Bergische Universität Wuppertal

Fakultät für Human- und Sozialwissenschaften, Psychologie

Methodenlehre und Psychologische Diagnostik, Prof. Dr. Ralf Schulze

An Investigation of Empirical Scoring Methods for Ability Measurement

Dissertation zur Erlangung des akademischen Grades Doktor der Naturwissenschaften (Dr. rer. nat.) an der Fakultät für Human- und Sozialwissenschaften der Bergischen Universität Wuppertal.

Angenommen am 08. Juni 2018.

Dissertation verfasst von

Anna-Lena Jobmann

Die Dissertation kann wie folgt zitiert werden:

urn:nbn:de:hbz:468-20180724-151304-8
[http://nbn-resolving.de/urn/resolver.pl?urn=urn%3Anbn%3Ade%3Ahbz%3A468-20180724-151304-8]

Acknowledgements

Thanks to the enduring support of a number of people I was able to accomplish this research project. I would like to thank Prof. Dr. Ralf Schulze, who gave me the perfect amount of freedom to process this project while at the same time he was constantly available for advice. He has likewise offered opportunities of scientific exchange and collaborations in different projects which very much helped to get along with the more challenging parts of this work. I also would like to thank Prof. Dr. Susanne Buch for taking over the secondary supervisory. Thanks to my colleagues Lisa Roese, Dr. Susan Hellwig and Dr. Philipp Engelberg for sharing their precious views and experiences. I also would particularly like to thank my colleagues Maike Pisters and Sally Ischebeck for always having encouraging and motivating words, and for always expecting the best of me. Thank you also to the student assistants, who supported several different parts of this work in the last years. I have been fortunate to work with you all.

For the constant and caring emotional support, I want to thank my family, my friends and, in particular, my husband Daniel. I am very grateful for your patience, your protective calmness and your firm loyalty. Thank you for understanding when I need time on my own, for being there, when I need time with you. Many thanks to my parents, Nicole and Olof, who let me grew up so liberally and safely giving me the confidence for this project.

Contents

Abstract

Measurement of new intelligences frequently relies on empirical scoring methods such as consensus based measurement (CBM). As the correctness of response options is not established on the basis of theories, but of empirical standards, these methods have been questioned in the past to indicate the true correctness. The present studies aim to systematically investigate CBM methods as well as an enlarged pool of supposed alternative methods - Consensus Analysis, HOMALS and the Nominal Response Model (NRM) - based on data for which the true scoring keys are known. For the systematic evaluation of the empirical scoring methods, two studies were conducted, one using simulated data, the other using real-world data. In the simulation studies, several characteristics of two- and five-categorical data were manipulated to investigate the influence of the independent variables sample size, number of items, ability of respondents, and difficulty of items. The methods were evaluated with dependent variables indicating the relative distance of true and reconstructed scoring keys and the correlation between true abilities and abilities based on the respective method. The results indicate that ability and difficulty are the key influencing variables for the consensus-based scoring methods. With high ability samples and low difficulty of items these methods performed well. CBM and Consensus Analysis showed only minor differences. The NRM was mainly independent of the manipulated variables, but was only observed to work when two response options were fixed by making plausible assumptions. HOMALS was not observed to provide satisfactory results in any realized data condition. The real-world data study used responses to 18 mathematics TIMSS 2011 items ($N = 15,992$). Consensus-based scoring methods as well as NRM with plausible fixation worked perfectly, whereas HOMALS and NRM with random fixation did not provide the correct scoring key. Moreover, dependence on the ability was supported. It is concluded that empirical scoring methods are only recommended for specific data conditions. The results support the need for more finely elaborated theories for the measurement of intelligences.

1 Introduction

Human intelligence is one of the most prominent as well as most important constructs in psychological research. It has been shown to be a successful predictor of real life outcomes, such as education, income, and health variables (Deary, 2012). The definitions, theories, and models of intelligence are heterogeneous, however, the construct is consistently theorized to include various interrelated subcomponents (Carroll, 1993; Cattell, 1963; Guilford, 1967; Jäger, 1967; Thurstone, 1938). Measuring most types of intelligence has been studied extensively, resulting in a diverse pool of possible assessment methods for verbal, numerical or spatial abilities, or for fluid and crystallized intelligence, for example. The psychometric quality and validity of many intelligence tests is widely supported by empirical evidence (Roberts, Markham, Matthews, & Zeidner, 2005).

Measurement of candidates for new cognitive ability constructs such as Emotional Intelligence (EI) or Social Intelligence (SI), however, yields new assessment challenges (Roberts, Zeidner, & Matthews, 2001; Schulze, Wilhelm, & Kyllonen, 2007; Zeidner, Matthews, & Roberts, 2001). The number of scientific studies of EI and SI has steeply increased in recent decades, as the topic has attracted the attention of many researchers (Weis, 2008). As a result, much effort has been invested in developing psychometrically sound assessments for various new ability constructs. However, some fundamental issues regarding the measurement of EI and SI remain unsolved, one of which will be addressed in the present paper. The measurement of these new intelligence constructs can be fraught with problems. Therefore, the validity of the constructs is limited by these challenges: The questions of whether the types of intelligences exist and whether an adequate measurement approach is available are not yet fully investigated. The predictive power of these constructs is partly an open research question, which, amongst other things, depends on psychometrically sound measurement.

The fundamental issue for the measurement of new intelligences addressed in the present studies refers to the evaluation of response behavior which is a requirement for maximum performance measures. The evaluation of response behavior with regard to an objective maximum performance criterion is denominated as *scoring* in the present studies. During the literature search, it became apparent that the term is not consistently used in the relevant literature. Alternative terms for scoring are rules, classification, ranking or grading (Guttman & Levy, 1991; Jensen, 1998). These terms refer to the same principle, that is, to the assignment of numerical values to response options with the purpose of quantifying what the option indicates about the underlying trait (Cronbach, 1990). Here, the focus is on scoring for ability measurement: A response indicates higher ability if it is better (i.e., correct), and lower ability if it is worse (i.e., incorrect). On the one hand, response options can be scored either as right or wrong, whereby a value of one indicates a correct option and a value of zero an incorrect option. On the other hand, response options can be ordered with respect to the information they provide about the underlying ability, that is, weights are assigned to options indicating a degree of correctness. Either way of scoring is common for ability testing, see for example Liepmann, Beauducel, Brocke, and Amthauer (2006) for the dichotomous case and Von Aster, Neubauer, and Horn (2006) for scoring in more than two categories. In any case, scoring does not refer to assigning respondents values indicating their ability or applying item response models or other measurement models where the scoring needs to be evident beforehand.

Scoring rules are said to be unequivocal and objective for standard intelligence tests (Guttman & Levy, 1991). Definitions and models of intelligence imply the concept of correctness, that is, intelligent behavior is often - but not solely - defined by the correctness of the behavior. As one of the most central characteristics of intelligence tests, the correctness of item responses is evaluated by scoring rules based on objective standards, for instance based on logical, scientific or semantic criteria (Guttman & Levy, 1991). Objective criteria for correctness have been set as a standard that assessments of new intelligences have to meet (Roberts et al., 2001).

Tests proposed for the measurement of new ability constructs commonly contain response options that have not been evaluated on the basis of theoretical or logical standards. Hence, determining the correctness of response options has been particularly challenging for these instruments. Faced with this problem, the development of psychometrically sound measurements has been cumbersome, progressing only slowly in spite of the attention given to EI and SI (Roberts et al., 2001). Amongst other things, the validity of such tests has been questioned, because rules for judging the correctness of response options are not based on a theoretical and objective rationale (Maul, 2012a). As the scoring of response options is a crucial element of ability measurement, the claims and implications of research into EI and SI are to some degree questionable. Scoring challenges are one of the reasons why some researchers have even claimed that studying these new ability constructs shows no promise (Landy, 2006).

In order to provide scoring rules for response options in tests of new ability constructs, different empirical procedures have been applied, of which one is known as consensus based measurement (CBM) (Legree, 1995; Legree, Psotka, Tremble, & Bourne, 2005). Using CBM, the correctness of a response option is determined by the proportion of respondents endorsing the option. This is based on the idea that respondents themselves provide at least some expertise for abilities which are relevant in daily social interaction. Moreover, Legree et al. (2005) assumed that for tacit knowledge constructs mean judgments of many journeyman do not differ from the mean judgments of a few experts. Knowledge about the correct responses of items with social-emotional content is accumulated in large groups and the consensual responses of these groups may be used in order to identify scoring rules (Mayer, Caruso, & Salovey, 2000; Mayer, Salovey, Caruso, & Sitarenios, 2001). CBM scoring procedures have been criticized because of a vague theoretical rationale as well as lack of empirical evidence for its validity, that is, for the assumption that consensus indicates true correctness of response options (e.g., Maul, 2012a; Roberts et al., 2001). Both points will be addressed in the present paper.

The idea that group consensus provides a valid indicator of true knowledge has found support in both non-psychological and psychological contexts for a long time, providing additional information with regard to a theoretical rationale for consensus scoring. In philosophical literature, consensus theories of truth addressed the validity of consensus among individuals. The consensus of individuals has been discussed as both a criterion for as well as a definition of truth (e.g., Habermas, 1984). However, consensus as a criterion for truth was critically discussed and questioned by several philosophers (Healy, 1987; Ferrara, 1987; Beckermann, 1972). In economic, anthropological and sociological science, the so called wisdom of the crowd (WoC) effect has been used extensively to predict events or to give estimates of physical quantities (Clemen, 1989; Larrick, Mannes, & Soll, 2012). The judgment of a group is hypothesized to be superior to that of a single persons, if the group provides at least some degree of knowledge about the target area and judgments of group members are given independently (Davis-Stober, Budescu, Dana, & Broomell, 2014). Across the areas discussing consensus, knowledge has been consistently argued as a precondition of key importance.

Because scoring rules for the measurement of new ability constructs are not based on theory or logic, consensus scoring procedures cannot be compared to objective standards (Schulze et al., 2007). Thus, it remains a mainly open research question under what conditions consensus might constitute a trustworthy knowledge source for scoring of response options in psychological ability tests. Indeed, empirically based scoring methods have been one of the crucial reason why constructs like EI and SI still remain problematic for psychological measurement (Roberts, Schulze, & MacCann, 2008). In order to develop psychometrically sound tests of all kinds of abilities affected by this problem, it seems necessary to investigate the advantages and disadvantages of empirical scoring methods in a more objective way than has been done in previous studies. Although there have been attempts to study the quality of different empirical scoring methods, these studies have often used rather subjective criteria and been restricted to highly specific research fields (Barchard, Hensley, & Anderson, 2013; Mohoric, Taksic, & Duran, 2010).

In order to improve the scoring of new intelligence tests, alternative empirical scoring methods have been suggested which will also be addressed in the present studies. Empirical and consensus-based methods use the respondents test behavior to estimate an answer key that is hypothesized as existent, yet unknown. The alternative methods have been - if at all - used quite rarely and have never been comprehensively and systematically compared to general objective standards. Investigating the methods, namely (1) Consensus Analysis, (2) HOMALS, and (3) the NRM, aims to advance test development for EI and SI.

(1) Consensus Analysis was developed in anthropological research aiming to identify potentially different scoring keys in various cultures (Batchelder & Romney, 1988). Consensus Analysis intends to estimate the culturally correct answer. It covers a vast array of models and estimation methods for the identification of correct response options. The toolbox of methods has to date never been used for the assessment of intelligence.

(2) The HOMALS algorithm allows quantification of nominal categories within a principal component analysis, as well as outputs quantifications for response categories that might be used as weights for a scoring key (Gifi, 1990).

(3) The NRM - an item response model for categorical data without information about the ordering of response options (Bock, 1972) - has been suggested for use in areas with scoring problems (Guo, Zu, Kyllonen, & Schmitt, 2016). The estimated slope values of the model inform about the degree of the ability that is needed to endorse the option.

