoeftelijst [personality and work experience: The validity of personality questionnaires, and in particular the validity of a sensation-seeking questionnaire.]. *Unpublished Ph. D. thesis, Vrije Universiteit, Amsterdam*.

Van den Brink, W., & Mellenbergh, G. J. (1998). *Testleer en testconstructie [Testing and test construction]*. Amsterdam, The Netherlands: Boom.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*, 1–19. doi: 10.18637/jss.v020.i11

Van der Ark, L. A. (2010). Computation of the Molenaar Sijtsma statistic. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 775–784). Berlin, Germany: Springer. doi: 10.1007/978-3-642-01044-6_71

Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*, 1–27. doi: 10.18637/jss.v048.i05

Van der Ark, L. A., Van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement, 35*, 380–392. doi: 10.1177/0146621610392911

Van der Veen, G. (1992). *Principes in praktijk: CNV-leden over collectieve acties [Principles into practice. Labour union members on means of political pressure]*. Kampen, The Netherlands: J H Kok.

Van Maanen, L., Been, P., & Sijtsma, K. (1989). The linear logistic test model and heterogeneity of cognitive strategies. In E. E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 267–287). Berlin, Germany: Springer. doi: 10.1007/978-3-642-83943-6_17

Verweij, A. C., Sijtsma, K., & Koops, W. (1999). An ordinal scale for transitive reasoning by means of a deductive strategy. *International Journal of Behavioral Development, 23*, 241–264. doi: 10.1080/016502599384099

Vingerhoets, A. J. J. M., & Cornelius, R. R. (Eds.). (2001). *Adult crying: A biopsychosocial approach*. Hove, UK: Brunner-Routledge. doi: 10.4324/9780203717493

Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods, 4*, 361–375. doi: 10.1177/109442810144003

Wanous, J. P., & Reichers, A. E. (1996). Estimating the reliability of a single-item measure. *Psychological Reports, 78*, 631–634. doi: 10.2466/pr0.1996.78.2.631

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology, 82*, 247–252. doi: 10.1037//0021-9010.82.2.247

Weiss, D. J. (1976). Multivariate procedures. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 327–362). Chicago, IL: Rand McNally. doi: 10.4135/9781848608320

Williams, G., Thomas, K., & Smith, A. P. (2017). Stress and well-being of university staff: An investigation using the demands-resources- individual effects (DRIVE) model and well-being process questionnaire (WPQ). *Psychology*, *8*, 1919–1940. doi: 10.4236/psych.2017.812124

Yohannes, A. M., Willgoss, T., Dodd, M., Fatoye, F., & Webb, K. (2010). Validity and reliability of a single-item measure of quality of life scale for patients with cystic fibrosis. *Chest*, *138*, 507A. doi: 10.1378/chest.10254

Zapf, D., Vogt, C., Seifert, C., Mertini, H., & Isic, A. (1999). Emotion work as a source of stress: The concept and development of an instrument. *European Journal of Work and Organizational Psychology*, *8*, 371–400. doi: 10.1080/135943299398230

Zegers, F. E., & Ten Berge, J. M. (1985). A family of association coefficients for metric scales. *Psychometrika*, *50*, 17–24. doi: 10.1007/bf02294144

Zijlmans, E. A. O., Tijmstra, J., Van der Ark, L. A., & Sijtsma, K. (2018a). Investigating the relationship between item-score reliability, its estimation methods, and other item indices. *Manuscript submitted for publication*.

Zijlmans, E. A. O., Tijmstra, J., Van der Ark, L. A., & Sijtsma, K. (2018b). Item-score reliability in empirical-data sets and its relationship with other item indices. *Educational and Psychological Measurement*, *78*, 998–1020. doi: 10.1177/0013164417728358

Zijlmans, E. A. O., Van der Ark, L. A., Tijmstra, J., & Sijtsma, K. (2018). Methods for estimating item-score reliability. *Applied Psychological Measurement*, *42*, 553–570. doi: 10.1177/0146621618758290

Zimmerman, M., Ruggero, C. J., Chelminski, I., Young, D., Posternak, M. A., Friedman, M., . . . Attiullah, N. (2006). Developing brief scales for use in clinical practice: The reliability and validity of single-item self-report measures of depression symptom severity, psychosocial impairment due to depression, and quality of life. *The Journal of Clinical Psychiatry*, *67*, 1536–1541. doi: 10.4088/jcp.v67n1007

Summary

This dissertation deals with item-score reliability. The goal was to create an overview of studies that had touched upon this subject so far, to develop and investigate the performance of methods to approximate item-score reliability, and to evaluate the usefulness of item-score reliability in practice.

A literature review was performed, where the current status of methods to estimate item-score reliability and the use of these methods was assessed. The main part of the literature used item-score reliability as a quantity to justify the use of single-item measures. The method that was used most frequently was the one developed by Wanous and Reichers (1996), based on the correction for attenuation (Nunnally & Bernstein, 1994, p. 257), which we refer to as method CA.