The major aim of the present studies is to examine consensus-based scoring procedures from a theoretical as well as an empirical perspective as well as to explore the proposed alternative empirical scoring methods. Using objective criteria, the four above mentioned scoring methods are systematically studied in this thesis. To provide objective criteria, scoring methods are investigated (1) using simulated data, where the simulation model provides the true answer key, and (2) using real world data from ability tests, where the scoring key is provided by unequivocal, theory-based or logical rules. In the simulation studies, several data characteristics are varied to determine the influence of these independent variables on the measures of scoring quality. Important independent variables are hypothesized to be a) number of respondents, b) number of variables, c) difficulty of items, d) ability of respondents, and e) number of response categories. In addition, selected data from Trends in International Mathematics and Science Study (TIMSS) 2011 is used to investigate the scoring methods under more realistic conditions.

The present studies contribute to the test development of various proposed new ability constructs. Major aspects of this contribution concern (1) the discussion of standards for scoring ability tests, (2) the theoretical rationale of empirical and consensus-based scoring methods, (3) most importantly, the empirical examination of the quality of empirical and consensus-based scoring methods, and (4) the resulting consequences for test development for new ability constructs.

Starting from these introductory statements and information, subsequent chapters will provide theoretical background information and describe, analyze and discuss the present studies. Chapter 2 describes measurement approaches suffering from scoring challenges, presents theoretical information about scoring rules in ability measurement, and outlines scoring strategies used for the new-ability constructs. The idea behind consensus based scoring methods in general is then described and referred to related discussions about consensus. Chapter 3 presents the three alternative scoring methods from the methodological perspective. In Chapter 4, the aims and objectives of the simulation and real-world data studies are presented. Here, hypotheses which have been inferred from the theoretical and methodological parts are summarized. Chapter 5 presents the methods and results for simulation studies and Chapter 6 the methods and results for the real-world data study. Chapter 7 discusses the results of simulation and real-world data studies focusing particularly on external validity issues.

2 Scoring Rules in Psychological Measurement of Abilities

Of the many different ways to organize psychological measurement, one very important was introduced by Cronbach (1990), who distinguished between maximum and typical performance tests. Maximum performance tests qualify for the measurement of abilities, in particular intelligence, achievement or aptitude. In maximum performance tests, respondents are explicitly asked to put maximum possible effort into solving the set tasks. The test taker also has to be willing to give his or her best as well as to understand what 'doing best' specifically means (Cronbach, 1990). Typical performance tests qualify to measure personality, including habits or interests. In typical performance tests, typical responses to stimuli are observed by others or reported by the subject of the test. Most commonly, respondents have to record to what extent the statement indicating a particular trait is applicable for themselves.

Scoring of response options of both typical and maximum performance tests has to be systemized by defining uniform scoring rules (Cronbach, 1990). Scoring for typical performance tests does not include setting rules for distinguishing poor from good answers. For instance, showing more gregarious behavior is not in general more correct (effective, good, etc.) than preferring to be alone. Instead, typical responses are most commonly ordered by quantity or magnitude statements using a Likert scale or similar scales. Responses to maximum performance tasks, however, usually can be evaluated by performance criteria. As Cronbach (1990) outlined, defining suitable credits for responses ranging from excellent to poor is difficult.

The example item in Figure 1 serves to illustrate a non-trivial scoring difficulty of a specific intelligence test tasks used for the measurement of new intelligences. Here, the test taker has to understand the situation represented in the left picture, that is, the thoughts, feelings and intentions of the presented person, and has to indicate what most likely follows this situation (O'Sullivan, Guilford, & deMille, 1965). However, the given information about the person as well as the situation is sparse (one picture) and, depending on traits of

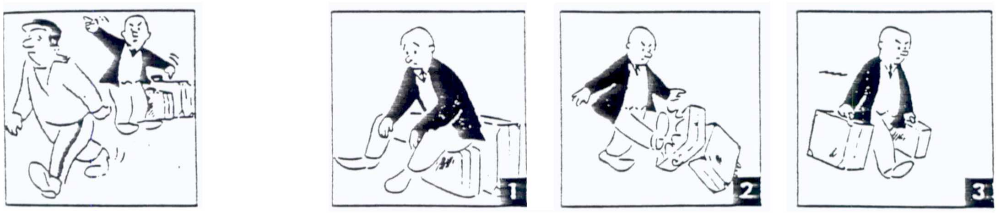


Figure 1. Example item (cartoon prediction) with scoring challenges. From *Measurement of social intelligence*, by M. O'Sullivan, J. P. Guilford, and R. deMille, 1965, Los Angeles, CA: The University of Southern California.

the person as well as the events happening, several likely reactions are plausible. As a consequence, one is able to find plausible explanations for every of the presented three reactions for this example: (1) sadness might follow anger as a characteristic of the presented individual, (2) the individual could also be someone who overreacts and needs to abreact immediately by kicking something, or (3) he could be still angry but leaving the situation or solving it for himself. Nevertheless, no theory easily tells us what is the most correct response option for the example item, leaving a researcher with the challenge of scoring.

The challenge of scoring in ability measurement is different in diverse ability areas. The particular focus of this thesis is on general abilities (intelligence). Hence, performance criteria in intelligence testing will be thoroughly discussed. In particular, scoring rules have been a controversial topic for the measurement of the candidates for the new ability constructs SI and EI (Weis, 2008). These areas will be described in the next section, before the solution strategies for these areas are presented, with a particular focus on one method and on the related criticism.

2.1 Scoring Challenges in Intelligence Research

2.1.1 Social Intelligence. SI has a long but rather less fruitful history in psychological research (Landy, 2006; Weis & Süß, 2005; Weis, 2008). Despite critique of theoretical concepts, as well as limited empirical support regarding the question whether SI is measurable, interest in SI is still intense. In recent years, some promising new measurement approaches, as well as theoretical discussions of the construct, have been published (Seidel, 2007; Weis & Süß, 2005; Weis, 2008). Regarding measurement of SI, scoring of tasks remains a major challenge and a limiting factor for test construction: moreover, it might be the reason for some fundamental psychometric, as well as validity-related problems.

The term *social intelligence* was first mentioned by Dewey (1909) and first stated as an aspect of human intelligence by E. L. Thorndike (1920). Thorndike mentioned three types of intelligence, abstract, mechanical and social intelligence, in a popular scientific article, and defined SI as "the ability to manage and understand men and women, boys and girls, to act wisely in human relations" (p. 228). Although E. L. Thorndike is frequently cited as the first to propose a definition of SI, he did not himself conduct research on this topic. No discussion of the construct took place in the following years, nor was any detailed theoretical definition elaborated, although several tests of SI were developed (for an overview see Landy, 2006). These tests were often based on rather vague definitions of SI, which confused cognitive, behavioral, and knowledge structures (Seidel, 2007). Presented many years after the first tests of SI, two theoretical examinations are in particular noteworthy: behavioral cognition as a part of the Structure Of Intellect (SOI) model (O'Sullivan et al., 1965), and the cognitive performance model of SI (Weis & Süß, 2005). As the term *intelligence* implies cognitive operations (Seidel, 2007), both theories focus on cognitive, rather than on behavioral or knowledge aspects of SI. Other cognitively based theories might be worth mentioning, such as that of Gardner (1983) who introduced the concept of interpersonal intelligence. The two probably most important theoretical models of SI are introduced briefly in the following to provide information about the theoretical status of the construct and the fundament for test construction.

In 1965, O'Sullivan, Guilford and De Mille defined SI within the SOI model of

Guilford (1956, 1981). At that time, their theoretical discussion was a resurrection of the construct, as previous attempts to measure SI had not been very successful, and a theoretical discussion of the construct had so far been neglected. O'Sullivan et al. (1965) hypothesized "difficulties in determining the "right answer"" to be one of the reasons for the paucity of research on the topic in previous decades. The authors hypothesized that SI is a multidimensional construct including a group of intellectual abilities. As established in a re-analysis of inter-correlations among intelligence test tasks, they introduced behavioral intelligence as part of the SOI, which includes 30 factor of behavioral (social) intelligence. The model resulted in, amongst others, the six factor test of behavioral cognition (O'Sullivan & Guilford, 1975). In order to provide a more comprehensive theoretical model of SI, Weis and Süß (2005) presented a synthesis of the many different definitions based on an extensive literature review. As Weis (2008) stated, the many different measures and theories of SI can be categorized into the operation classes of reasoning, memory, perception, creativity and knowledge. Most fall in the first category, also referred to simply as social understanding. In the Cognitive Performance Model of Social Intelligence (Weis & Süß, 2005), SI is defined as a multifaceted construct with the sub-abilities of (1) social understanding, (2) social memory, (3) social perception, (4) social creativity and (5) social knowledge. This model constitutes the basis for the Magdeburg Test of Social Intelligence (MTSI) (Seidel, 2007; Weis, 2008).

In the course of the long history of SI, different measurement approaches have been used, including cognitive ability tests, self-report inventories, self-report of implicit theories of laypersons, or behavioral tests (Weis, 2008). Without a solid theoretical framework, it long was difficult to judge the adequacy of different measurement approaches; they can, however, be evaluated on the basis of the theoretical models of SI described above. On this basis, only cognitive ability tests allow measurement of SI. In contrast, self-report measures, as well as behavior-based ratings are inappropriate for the measurement of intelligence. Self-report measures are appropriate to assess typical behavior and have been used, for example, in Ford and Tisak (1983), Marlowe (1986), Keating (1978) and Riggio (1986). Behavior-based ratings aim to assess social behavior (Ford & Tisak, 1983; Frederikson, Carlson, & Ward, 1984; Stricker & Rock, 1990). The focus of the present section is, however, on cognitive ability tests of SI: these are maximum performance tests which require performance-related scoring rules.

Until 1983, six of the early tests of SI can be classified as ability tests, as their responses range from correct to incorrect, or scoring was based on other maximum performance criteria (Landy, 2006). According to Landy, these tests were all scored by consensus, that is, the response which was seen as correct by the majority of a student or expert sample was classified as correct. The most important early tests as well as one recent test of SI will now be described briefly, to illustrate the related scoring issues as well as the approaches of dealing with these issues - as far as they are known. The description of SI tests is by no means claimed to be complete, but serves to illustrate common testing approaches as well as the related scoring problems. Seidel (2007) and Weis (2008) provided a more thorough list of assessment tools.

The George Washington Test of Social Intelligence (Hunt, 1928; Moss, Hunt, Omwake, & Woodward, 1955) was the first test of SI and for a long period of time the only available assessment tool. The test has - depending on its version - several subtests, namely (1) judgment of social situations, that is, ability to provide solutions for social problems, (2) memory for names and faces, (3) recognition of mental states from facial expressions, (4) observation of human behavior or behavioral knowledge (5) social information, and (6) recognition of mental states behind words, that is, emotion implied in quotations or speech (Landy, 2006). Table 1 presents sample items for some subtests taken from Moss et al. (1955). According to Landy (2006), the correctness of the response options is defined by some kind of consensus. However, very little explicit information or critical discussion of the scoring procedure is given in the relevant publications, with one exception: Strang (1932) investigated the errors made by respondents of the George

Table 1

Sample Items from the George Washington Test of Social Intelligence

Subtest	Item	Response options
Judgment of Social Situations	Assume that you are a teacher of a third grade and while going to school after the first snow of winter some of your pupils throw snowballs at you. From the standpoint of good school management you should:	(A) Punish them then and there for not treating you with the proper respect. (B) Tell them that if they ever do it again you will punish them. (C) Report them to their parents. (D) *Take it as a joke and say nothing about it.*
Observation of human behavior	A person of strong character usually makes firm friends and bitter enemies.	(A) *True*, (B) False
Recognition of mental states behind words	Which one of them shall I take? Both? One? Or neither?	(A) Disappointment, (B) Hypocrisy, (C) *Indecision*, (D) Love

Note. Correct response options are presented in italics. Adapted from *George Washington University Series Social Intelligence Test* (2nd ed.), by F. A. Moss, T. Hunt, K. T. Omwake, and L. G. Woodward, 1955, Washington, DC: Center for Psychological Service. Copyright 1949 by Center for Psychological Service.

Washington Test. Strang's results indicate that the correctness of responses is questionable with regard to the inferences they allow about SI. For instance, one of the tasks requires respondents to indicate appropriate behavior when speaking with an acquaintance who has just lost a close relative. 31% of respondents incorrectly chose to speak well about the departed relative, instead of the correct response to talk about current events of general interest. As Strang (1932) pointed out, the correctness of a behavior depends on the particular relation of the two people speaking, which is not described in sufficient detail in the task. The inspection of the response options in Table 1 supports the conclusions made by Strang (1932): The correctness of the response options might depend on missing information, for instance, about the particular school management rules or about the nature of a strong character. Not only due to scoring issues, but also due to the many psychometric problems of the test, and in particular because of the high overlap with abstract verbal intelligence (Broom, 1928; Strang, 1930; R. L. Thorndike, 1936), the George Washington Test is only historically significant.

The Chapin Social Insight Test (Chapin, 1942, 1960; Gough, 1965) was developed to measure social insight, defined as the ability to recognize the two mechanisms "psychological dynamics underlying a particular behavior" (Gough, 1965, p. 356) and "the stimulus, compromise, or innovation necessary to resolve the situation or to carry it through to a constructive conclusion" (Gough, 1965, p. 356). Chapin (1942) developed 45 (20 after item analyses) items from material like case histories of individuals, novels, existing scales, conference discussions, and so forth. Table 2 presents a sample item from the test (Chapin, 1960). Problem situations were briefly described, for which several reactions might be possible but only one was chosen as correct. Chapin (1942) stated that the items were "self-validating", that is, the correct response options were validated by a "competent observer of the problem as the most likely explanation, interpretation, or comment" (p. 216). Although not referred to it in the original literature on the tests, the scoring key is hypothesized to be based on consensus (Landy, 2006; Weis, 2008). The presented example item again demonstrates the difficulties which are aligned to scoring; for the example item, depending on the specific history of the woman, different explanations might be correct.

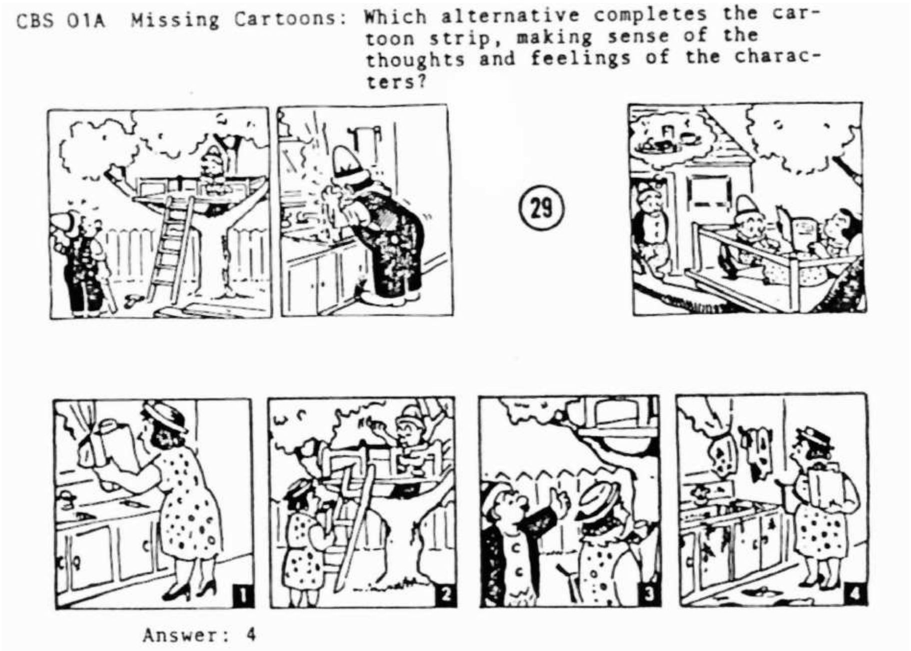
Table 2

Sample Item from the Chapin Social Insight Test

Ability	Item	Response options
Social insight	A young woman reacted with intense emotion to any indulgence in alcoholic drinks. If any of her friends as much as took a single drink, she went out of her way to denounce them in most emphatic terms. The explanation was:	(A) That her mother had been a leader in the Mother's Against Drunk Driving. (B) That her father had been an alcoholic, who had treated her mother brutally and finally deserted her. (C) She was herself a secret drunkard at late parties. (D) Her ancestors came from a strict religious background that denies the use of alcohol.

Note. No information about the correct response options was available. Adapted from *Chapin Social Insight Test*, by F. Stuart Chapin, 1960, Redwood City, CA: Mind Garden. Copyright 1960, 1963 by Consulting Psychologists Press, Inc.

The Six Factor Test of Social Intelligence (O'Sullivan & Guilford, 1975) measures the six factors of behavioral cognition as stated in the SOI model, defined as the "ability to understand other people's thoughts, feelings, and intentions" (p. 256). As cross classified in the SOI model, behavioral cognition includes the product classes units, classes, relations, system, transformation, and implication (O'Sullivan et al., 1965). Test materials for the six factors were not based on verbal stimuli, but likewise included material based on pictures of faces and bodies, cartoons, and recorded words or sounds. Figure 1 presented an example item for which the scoring problematic was illustrated in the beginning of this chapter. Figure 2 and 3 present two additional item examples with similar scoring challenges. In Figure 2, a short cartoon story is presented and the missing element of the story should be identified. In the other task, a photographed story is presented for which the last part ought to be selected out of three options. The information given in the stories is limited and different developments of the presented stories might be plausible. Scoring of tests can hypothesized to be mainly done by consensus scoring, as the items were pre-tested and classified as good "in terms of consensual validation of the key and discrimination between high- and low-scorers on the total test" (O'Sullivan et al., 1965, p. 6). For some subtests, however, scoring might be based on target or expert scoring as well, or as O'Sullivan et al. (1965) stated as "rational" (p. 17).

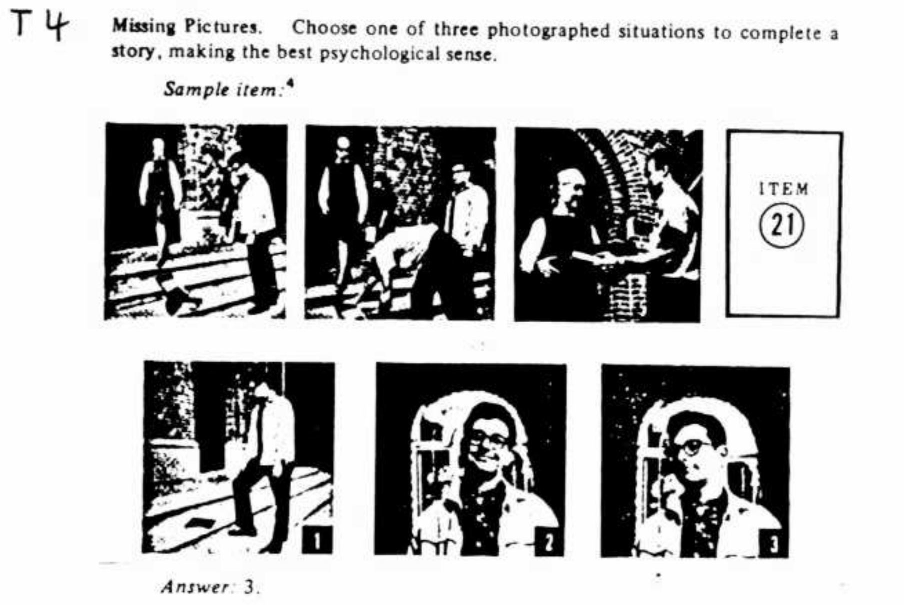


CBS 01A Missing Cartoons: Which alternative completes the cartoon strip, making sense of the thoughts and feelings of the characters?

Answer: 4

Figure 2. Example item (Missing Cartoons) from the Six Factor Test of Social Intelligence. From *Measurement of social intelligence*, by M. O'Sullivan, J. P. Guilford, and R. deMille, 1965, Los Angeles, CA: The University of Southern California.

Based on the SI model of Weis and Süß (2005) described above, Süß and colleagues



T 4 **Missing Pictures.** Choose one of three photographed situations to complete a story, making the best psychological sense.

Sample item:[4]

Answer: 3.

Figure 3. Example item (Missing Pictures) from the Six Factor Test of Social Intelligence. From *Measurement of social intelligence*, by M. O'Sullivan, J. P. Guilford, and R. deMille, 1965, Los Angeles, CA: The University of Southern California.

developed an MTSI ability test (Baumgarten, Süß, & Weis, 2015; Conzelmann, Weis, & Süß, 2013; Seidel, 2007; Weis, 2008; Weis & Süß, 2007). The test measures three ability areas of the cognitive performance model of SI, namely social understanding, social memory and social perception. The test development for two ability areas, social creativity and social knowledge, was omitted due to severe challenges in test construction: both sub-abilities would require a far more highly elaborated theory to allow performance-based scoring (Weis, 2008). In social memory tasks, respondents are confronted with text, picture or video stimuli material including socially relevant details which they have to memorize. Six memory tasks are included and relevant details are correspondence/conversation, voices, couples, or situations. Correct responses to social memory tasks are provided on the analogy to classical memory tasks, that is, the correct recognition of presented social stimuli defines a correct response. In social perception tasks, respondents have to quickly identify and classify social target cues, for instance a person's body, face, or voice, or other social cues like eye contact, specific emotions, and so forth (Conzelmann et al., 2013). The MTSI contains eight social perception tasks, namely perception in texts, language, voices, pictures and videos. For social perception tasks, the reaction time for correct responses determines performance in respective tasks. Both for social memory and perception tasks, scoring does not constitute a key challenge, as for both type of tasks the performance is not mainly defined based on the theory-based correctness (but on memory and speed) of test behavior. For social understanding, however, responses are categorized regarding their correctness and scoring challenges are more severe. Social understanding is measured with eight scenarios, where respondents are required to judge target persons regarding emotions, cognition or relationships (Conzelmann et al., 2013). The tasks are presented on the basis of written or spoken language, pictures or videos. The target person defined the correct response option for the tasks. For the test construction, the targets were accompanied for several days. They had to indicate their cognitions and emotions and had to give information about their relationships. Moreover, the test constructors evaluated the correct responses with regard to their own knowledge about the target person as well as with regard to consistency between their evaluations. The authors pointed out that the psychometric quality of the social understanding tasks has to be further improved, as some items had negative or zero item-total correlations. As one reason for these problems, Conzelmann et al. (2013) mentioned possible negative effects of target scoring. Targets might not be capable of correctly identifying and labeling their own cognitions and emotions, or they might be biased by social desirability. However, alternative scoring procedures are not favored, as target scoring allows the most objective scoring (Conzelmann et al., 2013), above all when cross-validated by expert scoring. Although some scoring issues remain unsolved for the measurement of social understanding, in comparison to the older SI tests, the MTSI has adequately addressed scoring challenges for its measurement, however, without being free from effects of these challenges.

To briefly summarize, SI is a construct that continues to inspire many researchers as well as laypeople; however, severe theoretical as well as measurement-related challenges have not yet been overcome. The example items illustrate the difficulties of defining the correctness of response options, which is why various different scoring procedures have been used for SI measurement, the possibly most important being empirical methods such as scoring based on target information, expert judgment or group consensus. For some tests scoring procedures were simultaneously used; for older tests, it is moreover difficult to clearly identify scoring procedures used, as these issues were not as transparent as they are nowadays. In fact, issues of scoring are only considered with required depth in the newest studies. Scoring issues still are a key problem for SI measurement, which need to be further addressed.