The available method CA was considered in the next study, as well as three test-score reliability estimation methods that were adjusted such that they could be used for approximating the reliability of an item score instead of the reliability of a test score. This led to four methods for estimating item-score reliability: method MS, Guttman's method λ_6, the latent-class reliability coefficient (method LCRC), and method CA. A simulation study was performed to compare the methods with respect to median bias, variability (inter-quartile range; IQR) and percentage of outliers. The simulation study consisted of six conditions: standard, polytomous items, unequal α-parameters, two-dimensional data, long test, and small sample size. Methods MS and CA showed to be the most accurate. Method LCRC showed almost unbiased results, but large variability. Method λ_6 consistently underestimated item-score reliability, but showed smaller IQR than the other methods.

In the next study, the three most promising item-score reliability methods (methods MS, λ_6, and CA) were compared to four well-known and widely accepted item indices, assessing item quantities other than reliability. These item indices were the corrected item total-correlation, the item factor-loading, the item scalability, and the item discrimination. Values that can be expected for item-score reliability in real-data sets were investigated, and the relation between the three item-score reliability methods and the four item indices were investigated. Based on the cutoff values of these item indices, a lower bound of $.3$ for item-score reliability was proposed.

A fourth study further investigated the relationship between item-score reliability, its estimation methods, and the four item indices assessing other quantities than reliability. Also, the feasibility of a lower bound for item-score reliability estimates of $.3$ was further investigated. In a simulation study, the item's difficulty parameter, variance of the other items' location parameters, and number of items in the test were varied. All methods showed increasing bias for higher item-score reliability values. Method CA showed good results for items with a non-deviant location parameter. The results showed a one-to-one relationship between the item factor-loading and item-score reliability. It was concluded that a lower bound of $.3$ seemed to be too high in practice.

In a final study, the usability of item-score reliability as a criterion for item selection in test construction was investigated. Item-score reliability methods MS, λ_6, and CA were taken into account and compared to the corrected item-total correlation, which was added as a benchmark method. An ideal ordering to add items to the test (bottom-up procedure) or omit items from the test (top-down procedure) was defined based on the population test-score reliability. The orderings the four item-assessment methods produced in samples were compared to the ideal ordering, and the degree of resemblance was expressed by means of Kendall's τ. To investigate the concordance of the orderings across 1000 replicated samples, Kendall's W was computed for each item-assessment method. The results showed that for both the bottom-up and the top-down procedure, item-assessment method CA and the corrected item-total correlation most closely resembled the ideal ordering. Generally, all item assessment methods resembled the ideal ordering better, and concordance of the orderings was greater, for larger sample sizes and greater variance of the item discrimination parameters.

It can be learned from the studies in this dissertation that it is impossible to estimate the item-score reliability of a test containing a single item; One always needs other items in the test to estimate item-score reliability of a particular item, or another test allegedly measuring the same attribute as the target item. Even though until now item-score reliability is mainly used for estimating the reliability of single-item measures, it was argued that multi-item measures are the better option, because making important decisions based on single-item measures seems to be too much prone to error. Therefore, item-score reliability is a useful tool for assessing the repeatability of an item score, and thereby the quality of the item.

Acknowledgements

Ondanks dat mijn naam als auteur op de kaft van dit boekje prijkt, had ik dit proefschrift nooit kunnen schrijven zonder de hulp van een aantal mensen. In de eerste plaats zijn er natuurlijk mijn drie (co-)promotores. Begeleid worden door drie personen klinkt misschien als een uitdaging. Desalniettemin was ik na een afspraak met zijn vieren altijd weer gemotiveerd om verder te gaan werken aan mijn onderzoek.

Klaas, toen ik begon aan mijn proefschrift was je nog erg druk als decaan. Ondanks dat ben je altijd erg betrokken geweest bij mijn project. Je deur stond altijd open, mocht ik hulp nodig hebben. In mijn laatste jaar, tijdens jouw sabbatical, heb je er mede voor gezorgd dat het proefschrift op tijd af kwam en dat ik niet te lang bleef zwoegen op dezelfde stukken. Dank voor al je adviezen en de fijne begeleiding.

Andries, toen ik kwam werken in Tilburg was jij net vertrokken naar Amsterdam. Je kreeg het steeds drukker en gaf aan soms het gevoel te hebben me in de steek te laten, maar dat gevoel heb ik nooit gedeeld. Zodra ik een vraag had, vaak over wiskunde of een stuk code, kreeg ik snel antwoord en kon ik weer op weg. Bedankt voor alles.

Jesper, we leerden elkaar kennen in Utrecht toen je mijn tutor werd tijdens de master. Deze rol heb je ook vervuld in de afgelopen vier jaar. Ik kon altijd bij je binnen lopen met vragen of voor adviezen, van onderzoek tot hoe om te gaan met alles wat bij het schrijven van een proefschrift komt kijken. Bedankt hiervoor.

De leescommissie wil ik graag bedanken voor het voor het lezen en beoordelen van mijn proefschrift en voor het naar Tilburg reizen voor de verdediging. Het IOPS wil ik graag bedanken voor alle interessante congressen en cursussen en voor alle leuke contacten die ik daar heb gelegd. Daarnaast was het erg leerzaam om twee jaar plaats te mogen nemen tijdens de bestuursvergaderingen als (assistent) PhD representative. Ook de TSB PhD council en TiPP wil ik bedanken voor de fijne samenwerking en de leerzame ervaringen.