2.1.2 Emotional intelligence. In comparison to SI, models and measurement approaches of EI do not have a very long history; however, the last two decades of research on EI have been more intensely engaged with scoring challenges. Here, one ability model has dominated the research: EI has most frequently been conceptualized as a system of four abilities (Mayer & Salovey, 1997), that is, as the ability to correctly perceive, use, understand and manage emotions in oneself and others. Like SI, the measurement of EI requires maximum performance tests, since EI tests are intended to meet the standards for intelligence (Mayer et al., 2001; Roberts et al., 2001). Among other criteria they should have a solid scoring system based on veridical standards (Roberts et al., 2001; Schulze et al., 2007). Scoring is a prevailing root problem of EI measurement (Barchard et al., 2013; MacCann, Roberts, Matthews, & Zeidner, 2004; Mayer et al., 2000; Mohoric et al., 2010). Accordingly, research is characterized by the search for and discussion of ways to determine the correctness of response options, possibly even more than it is discussed for SI measurement.

EI has been a hot topic in scientific communities and in popular science over the past 27 years. Salovey and Mayer (1990) first theoretically framed the term. In their article, they argued that EI is the ability of subjects to "understand and express their own emotions, recognize emotions in others, regulate affect and use moods and emotions to motivate adaptive behavior" (p. 16). The concept of EI was based on research in various areas and incorporated the view that emotions might function in an adaptive way to enhance cognition. From early on, EI was defined as a component of intelligence, specifically as a potential sub-construct of social or personal intelligence (Mayer & Salovey, 1993; Mayer & Geher, 1996; Salovey & Mayer, 1990). In the first years of research on EI, two routes evolved and appeared to diverge quickly (Mayer, Roberts, & Barsade, 2008). With the popular bestseller by Goleman (1995), EI promised to become a widespread and fruitful research area that attracted researchers and lay-people alike. Some researchers focused on traits, motivations, or attitudes that were commonly assessed via self-reports (Mayer et al., 2008). This line of research has been named trait EI (Petrides & Furnham, 2000, 2001). Without convincing evidence, it was claimed that trait EI was an even better predictor of key outcome variables than intelligence and personality. However, studies showed that trait EI might not be capable of being differentiated from the Big Five personality factors, especially neuroticism and extraversion (MacCann, Matthews, Zeidner, & Roberts, 2003; MacCann, Schulze, Matthews, Zeidner, & Roberts, 2008; Matthews, Zeidner, & Roberts, 2005). Trait EI should be investigated separately from ability models because different criteria and standards apply for validation (Kaufman & Kaufman, 2001; MacCann et al., 2008). Since identifying correct response options is usually of little or no interest for personality measurement, trait models and measurements will not be further discussed here.

During the late 1990s, Mayer and Salovey (1997) refined the ability model, which is now well known as the four branch-model of EI. They defined EI as the "ability to perceive accurately, appraise, and express emotion; the ability to access and/or generate feelings when they facilitate thought; the ability to understand emotion and emotional knowledge; and the ability to regulate emotions to promote emotional and intellectual growth" (p. 10). Accordingly, the ability model of EI contains four sub abilities, (1) perception of emotions (for instance in faces) as well as expressions of emotions, (2) facilitating emotional conditions to enhance thought, (3) understanding emotions and emotion transition, and (4) regulating emotions in oneself and others. Although Mayer, Salovey, and Caruso (2002) provided evidence that supported the intended structure of MSCEIT, independent confirmatory tests of the model cast doubts whether the second branch is replicable. Empirical results support a three-factor structure with the factors perception, understanding and management, or a two factor structure with the factors experiential and strategic EI, where strategic EI is represented by the understanding and management branches, whereas experiential EI is only represented by the perception branch (H. Fan, Jackson, Yang, Tang, & Zhang, 2010; Palmer, Gignac, Manocha, & Stough, 2005; Rossen, Kranzler, & Algina, 2008).

The intense research of the recent years has produced a few ability EI tests that theoretically more or less incorporate the four-branch model (Mayer et al., 2008). The model also covers the theoretical basis of emotion-perception tests, which were developed partly independently of EI research (Roberts et al., 2006). Despite years of profound and prolific research, the first theoretically based, structured, and published ability test has remained the predominantly used and known test of the past decade. Test variety still remains limited, and important shortcomings in assessments have not been fully overcome. The most popular ability EI tests are the Mayer-Salovey-Caruso Emotional-Intelligence-Test (MSCEIT) (Mayer et al., 2002; Mayer, Salovey, Caruso, & Sitarenios, 2003) and its precursor the Multidimensional Emotional Intelligence Scale (MEIS) (Mayer et al., 2000). In recent years, alternative assessments have been developed, for instance the Situational Test of Emotional Understanding (STEU) and the Situational Test of Emotion Management (STEM) (MacCann & Roberts, 2008). The description and presentation of ability EI measurement does not claim to be exhaustive, but serves to illustrate the measurement status for EI with reference to the most important tools. For more general information about EI tests see for example Schulze and Roberts (2005) or Mayer et al.

(2008); for specific ability based tests see Blickle, Momm, Liu, Witzki, and Steinmayr (2011), Hellwig (2016), or Warwick, Nettelbeck, and Ward (2010).

Table 3

Fictive Sample Items for the Mayer-Salovey-Caruso Emotional-Intelligence-Test

Subtest	Item	Response options
Emotion facilitating	Your are excited, but nervous because of an important meeting at work. How much does this feeling match the following sensations?	(A) hot, (B) bright, (C) sour.
Emotion Understanding	Unreturned sympathy is most likely followed by	(A) disappointment, (B) depression, (C) grief, (D) shame, or (E) surprise
Emotion Management	Julia has been searching for a parking spot in center city for 30 minutes. As she just found one, someone cut in on her and takes the spot. Julia is in a rage. How effective are the following reactions?	(1) Scream and curse the thief, (2) Catch a breath and continue searching, (3) Park in front of the car that just had stolen the spot, (4) Get off the car and confront the car driver

Note. No information about the correct response option is available. Item are written by the author of this dissertation based on test material from *Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT): User's manual*, by J. D. Mayer, P. Salovey, and D. R. Caruso, 2002, Toronto, Canada: Multi-Health Systems.

The MSCEIT is an ability-based multi-faceted measurement tool for EI developed by Mayer et al. (2002) based on the results of studies on the precursor MEIS. Both the MEIS and MSCEIT are directly based on the four branch model of EI. The MEIS showed several shortcoming regarding reliability at subtest level, confirmation of the four branch model, and length (Mayer et al., 2002). Due to these shortcomings, the authors constructed the shortened MSCEIT which seeks to assess the four branch model with two subtests for each branch. In the emotion perception subtests, respondents view pictures of emotions expressed in faces or in landscapes and designs and have to rate the extent to which different emotions are present. Emotion facilitating is assessed via two subtests, where (1) respondents are requested to link a described emotion to sensations such as warm, cold, sweet, and so forth, and (2) respondents are confronted with a situation that requires cognitive work; they are asked to indicate on an answer scale from not-helpful to helpful how supportive the presented emotions are. An example item (rate-the-extent) for this branch is presented in Table 3. Understanding emotions is assessed via tasks, where (1) respondents indicate which combinations of emotions are similar to particular emotions, and (2) respondents have to choose among five answer alternatives which emotion most likely follows to a presented situation and emotion (see Table 3 for an example). In managing emotions items respondents rate reactions to situations according to their effectiveness with regard to emotion management. In the emotional relationship tasks, the respondent is again confronted with a situation where the emotions of other persons should be managed effectively, and respondents are requested to rate possible reactions according to their effectiveness. An example item for emotion management (rate-the-extent) is presented in Table 3. The four-branch model of EI does not provide detailed information about the emotion perception, facilitating, understanding, and mangagement, so that response options of the MSCEIT items are not evaluated based on theoretical assumptions. In contrast, the correct responses of MSCEIT (as well as MEIS) items are defined by two different empirical criteria: the responses of 21 (two for MEIS) experts (expert scoring), or of the sample (consensus scoring), which have been shown to correlate highly for MSCEIT (Mayer et al., 2002, 2003; Palmer et al., 2005). According to the authors of the MSCEIT, the empirical scoring methods are appropriate for the measurement of EI; however, the scoring procedures of the MSCEIT has been object of major criticism, as will be presented in section 2.2.

Scoring challenges were discussed more elaborately and more critically for the development of the STEU and the STEM by MacCann and Roberts (2008). The tests intend to measure two branches of the four branch model, namely understanding and managing emotions. Item stems of both STEU and STEM constitute short descriptions of situations. In STEU, respondents are requested to indicate the emotion that arises from the described situation (for an example item see Table 4). The situations and resulting emotions are inferred theoretically from Roseman's Appraisal Theory (Roseman, 2001). Hence, STEU can be scored theoretically, but also has been scored based on expert and group consensus judgment. In STEM, respondents are asked to rate several reactions with regard to their potential for effectively regulating the emotion in the situation (for an example item see Table 4). This test is based on the Situational Judgement Test (SJT) construction approach (Legree, 1995). SJTs represent a measurement methodology of judgment in specific situations (McDaniel & Nguyen, 2001; Weekley & Ployhart, 2006), which has most frequently been used in I/O psychology. SJTs can be used to measure specific ability or ability-related constructs (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; Christian, Edwards, & Bradley, 2010; Motowidlo, Hooper, & Jackson, 2005; Schmitt & Chan, 2006). The STEM is scored via expert scoring (MacCann & Roberts, 2008), that is, experts rate the effectiveness of each response option and their ratings, collected both in rate-the-extent and multiple choice formats, serve as criteria for response evaluation. However, the STEM can also be scored on the basis of group consensus for which the groups responses serve as criteria for response evaluation.

Table 4

Sample Items for the Situational Test of Emotional Understanding and the Situational Test of Emotion Management

Subtest	Item	Response options
Emotional Understanding	Clara receives a gift. Clara ist most likely to feel?	(A) *Happy*, (B) Angry, (C) Frightened, (D) Bored, (E) Hungry
Emotion Management	Jill's aunt Daria complains incessantly about her ill-health and impending death. Although Jill and her family thought Daria was a hypochondriac, it turns out she has luekemia. What could Jill do to deal with her guilt feelings?	(A) Talk to friends or family for support, (B) *Gather information about what will happen*, (C) Do other activities to distract himself, (D) Calming relaxation techniques, (E) Think positive

Note. Correct response options are presented in italics. Items are instruction or not used items from *New approaches to measuring emotional intelligence: Exploring methodological issues with two new assessment tools*, by C. MacCann, 2006, Sydney, Australia: University of Sydney.

In the quest to provide clear criteria for the assessment of ability EI, common characteristics and features of intelligence tests provide some orientation towards the formation of standards. Several researchers have claimed that as a part of the construct intelligence EI must meet these standards (Mayer et al., 2000; Orchard et al., 2009; Schulze et al., 2007). These standards include (1) a pattern of correlations that provides evidence for convergent, discriminant, and predictive validity for EI, that is, it should be moderately correlated to other forms of intelligence, be independent of personality, and show predictive validity for relevant criteria while controlling for personality and intelligence, (2) the development of EI over time, that is, an increase in ability similar to crystallized intelligence, and (3) unequivocal rules for scoring as they are common in intelligence testing. The first criterion is widely accepted and has been both empirically confirmed and criticized for ability tests (Austin, 2010; Brackett & Mayer, 2003; Farrelly & Austin, 2007; Matthews, Roberts, & Zeidner, 2004; Matthews, Zeidner, & Roberts, 2012; Maul, 2012a; Palmer et al., 2005; Roberts et al., 2008; Van Rooy, Viswesvaran, & Pluta, 2005). Ability EI tests have been shown to correlate moderately with intelligence, particularly crystallized intelligence, and to be largely independent of personality. In addition, MSCEIT and MEIS scores have been shown to predict real-life criteria (Roberts et al., 2008). However, definition of important real-life criteria for emotion processing remains a huge challenge. The second criterion has been criticized, because a certain developmental trend with age depends on the intelligence type (Roberts et al., 2001). Although empirical results support a stronger relationship with crystallized intelligence (Farrelly & Austin, 2007), the definition of EI does not clearly state that it has a high overlap with crystallized intelligence. In fact, the definition also includes reasoning with information in the context of emotion (Mayer & Geher, 1996); however, the development criterion is not common for reasoning ability. Along with this criticism, the biggest shortcoming of most tests is possibly the problem of scoring (Schulze et al., 2007). Indeed, the psychometric quality of EI measurement cannot be evaluated without addressing scoring challenges.

Mayer et al. (2001) outlined that knowledge about EI has not been systematized to the same degree as it has for other intelligences; agreement regarding emotional rules is therefore not yet available. Although the four branch model is the most widely known and used EI model, it is not well elaborated with regard to concepts of correctness, that would allow one to infer a scoring algorithm. According to Mayer et al. (2001), with growing research on EI, the systematization of knowledge will continue and EI will reach the standards that are common in intelligence research. Standards-based scoring - that is, unequivocal and objective rules based on logical or theoretical criteria - is prevalent for intelligence testing and should also be implemented in EI testing (Roberts et al., 2008).

Consequently, generic theories of EI that are able to provide veridical rules are essential for valid assessments of intelligence. For the sub-abilities of perception, understanding, and management, empirically supported theoretical standards are available that provide details on how to evaluate certain responses or alternative measurement approaches have been developed which provide standards by other means (Hellwig, 2016; MacCann, 2006; Mohoric et al., 2010; MacCann & Roberts, 2008; Roberts et al., 2008); however, using these standards to evaluate response options does not necessarily result in reliable assessments (MacCann, 2006).

Since most theories of EI do not provide means for veridically scoring tests with unequivocal rules, different ways of identifying scoring rules have been suggested. Very similar to SI, rules for scoring of EI tests have commonly been identified using empirical methods, such as group consensus scoring, expert consensus scoring or target scoring (Mayer, DiPaolo, & Salovey, 1990; Mayer & Geher, 1996). As the assessment of EI has mainly depended on the use of empirical scoring procedures, research on EI has addressed the use of these scoring methods. Before addressing these scoring procedures in more detail, the role of scoring in intelligence measurement is discussed, as measurement of new ability constructs is requested to be on a par with measurement of classic intelligences.

2.2 The Role of (Consensus) Scoring in Ability Measurement

As has been mentioned in the last sections, the measurement of EI and SI often relies on scoring methods that base on human judgement or consensus. Following the general criticism, the measurement of these new ability constructs does not incorporate scoring rules that are unequivocal and objective. From the perspective of measurement of new ability constructs, MacCann et al. (2008) stated "in standard intelligence tests, the correctness of an answer can be obtained from trusted sources of knowledge (e.g., arithmetic tests are scored according to mathematical rules, vocabulary items with thesauri or dictionaries)" (p. 141). The correct response of ability items is said to be defined on the basis of logical, semantic, empirical, or normative rules (Zeidner et al., 2001). For *new* ability constructs, the scoring rules do not appear to meet these standards, as their basis in trusted knowledge sources is questionable. In the following subsections, the scoring rationale for *standard* or *classical* intelligence tests are presented and discussed in more detail. In order to build a basis for discussing scoring challenges in ability measurement, two questions are investigated in the following subsections: (1) What standards should scoring meet from a philosophical perspective? and (2) In what way is scoring theorized in literature about theories and measurement of general ability?

2.2.1 Scoring in Intelligence Measurement from a Philosophical Perspective. The concept of psychological measurement is shaped by philosophy, which provides different perspectives - realist as well as non-realist - on measurement and how measurement and the construct to be measured are related (Borsboom, 2009). Although non-realist views still influence the understanding of measurement today, state-of-the-art measurement in psychology relies predominantly on a realist perspective (Borsboom, 2009). This is nowadays a widely accepted scientific mindset. The realist perspectives propose that a natural world exists independently of claims about it, and that knowledge of this world can be acquired through to scientific methods (Devitt, 2006; Maul, Wilson, & Irribarra, 2013). Hence, theoretical assumptions are said to be *true* if they correspond to facts, and measurement is an attempt to capture parts of the natural system or theory.

The two following realist conceptions are described to present the realist view on scoring for intelligence measurement. On the one hand, the realist view is incorporated in the framework of evidence-centered assessment design (ECD), which outlined the necessity of theory-based scoring for the construction of psychological tests in large-scale assessment situations (Mislevy, Almond, & Lukas, 2003; Mislevy & Riconscente, 2005). As a part of the ECD, the conceptual assessment framework (CAF) integrates necessary steps in the test development process. One essential element of CAF is the evidence model, which includes the evaluation of responses in light of psychological theory (Schulze et al., 2007).

Theories specifying constructs ought to indicate how behavior in terms of performance is to be evaluated, that is, what kind of behavior indicates a specific value (or area) of the trait. Observations that allow inferences about an underlying ability should be clearly designated in terms of scoring rules.

In line with the realist view on measurement, moreover, Borsboom, Mellenbergh, and van Heerden (2004) presented a definition of validity that focuses on the scoring of responses. According to these authors, a test provides validity if (1) the measured trait exists and (2) it causes differences in the observed behavior. This definition of validity focuses on the ontological status of the trait (i.e. the question whether the trait exists) rather than on epistemological questions of validity of inferences based on the measurement. Accordingly, the focus of the validity concept is the theory of the construct being measured. The theory should allow its user to connect certain (test) behaviors and values of the latent trait. If variations in the latent trait can, on the basis of a solid theory of the construct, be shown to be causal for the variation in the observed behavior, then the measurement is said to be valid.

These realist views on measurement and scoring result in certain assumptions for state-of-the-art psychological measurement. Psychological traits are assumed to exist independently of their measurement, as not directly observed attributes of individuals. Theories merely motivate hypotheses about the structure and content of their measurement. These assumptions include predictions about relevant behavior, as formalized in scoring rules. Realist measurement requires scoring rules which (1) are defined on the basis of theoretical descriptions of the construct, and (2) are supported by empirical evidence.

2.2.2 Scoring and Intelligence Measurement. Approaches of intelligence measurement include concepts of performance in terms of correctness, success, speed and originality. The correctness of a response is the most central element of performance in intelligence test tasks; however, performance is not solely defined by correctness for all types of tasks. For some tasks, performance is mainly defined by speed of producing correct responses or originality of correct responses. Here, the focus is on the tasks in which the evaluation of responses is based on correctness.

In the literature mentioning evaluation standards of intelligence test tasks, intelligence is frequently defined in terms of performance in tasks. A clear separation of theory and measurement is not always evident. One reason for this might be that definitions of intelligence are heterogeneous and little consensus has yet been reached on this issue (Jensen, 1998; Sternberg, 1986). Instead of specifying the nature of the construct itself, it appears more common to describe the (test) behavior that intelligence is hypothesized to cause. While seeking to avoid confusing theory and measurement of intelligence, the present subsection will refer to researchers who have more or less confounded these two areas. Their approaches will be described from the measurement perspective, assuming that the performance evaluation of behavior is a central element of theories of intelligence, albeit without claiming to define intelligence from a theoretical perspective.

While some approaches to intelligence measurement only implicitly state evaluation of performance as a definitional aspect, other approaches discuss this aspect of intelligence measurement more explicitly. Examples for the first case can be found in Carroll (1993), Sternberg (1978), Thurstone (1938), and Wechsler (1944). For instance, according to Carroll (1993), different kinds of abilities share a common definition in terms of performance. Carroll defined ability as "*possible variations over individuals in the liminal levels of task difficulty (or in the derived measurements based on such liminal levels) at which, on any given occasion in which all conditions appear favorable, individuals perform successfully on the defined class of tasks*" (p. 8). As a determinant of successfully performing an ability task, Carroll assigned the correct or appropriate processing of mental information a major role, without, however, explicitly stating what defined successful performance.

Only few intelligence researchers, that is, E. L. Thorndike, Bregman, Cobb, and Woodyard (1926), Guttman and Levy (1991), Nevo (1993), and Jensen (1998), have mentioned the types of scoring rules more explicitly. The types of rules as defined by these researchers will be described in the following.

As one of the early intelligence researchers, E. L. Thorndike commented on performance evaluation standards for intelligence measurement. In the symposium on intelligence "The nature and measurement of intelligence", E. L. Thorndike defined intelligence in his article as "the power of good responses from the point of view of truth or fact" (Intelligence & its measurement: A Symposium, 1921). In "The measurement of intelligence", E. L. Thorndike et al. (1926) outlined the measurement of intelligence in terms of the evaluation of intellectual products and stated: "A is rated as more intelligent than B because he produces a better product" (p. 11). The evaluation of this product is for some tasks "so clear that everyone must admit it" (p. 13), for others it is based on human judgment, for example, for identifying the opposites of words, completing sentences or answering questions about a text paragraph. According to E. L. Thorndike et al. (1926), the sources of evaluation are truthfulness or wisdom, grammatical form, rhetorical excellence, originality, rate of production or "a subtle sense of their significance as evidence of intelligence" (p. 13). For more than half of the discussed intelligence tests, truth was said to be the evaluation standard and the authors concluded:

> It may be that truth-getting is what we unwittingly do measure by our intelligence tests, or what we ought to try to measure, but very few of those who devise or apply the tests think so. And it is surely wise to find out what we do measure before deciding that it is or ought to be truth-getting. E. L. Thorndike et al. (1926, p. 16)

Seemingly, these valuation standards were not of great interest throughout the following decades, until Guttman and Levy (1991) specified general characteristics and scoring rules of ability items that build upon a definitional framework for intelligence test items within the Facet Theory of Intelligence (Guttman & Levy, 1991; Süß & Beauducel, 2005). Facet Theory is a meta theory of intelligence that allows one to transparently formalize a theoretical system, including its hypotheses, in terms of a categorization of observations (facets) (Süß & Beauducel, 2005). Facets of intelligence are those characteristics in which items can differ systematically from each other although they still show a positive manifold, that is, they show positive correlations. The different facets of intelligence are stated in the mapping sentence, where, for example, content information of items (e.g., verbal, numerical, figural) or information about their mode of presentation (e.g., paper-pencil, verbally) is given. According to Guttman and Levy (1991), an item is specified as an ability test item only if it represents a specific domain for which an objective rule can be stated that allows possible response options to be ordered from *very right* to *very wrong.* The range of scoring constitutes a basic characteristic of ability items, rather than a possible varying facet of those items. The scoring range appears to be the common ground of all traditional ability items (Guttman & Levy, 1991; Süß & Beauducel, 2005). According to Guttman and Levy (1991), scoring rules for items can be logical, scientific or semantic, the first type of rule being the most unequivocal and the last type the most ambiguous. These cases have in common that "an outside body of experts decide what the (current) right answer is" (Guttman & Levy, 1991, p. 8). Subjective rules such as aesthetic, religious or cultural norms are not regarded as appropriate for intelligence testing, although studying these subjective rules might be worthwhile within other psychological theories.

Based on Guttman's earlier work, Nevo (1993) attempted to define the nature of evaluation rules in more detail. Accordingly, considerations about such rules in terms of correctness can be seen as closely related to a very fundamental field of philosophy: theories of truth and the question how truth (i.e. the true correct answer) can be defined. In a small study, he investigated the implicit theories of 22 students about scoring rules. These implicit theories of correctness largely corresponded to philosophical criteria for truth. Hence Nevo introduced a typology of scoring rules, without, however, concluding that this typology might be exhaustive. According to the typology, correctness can be defined by (1) correspondence, that is, relations to the physical-empirical world, (2) coherence, that is, formal relations based on mathematical and logical systems, (3) authority, that is, knowledge of an expert or an authority, and (4) semantic congruity, that is, closeness in meaning.