Wanneer ik mensen vertelde over wat ik nu precies deed op de universiteit oogstte ik vaak medelijdende blikken. Ze leken dan wel weer enigszins gerust wanneer ik zei: “Maar ik heb wel hele leuke collega’s!”. Het schrijven van een proefschrift gaat met pieken en dalen, maar er is in de afgelopen vier jaar geen dag geweest dat ik met tegenzin naar mijn werk ben gegaan. De leuke lunches (met gespreksonderwerpen die binnen no-time uit de hand liepen), boswandelingen, borrels, PhD-trips, outings, pubquizzen en gesprekken bij de koffieautomaat ga ik enorm missen.

Mijn kamergenoten, Davide, Laura, Sara en later ook Andrea, bedankt voor de leuke jaren, eerst in P 1.113 en later in S 709. Jules en Niek, bedankt voor de inspirerende boswandelingen. Robbie, bedankt voor al je hulp met mijn code, als ik weer eens code wil parallelliseren weet ik je te vinden. Marieke en Anne-Marie, bedankt voor al jullie ondersteuning en de gezellige praatjes. Chris, Elise, Erwin, Esther, Florian, Geert, Hilde, IJsbrand, Inga, Jaap-Joris, Leonie, Lianne, Mattis, Michèle, Ylva, Paul, Paulette, Pieter, Robert, Soogeun en Zhengguo bedankt voor alle gezelligheid!

Aan de tijd gedurende mijn master in Utrecht heb ik erg leuke herinneringen. Het was voor mij een moeilijke tijd en ik denk dat mijn studiegenoten niet door hebben gehad hoe zij mij door een lastige periode gesleept hebben. Samen dagenlang studeren op de Uithof en samen ervoor zorgen dat we uiteindelijk allemaal verder kwamen; ik ben blij dat we nog steeds contact hebben. Bente, Danielle, Jolien, Kees, Kirsten, Laura, Mariëlle, Rob, Sanne en Timo, bedankt!

Tijdens mijn master mocht ik, als student-assistent, verschillende onderzoekswerkzaamheden uitvoeren. Peter, Edith en Joop, bedankt dat ik met jullie aan onderzoek mocht werken. Ik vond het erg leerzaam en bovenal ook gezellig!

Mijn vriendinnen wil ik graag bedanken voor alle afleiding die zij mij de afgelopen jaren geboden hebben. Sommige van hen ken ik al zo lang als ik me kan herinneren, anderen kwamen later. Ik ben dankbaar voor de langdurige vriendschappen, de gezelligheid en de herinnering dat er meer is in het leven dan het schrijven van een proefschrift.

Lieve Bente en Sanne, mijn paranimfen. Op de dag van mijn verdediging zullen jullie achter mij staan om het van me over te nemen mocht ik bezwijken onder de druk. De afgelopen jaren waren jullie eigenlijk ook al mijn paranimfen: wanneer het tegen zat kon ik altijd op jullie rekenen. Bedankt voor alle steun maar ook zeker voor alle gezelligheid! Verder wil ik hier graag nog Stephanie bedanken, mijn buurmeisje. We kennen elkaar al ons hele leven en je bent er altijd voor mij geweest, vooral tijdens moeilijke momenten de afgelopen jaren. Ik hoop dat onze vriendschap nog heel erg lang mag duren.

Mark, Renske, Roy, Niek, Anne, mijn gezin. Bedankt voor alle geboden steun de afgelopen jaren, in welke vorm dan ook. Jullie hebben ervoor gezorgd dat ik altijd met beide benen op de grond ben blijven staan. Lieve Fleur, Britt, Mick en Evy, bedankt voor al jullie lachjes, grappige acties en vooral voor jullie vrolijkheid.

Ik vind het heel leuk om jullie tante te mogen zijn.

Beste Jos en Odie, lieve papa en mama. Jullie hebben me altijd mijn gang laten gaan en me vrij gelaten om te doen wat ik wilde. Het behoeft geen uitleg dat zonder jullie dit boekje er niet had gelegen en ik nooit zo ver was gekomen. Bedankt dat ik jullie altijd mag bellen en voor alles dat jullie voor me hebben gedaan, in het bijzonder in de afgelopen jaren.

Lieve Michiel, tijdens onze studie in Utrecht probeerde je me altijd over te halen een vak in programmeren te komen volgen aan jouw faculteit. Wie wist dat ik dit uiteindelijk een van de leukste aspecten van mijn studie en later mijn werk zou gaan vinden. Wat had ik graag gewild dat je hier bij kon zijn. Ik mis je.

Lieve Koen. Samen hebben we de afgelopen jaren heel wat dalen maar zeker ook pieken beleefd. Tijdens een groot gedeelte van mijn PhD zat je in het buitenland. Ik ben heel blij dat je weer thuis bent en dat we aan ons leven samen kunnen gaan beginnen. Waar jij bent voel ik me thuis. Ik hou van je.

Eva Zijlmans, *Terschelling, december 2018*