The fourth intelligence researcher who explicitly referred to evaluation standards is Jensen (1998) who, given the different definitions of these standards, rejected a consensual definition of intelligence altogether and focused instead on aspects of its measurement. Jensen (1998) argued that, to infer to an ability, item performances - voluntary behavioral acts discretely classifiable or ratable on a continuum (hence excluding unconscious or automatic processes) - have to meet three requirements. An ability is defined by item performances if these performances are (1) stable across time, (2) can be ranked or scored by a "standard of proficiency" (Jensen, 1998, p. 51) and (3) generalize across different, but similar items. Regarding the second criterion, Jensen (1998) stated: "An essential element in the definition of ability is that the IP [item performance] must be an act that can be objectively classified, ranked, graded, or scored in terms of an objective standard of proficiency" (p. 51). If an item performance cannot be scored according to an objective rule, it does not qualify as an ability item. Defining the characteristics of abilities in terms of criteria of item performances, Jensen went on to outline examples for objective standards of proficiency, such as comparison of reaction times, calculus, or a certain demonstration of physical strength for a fixed time frame. Of course, physical strength is not seen as a mental ability, but served as an example for objective evaluation standards. In contrast, personal, moral, social, or economic standards did not qualify as appropriate rules. In addition, Jensen (1998) stated:

> Ratings or rankings with a specified degree of agreement among several judges (as quantified by the intraclass correlation or the coefficient of concordance) can also qualify as an objective standard and may serve to rate IPs [item performances] that do not lend themselves to direct measurement, such as performance in figure skating,

> playing a musical instrument, singing, art work, influencing people, and the like (Jensen, 1998 p. 51)

To summarize, these performance criteria are organized as shown in Figure 4. Ability is measured by (maximum) performance, which can be evaluated by different standards, namely correctness, speed, and originality as well as possible other standards. Correctness, which according to Guttman and Levy (1991) is the most common evaluation standard for intelligence test items, can be defined on the basis of different types of rules presented in the right hand column in Figure 4.

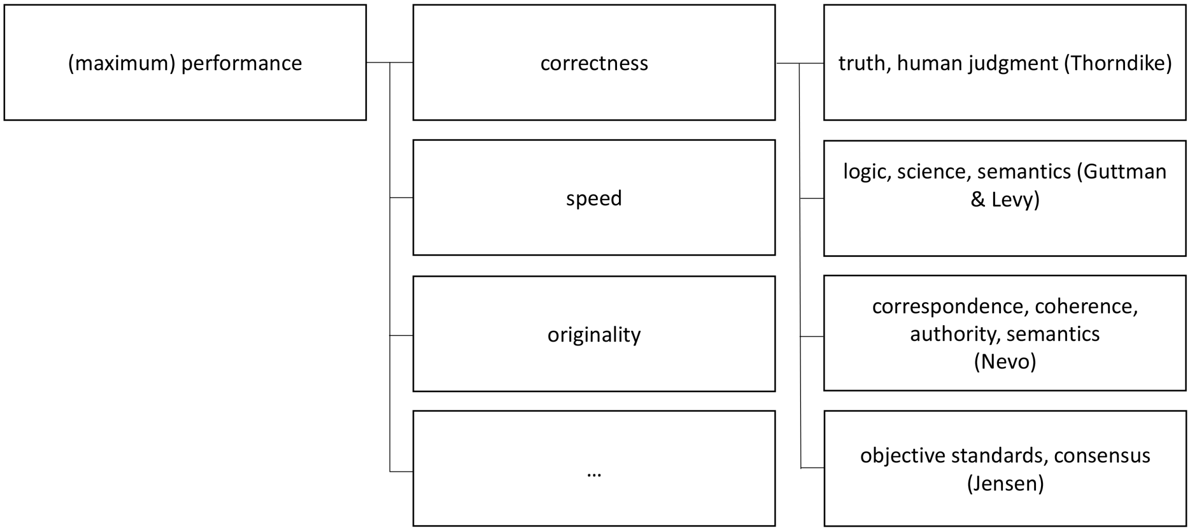

(maximum) performance
correctness
speed
originality
...
truth, human judgment (Thorndike)
logic, science, semantics (Guttman & Levy)
correspondence, coherence, authority, semantics (Nevo)
objective standards, consensus (Jensen)

Figure 4. Schematic illustration of performance criteria for ability measurement.

Three essential points can be summarized based on this literature review and will be further addressed in the following:

1. The number of references discussing scoring rules is low compared to the total number of articles and books on intelligence; hence, scoring seems to be at no point controversial for *classic* intelligence tests. However, for some intelligence tasks measuring *standard* intelligence, scoring is not necessarily unequivocal.

2. For intelligence measurement the realist measurement approach is only partly

realized with respect to scoring rules. Intelligence theories state that higher latent ability is related to higher task performance. For some intelligence theories and their measurement this link is quite clear: For instance, higher speed-related abilities relate to higher speed in respective tasks. For other theories and their measurement, standards for evaluation of performance (e.g., correctness) - and hence definitions of specific categories of test behavior - are more imprecisely grounded on theories of intelligence. That is, it remains unclear for some tasks, why a specific response option is the correct one. As E. L. Thorndike et al. (1926) and Nevo (1993) stated, philosophical theories of truth might contribute to a concept of correctness, however, only if performance is evaluated in terms of correctness (not speed or originality), which is itself not defined on the basis of logic.

3. Although differences between the typologies remain - which might be due to lack of discussions - , high overlap is observed in Figure 4 for one objective evaluation standard: human judgment, consensus, authority, and the "outside body of experts" (Guttman & Levy, 1991, p. 8). While Guttman and Levy (1991) and Nevo (1993) explicitly referred to experts as a source of evaluation, E. L. Thorndike et al. (1926) more vaguely referred to human judgment, and so did Jensen (1998), with the additional requirement that judges need to agree. Hence, human judgment appears to be a common source for objective evaluation.

To address the first concluding point in more detail, scoring of inductive reasoning tasks is discussed in the following. Number series tasks - a typical inductive reasoning task - were mentioned by a anonymous reviewer of the article by Roberts et al. (2001) as an example, where different rules might justify different correct responses. Hence, these tasks entail a potential non-trivial scoring challenge in *classical* intelligence measurement. The ability to reason is part of all major intelligence theories and its measurement is a common component in most intelligence tests (Wilhelm, 2005). Reasoning as an aspect of fluid intelligence includes inductive as well as deductive reasoning processes which, however, cannot be psychometrically differentiated (Colberg, Nester, & Trattner, 1985; Shye, 1988; Wilhelm, 2000). The definition of inductive and deductive reasoning is based on the cognitive psychology concept of mental models (Johnson-Laird, 1994b, 1994a, 2001). Mental models are cognitive representations of the process of understanding a situation in which certain premises are stated. Deductive conclusions are defined as valid conclusions which "must be true given that their premises are true" (Johnson-Laird, 1994b, p. 10). Deductive conclusions preserve the semantic content of the premises, which justify the conclusion if premises are true and the conclusion is rational. Thus, items that mainly require deductive reasoning should have a logically correct and unequivocal scoring key. Inductive conclusions, however, are not necessarily true given their premises. Johnson-Laird (1994b) defines induction as *"any process of thought yielding a conclusion that increases the semantic information in its initial observations or premises"* (p. 11). They increase the semantic information of the premises in such a way that content is added to the premises without necessarily following from them; but such content must be plausible. Hence inductive conclusions may be true if their premises are true, but the premises do not allow final evaluation of their truth. Items that mainly require inductive reasoning processes will predictably entail scoring difficulties, because inductive conclusions cannot be evaluated by any absolute criterion for correctness (Wilhelm, 2005).

Examples of inductive tasks, which are used for the measurement of individual intelligence differences, are verbal analogies, number series, or figural matrices. Figure 5 represents an example item[1] for which a conceptual scoring challenge of inductive reasoning tasks is illustrated. The item represents a typical matrix task consisting of three elements and a fourth empty element. The elements of the matrix are arranged by specific rules which define the correct missing element. The test taker has to infer these rules and has to choose the correct missing element from five response alternatives. The following rules (mental model), which ought to define the correct response, can be inferred from the item

[1] The example item is adapted from the IST-2000-R intelligence test (Liepmann et al., 2006)

stem material: The circles are separated into three uniform parts, of which the bottom left segment is always empty. The right bottom and top segment are each filled with a pattern. Given these observations, it allows to infer that the correct circle should have (1) an empty left bottom segment, (2) a right bottom segment containing a pattern, and (3) a top segment containing a pattern. The semantic information of the mental model may be increased in different ways, two of which will be described in the following. One plausible way of increasing the semantic information is to concentrate on the pattern that appeared in the top segment in the left circles and switched to the right bottom segment in the right top circle. Focusing on this pattern, response option b) is a plausible correct response similar to the correct response in the original test material. Another plausible way of increasing the content is to take the occurrence of all pattern into account. In total, four different patterns appeared and the same pattern never appeared in the same circle twice. So far, two of the four patterns appeared two times, whereas the other two patterns appeared only once. The patterns that appeared two times were never in the same segment twice, that is, they were either in the top or on the right bottom segment of the circle. Hence, Figure 1c is another plausible correct response. Other ways of increasing the semantic content makes other responses plausible, which results in difficulties defining one unequivocal correct response.

The illustration shows that there is not only one correct response option for inductive tasks, resulting in challenges regarding response evaluation in these tasks. Although not only one specific inductive conclusion can by definition be classified as correct, and the number of possible conclusions increases with the complexity of the model, the challenges are only rarely discussed (Mittring & Rost, 2008; Preckel, 2003), and there has been no published general discussion of scoring challenges associated with inductive reasoning tasks. A reason for this lack of discussion may be because inductive reasoning tasks do not show any psychometrically conspicuous features (Wilhelm, 2000). During test construction,

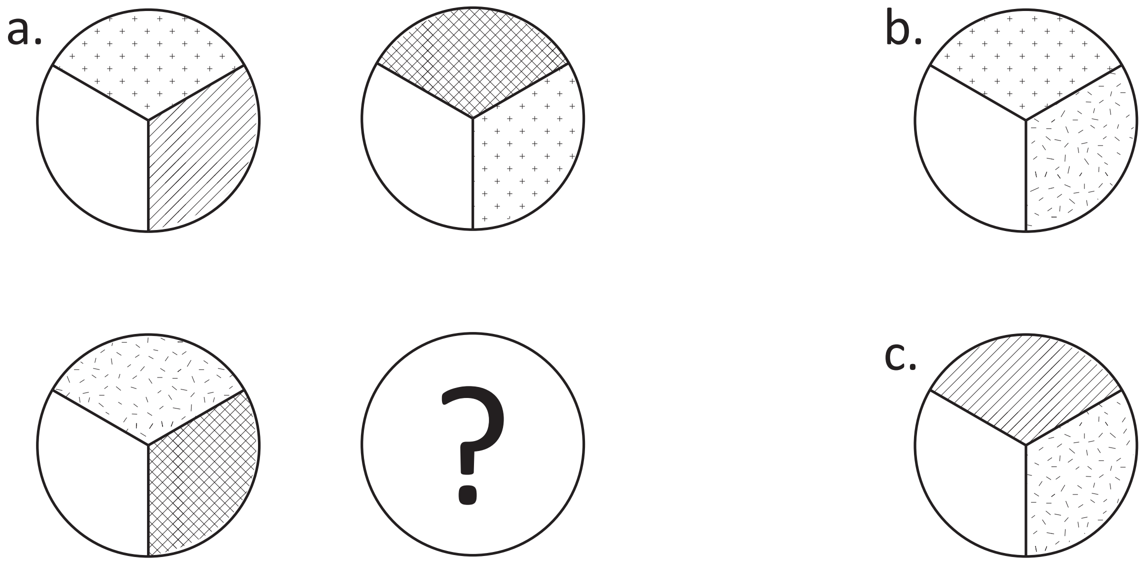


Figure 5. Example item with scoring challenges. Adapted from *Intelligenz-Struktur-Test 2000 R* (2nd ed.), by D. Liepmann, A. Beauducel, B. Brocke, and R. Amthauer, 2006, Göttingen, Germany: Hogrefe.

items with two equally likely and plausible answer alternatives are supposedly eliminated on the basis of item statistics, such as item-total correlations, which is a procedure recommended for test construction (Lienert & Raatz, 1998). It is likely, therefore, that items with only one plausible response option remain in the tests, or items with distractors or other information sources that provide additional premises, making these tasks deductive rather than inductive (Mittring & Rost, 2008; White & Zammarelli, 1981; Wilhelm, 2000, 2005). When complex inductive reasoning tasks are used without distractors scoring might become more challenging, as more than one or partly correct responses are possible. However, providing rules or other relevant information for solving the tasks - for examples see Becker, Preckel, Karbach, Raffel, and Spinath (2014) providing the elements of the correct response or Wilhelm (2005) for an number series item with additional information about the possible operations - scoring challenges are also prevented by producing a deductive character. This illustration reveals that the concept of correctness is theoretically challenging for inductive reasoning tasks; scoring is not necessarily unequivocal.

The second concluding point presented above addressed the question to what degree typical sources for scoring rules for intelligence measurement are grounded on realist scientific standards. Defining the correctness of response options - as one possible evaluation criteria for intelligence measurement - has been associated with philosophical discussions about truth (Nevo, 1993; E. L. Thorndike et al., 1926). This link also gets evident in the requirement that scoring rules should be *veridical*, or in stating *science* as a source for the scoring rules. However, truth has only mentioned as an aspect of the definition of intelligence by E. L. Thorndike et al. (1926), but has not been discussed in a more elaborated way providing a theoretical link between the ability and the test behavior. Philosophical truth theories can also be evaluated on the basis of fundamental philosophical perspectives, whereas the realist perspective is the one measurement of intelligence should be aligned with. Definitional (realist) theories about truth address the meaning of the term and give semantic-logical definitions of truth, whereas criteria for truth (non-realist theories) provide means to test if a proposition is true (Rescher, 2012). Realist theories are defined by two conditions: (1) an individually necessary and jointly sufficient condition for the truth of a statement is that the fact expressed by the statement must exist in the natural or extramental world, and (2) that the fact in the extramental world is mind-independent, that is, it exists independent of human beings thinking about it (Kirkham, 1999). Theories that do not fulfill these conditions are non-realist. In particular, non-realist theories do not hold that facts occur independently of our mind, hence it is, in such theories, neither a necessary nor a sufficient condition for a statement to be true that actual facts occur.

Nevo (1993) did not differentiate between realist and non-realist truth theories, however, presenting both types of theories: Coherence theory - which is a non-realist theory - and correspondence theory (Kirkham, 1999) and the semantic theory of truth Tarski (1935) - which are realist theories. Authority, that is, consensus theory of truth (Habermas, 1984) was discussed to be both realist and non-realist and will be addressed later in section 2.3.3.1. The attempt of Nevo (1993) to equalize truth theories and types of scoring rules should be seen rather critically, as it vastly simplified the philosophical discussion of truth and neglected several important theoretical consequences. For one thing, logic is a method of philosophy that is used in coherence theory, but it is not identical with coherence (Kaufmann, 1940; Kirkham, 1999). Coherence theory defines a statement as true if it is an element of a set of statements or beliefs that entail each other and as a whole give a complete picture of the world (Kirkham, 1999). Coherence of beliefs or statements is seen as evidence for their truth. In addition, the semantic theory of truth does not formalize truth in natural languages (Tarski, 1935), as Nevo's typology would suggest. Semantic theory as described by Tarski (1935) rather states that truth can be defined in formal language (Field, 1972). To define truth, one needs to assume two types of formal languages, (1) the object language, that is, the language in which the statement is discussed, and (2) the metalanguage, which is used to give the definition of truth. The metalanguage needs to copy the object language in order to describe the truth of statements in the object language. A sentence in the object language is said to be true if there is a sentence in the metalanguage that copies it.

Due to the mentioned critical points, the scoring rules typology based on truth theories is not appropriate for a realist measurement approach, but however constitutes a useful starting point for discussing sources of scoring rules from a theoretical perspective. Aiming for a realist measurement approach, the correctness (truth) of response options should be aligned with realist truth theories. Therefore, a response option is correct if the fact stated in the option does exist and its existence is independent of individual human beings.

The third concluding point, which is the most central here, will be addressed in more detail in the following subsection where the most common solution strategies for scoring of tests of *new* intelligences are presented: empirical scoring methods based on human judgment.

2.3 Solution Strategies to Scoring Challenges: Empirical Scoring

Scoring based on theoretical rule-systems is a gold-standard for psychological measurement. But comprehensively elaborated theories of EI and SI are not available, so although attempts have been made to base scoring of measurement of new ability constructs on theory, standards-based scoring has not been thoroughly established for the measurement of these constructs.

For the measurement of SI and EI, several different empirical scoring procedures are common. Empirical scoring procedures identify correct responses not on the basis of theory but of information from one or more respondents. Hence, the conclusions about the underlying ability, inferred from the selection of single response options, are not based on theory but on the judgment of a single individual or a group of individuals. Judgment of response options is gathered during, or even after, test construction. Popular empirical methods used to score EI and SI tests as well as SJTs, are target, expert consensus, and group consensus scoring (Bechtoldt, 2008).

Target scoring, as one of the empirical scoring procedures, is only introduced very briefly here as it will not be further investigated. Target scoring uses the information of a target individual to score responses to stimuli that were provided by that individual (Mayer & Geher, 1996; Weis, 2008). For instance, the target person reports the emotion she or he feels while taking a picture which is used as an item stimulus in an emotion identification task. Target scoring is suitable whenever human thoughts or feelings are part of the item stimulus. A fundamental idea of target scoring is that the target has the best knowledge about her or his mental state. Target scoring is not suitable for higher level aspects of EI, such as reasoning with emotions (Roberts et al., 2001), and might be problematic in respect to the targets ability of introspection (MacCann et al., 2004). On the supposition that targets possess expertise in describing their own status, target scoring represents a variant of expert scoring (Mayer & Geher, 1996). However, targets can only be regarded as experts if they are capable of reflecting on their own emotional reaction, recognizing it, and labeling it correctly. The focus here, however, is on expert and group consensus scoring, which will be described in the following.

Consensus scoring represents an evaluation standard of response options which is based on the judgment of a group of respondents. In the field of EI, consensus scoring is often referred to as consensus based measurement (CBM). As Legree et al. (2005) state, CBM constitutes a maximum performance scoring technique for the assessment of knowledge-based constructs. For tacit knowledge areas, expert opinion has been used to define the correctness of the response options (Legree, 1995). Using consensus of a particularly highly able or educated sample as a criterion against which responses are evaluated is referred to as expert consensus scoring (Legree, 1995; Mayer & Geher, 1996). Although expert consensus scoring may be suitable in many areas of psychological measurement where experts can be defined easily, the identification of genuine experts is said to be problematic in the field of social and emotional abilities (Legree et al., 2005). Legree (1995) and Legree, Martin, and Psotka (2000) suggested to extend the idea of expert consensus scoring and to use the opinion of a wider group for scoring of EI or SI tests. Using group consensus scoring, the response is evaluated against the most frequent response (for multiple-choice items) or the mean (for rating-scale) of the group as a criterion.

There are many detailed CBM methods (Legree et al., 2005); the most popular and psychometrically promising in the field of EI are proportion CBM and mode CBM (MacCann et al., 2004). CBM methods using percentage agreement or endorsement are suitable for either multiple Choice or Likert type (e.g., rate the extent) items. Mode CBM states that the modal response of a representative sample is the correct answer. Proportion CBM uses the relative frequency with which respondents endorse each response option as weights indicating correctness. For instance, five response options of an item are endorsed with the following relative frequencies: a) .05, b) .10, c) .40, d) .20 and e) .25. Using proportion CBM, respondents who endorsed answer option a) would earn a credit of .05

and respondents who endorsed option c) would earn a credit of .40. In the case of mode CBM, respondents endorsing option c) would earn a credit of one, while respondents endorsing other response options would earn no credit. Comparing several consensus scoring methods using MSCEIT Emotion Perception subtests Faces and Designs, MacCann et al. (2004) concluded that only proportion and mode CBM scores showed convergent validity. Other CBM scoring methods are lenient mode, distance, squared, or adjusted distance or correlation methods (Legree et al., 2005). Lenient mode consensus scoring might be usefully applied in the case of ordered response options (e.g., rate the extent format). Here the most frequently endorsed answer option and the response options next to that option earn credits. In addition, for Likert type items, several distance CBM methods are available (Legree et al., 2005; MacCann et al., 2004). Distance CBM methods use distances between individual ratings and mean ratings of the whole sample as a score for each item, whereby smaller distances represent higher consensus with the sample. These simple distances can be adapted in different other distance CBM methods, such as standardized distances (for more information see Legree et al. (2005)). Correlation CBM scoring represents another method that uses the correlation of individual scores on each item and mean scores across the sample as scoring weights (Legree et al., 2005). According to MacCann et al. (2004), lenient mode, distance and adjusted distance failed to result in one-dimensional scores.

As consensus scoring has been mainly used and discussed in the field of EI, the following discussion of empirical results refers to studies in this field. Consensus scoring has also been used for the measurement of SI, but criticism has not been as extensively focused on this procedure. In recent EI research consensus scoring has received both support and criticism. The following subsections outline the opposing positions.

2.3.1 Supporting Consensus Scoring in EI Research. The use of consensus scoring was motivated by (1) considerations about the construct, (2) assumptions about knowledge development, (3) empirical findings, and (4) comparisons with standard intelligence tests.

From the theoretical viewpoint, the nature of the construct in consideration was said to give reasons for using consensus to define correctness of reactions. In the early days of EI research, Mayer et al. (1990) defined group consensus as "the ability to perceive emotions that were consensually viewed as present, and the equally weighted ability to consensually agree when emotion was not present" (p.776). Thus it was not assumed that consensus scoring would provide the objective rule required for ability testing, but that the consensus itself constituted an ability (i.e., EI). Asking a representative group of people to define the emotion present in a subject's face, to define the most effective strategy for regulating an emotion, and so forth, and utilizing the modal answer as the scoring rule, was justified by the assumption that social groups define correct emotional reactions. Individuals were said to learn correct emotional behavior by looking at the consensus of the group (Mayer et al., 2001). Moreover, even experts acquire their knowledge by observing the consensual information provided by a group (Legree, 1995; Mayer et al., 2000). Because correct social emotional behavior was theorized as learned in group processes, these groups were also said to be qualified to judge emotional stimuli and reactions.

However, as research proceeded, reasons for consensus scoring were discussed in greater detail and lines of argumentation changed in their underlying assumption. A different theoretical perspective was presented by Legree et al. (2005), focusing on the general meaning of group consensus. These authors were the first to present a potential rationale for group consensus scoring according to which consensus scoring can be used for maximum performance measurement of tacit knowledge or of areas where knowledge is gained through experience. These knowledge areas are insufficiently formalized in terms of knowledge sources (i.e., experts), but the respective knowledge is important for everyday interaction so that everyone is accumulating it with growing experience. Thus, most individuals have some knowledge about these social interactions, which in turn can be revealed by the test behavior of these individuals. With increasing experience, the knowledge develops in terms of mean tendency as well as variance, and agreement among individuals increases. As agreement is a function of knowledge, experts are said to differ less in their responses, whereas knowledge of journeyman is hypothesized as varying more widely. However, according to Legree et al. (2005), the variety of experience of many journeyman is expected to exceed the knowledge of a few experts. Regarding their mean tendency in this respect, journeyman and experts are hypothesized as converging. Thus the judgment of journeyman is said to reliably reveal knowledge structures.

Empirical evidence was provided by Mayer et al. (2002) and Mayer et al. (2003) showing that MSCEIT test scores based on expert consensus and group consensus scoring correlate very highly ($r = .96$ to $r = .98$) and that MSCEIT scoring keys correlated $r = .88$ to $r = .91$. Other studies have indicated that experts agree more readily than novices (Mayer et al., 2003; MacCann, 2006). However, MEIS was criticized because of not high enough correlations ($r = .48$) between expert and consensus scores (Roberts et al., 2001). In the case of MEIS, the two test authors served as experts, whereas for MSCEIT a total of 21 experts provided judgments. Supported by the high correlation of expert and consensus scoring, Mayer et al. (2002) recommended using consensus scoring. However, in a later article, Mayer et al. (2003) stated experts as more reliable judges because of their higher agreement. Legree et al. (2005) concluded that group consensus scoring is a valid alternative to expert scoring. These ambiguous views will be readdressed in the next paragraph.

As Mayer et al. (2001) and Mayer, Salovey, and Caruso (2012) argue, the correct responses of most existing intelligence tests represent group or expert consensus. Hence, Mayer and colleagues conclude that expert and group consensus scoring is in general adequate for EI measurement. Specifically, they refer to subtests of the WAIS-III for which scoring was discussed in an expert panel. Here, several correct responses were theoretically possible for subtests of vocabulary, similarities, information and comprehension and scoring rules were not provided from objective knowledge sources. Although Mayer and colleagues did point to a probable possibility - as has been shown in section 2.2.2 - however, referring to what has probably been done in the past is not necessarily a valid argument for the use of consensus scoring.

2.3.2 Critical Investigation of Consensus Scoring in EI Research. While group and expert consensus scoring have been advocated by some researchers, others view these methods as a fundamental blemish of ability EI (and SI) tests and as one of the reasons why only a few ability tests have been developed (Brody, 2004; Matthews et al., 2005; Maul, 2012b; Roberts et al., 2008). There has been distinct doubt as to whether group consensus scoring provides reliable information about the correctness of response options for intelligence testing (Davies, Stankov, & Roberts, 1998). As Roberts et al. (2001) observe, "in the worst case, consensus scoring may simply indicate the extent of agreement with cultural or gender-based prejudices" (p. 203). Consensus may merely reflect the ability of an organism to adapt to the environment, and hence merely indicate greater adaption. This person-environment fit is both interesting and relevant, but does not necessarily indicate intelligence (Roberts et al., 2001). Maul (2012b) questioned whether emotion is consensually defined and stated that "nothing inherent to the logic of consensus-based scoring of emotional stimuli indicates either that consensus *determines* the correctness of answers, or that consensus will reliably *discover* correct answers" (p. 397). Based on the used empirical scoring procedures, Maul (2012b) questioned different aspects of the validity of the MSCEIT. He raised concerns regarding whether the scoring of different responses reflect different levels of EI and whether test content and theory are sufficiently interrelated. Zeidner et al. (2001) argue that to constitute an intelligence, EI measurement requires objective scoring rules, however they question whether these rules are generally definable. According to these authors, knowledge about emotions can depend on cultural influences and is not necessarily veridical. Hence, serious doubt has been raised concerning (1) whether veridical, objective scoring rules exist for EI, and (2) whether consensus scoring can provide such rules.

The first question depends upon theoretical considerations of the construct, and the question whether socio-emotional abilities are measurable. In the context of ability testing standards, veridical scoring is a definitional criterion of ability measurement; hence assuming that this criterion cannot be met, that is, that veridical scoring rules do not exist, one might question whether a measurement approach to these abilities does in fact exist. The assumption is maintained here that veridical scoring rules do exist for the measurement of social-cognitive abilities.

The second question, however, is of major importance for the present studies. To answer the question whether group or expert consensus scoring are capable of establishing an objective scoring key, consensus scoring needs to be compared to objective standards (Schulze et al., 2007). To date, two studies have compared consensus scoring methods to veridical scoring methods.

Attempting to investigate this research question, Mohoric et al. (2010) compared different versions of group consensus scoring ($N = 197$) and expert consensus scoring ($N = 10$) with a veridical scoring key for a specific emotion understanding test. They used the Vocabulary Emotion Test (VET) where respondents were asked to choose from six options the emotion word that was closest in meaning to a presented emotion word. The presented emotion words describe emotional states and moods. The 102 items of the VET have shown to be internally consistent, with a Cronbach's $\alpha = .91$, and to have unique variance independent of intelligence (Mohoric et al., 2010). According to the authors, the Croatian dictionary provided veridical, objective scoring for the items. The veridical scoring key was compared to group and expert consensus scoring, both in the proportion and mode version. For proportion CBM, Cronbach's alpha was rather low ($\alpha = .53$), whereas for the other scoring methods alphas were greater than .80. The experts consulted for expert scoring fully agreed with the veridical key in their modal response. Hence, the sum scores based on mode expert consensus correlated $r = 1.00$ with scores based on the veridical key, whereas the correlation was lower for proportion expert scoring ($r = .88$).

Sum scores based on veridical scoring and on group consensus scoring were correlated $r = .88$ for mode consensus scoring and $r = .52$ for proportion consensus scoring. Group proportion and mode consensus scoring did not fully converge, as their correlation was $r = .69$. To summarize, group consensus only partly succeeded in identifying the correct response options, whereas expert consensus showed better results. However, even here, experts did not agree for every item. The mean agreement of experts was, however, higher compared to novices ($r = .82$ versus $r = .32$). These results provided evidence for consensus scoring to be less reliable compared to expert scoring. Moreover, mode consensus scoring should be preferred according to the authors. Nevertheless, the authors pointed out some shortcomings, such as outstanding validation of the VET and characteristics of the consensus and expert scoring sample.

Barchard et al. (2013) investigated whether group and expert consensus scoring converge with a veridical scoring key for easy, medium, and difficult items, as well as for different number of items. They scored the 60 item Las Vegas Vocabulary Test, for which the veridical scoring key was again provided by the dictionary, with three versions of consensus scoring: (1) regular proportion consensus scoring, (2) two step proportion consensus scoring, where the top 10% of the respondents of proportion consensus scored data provided the relative weights for scoring, and (3) expert consensus scoring, where the top 10% of the respondents of the veridical scored data provided the relative weights for scoring. A total of $N = 353$ individuals were used for scoring. The results for the different scoring methods were investigated on the level of item keys and on the level of scored data. For proportion scoring, number of correctly identified items, that is, for which consensus scoring assigned the highest weight to the correct response option, varied with difficulty. For difficult items, both group consensus methods did not assign the highest score to the correct option in more than 50% of items. Expert consensus scoring worked better, as it assigned the highest value to the correct option in 75% of items. For easy and medium difficult items, results were more promising, as 90% to 100% of the correct response options were identified for these items by all methods. On the level of scored data, these results were confirmed: correlation of sum scores based on the veridical key and regular group proportion consensus scoring were $r = .53 - .66$ for difficult items, $r = .85 - 1.00$ for medium items and $r = 1.00$ for easy items. The value of the correlations also varied with number of items; it was higher with increasing number of items. For two step group consensus scoring and expert scoring, the correlations were higher and also varied with number of items. The authors concluded that proportion consensus scoring has not been successful for difficult items and should not be used for individual testing with difficult items; however, for easy items it worked appropriately. Although the authors do not claim that their vocabulary test is equivalent to EI tests, they, however, do not see any reason for limitations in generalizing the results to the measurement of EI.

In addition to the question of validity of consensus scores, psychometric problems have been outlined for consensus scoring. Assuming that response behavior can be modeled on the basis of common test-theoretical standards, consensus scoring will not allow the identification of difficult items (Zeidner et al., 2001; Zeidner, Roberts, & Matthews, 2004). The most correct response option is always the option that was chosen by the majority. This is why the most correct response option is automatically the easiest one, which results in a strange relationship between item variance and item difficulty. As item variance and item difficulty are positively correlated, easy items provide more information for the total score than difficult items, in particular for proportion CBM (MacCann et al., 2004). This effect was similar, albeit weaker, when using mode consensus scores. Consequently, the distribution of test scores of consensually scored tests is usually highly (left) skewed and has high kurtosis. As MacCann et al. (2004) showed, consensus scores cannot be both normally distributed and reliable. Highly skewed scores (e.g., proportion) show adequate reliability, while more normally distributed scores (e.g., mode) show weak reliability. Highly left-skewed test scores, however, do not allow discrimination of high ability levels. For instance, consensually scored tests might not allow determination of very fine gradings in emotion perception ability where highly able people can be hypothesized to perceive slightly different expressions than can less able respondents (MacCann et al., 2008; Roberts et al., 2008; Zeidner et al., 2001). Consensually scored tests are not, therefore, suitable for situations in which individual decisions are made based on test results, especially for highly able individuals which cannot be reliably identified based on consensually scored tests.

Barchard and Russell (2006) illustrated how mode CBM scoring can result in bias for smaller subgroups of the scoring sample. If the smaller subgroup differed from the larger group in their modal response, the total test score of the subgroup was automatically smaller than the test score of the major group. In comparison, proportion CBM was not as strongly biased as mode CBM, because even those answers that were not chosen by the majority group received a nonzero score.

Consensus scored scales have also been criticized for low reliability (Ciarrochi, Chan, & Caputi, 2000; Davies et al., 1998; MacCann et al., 2004). In order to improve reliability, MacCann et al. (2004) suggested using Multidimensional Reciprocal Averaging (MRA), a method allocating weights to response options that maximize the homogeneity of the scale. MRA succeeded in slightly improving the reliability as well as in improving the distribution of consensus scores. In addition, the relationship between item difficulty and item variance was corrected with MRA. However, difference between consensus scores and MRA consensus scores were not large. Keele and Bell (2009) investigated if optimal scaling could increase reliability estimates of consensus scores of MSCEIT sub scales. Optimal scaling is a procedure very similar to MRA, where nominal answer categories are metrically quantified within a principal component analysis with a specified number of components. Like MacCann and colleagues, Keele and Bell (2009) used consensus sores as starting values in this algorithm. They showed that Cronbach's alpha increased using optimal scaling, however the increase was significant only for one of the two sub scales. Results indicated that optimal scaling scores differed most from consensus scores for items where consensus scores were somewhat equally distributed and no response option was clearly preferred by group consensus. Overall, correlations between optimal scaling scores and consensus scores were high, indicating that scores based on optimal scaling and group consensus were very similar.

Due to doubts regarding the validity of consensus scoring, as well as the psychometric issues described above, the status of consensus scoring has been questioned. Moreover, the theoretical rationale behind consensus as a scoring criterion has not been discussed with the required breath. A more thorough elaboration of relevant theories is needed, which is the major aim of the following subsection.

2.3.3 Extending the Rationale of Consensus Scoring. Consensus scoring, that is, human judgment, authority, or consensus as a source for scoring rules, was mentioned early in the intelligence literature and has been used for the measurement of general intelligence: Jensen (1998) and Thurstone (1938) assigned group consensus the role of an objective standard of proficiency for the measurement of abilities, and Nevo (1993) affirmed expert knowledge as a source of scoring rules for intelligence tests.

In research on EI and SI, however, it has been questioned whether consensus provided objective and veridical scoring rules (Roberts et al., 2001; Schulze et al., 2007; Zeidner et al., 2001). Aside from the pending question of the empirical quality of consensus scoring as described in subsection 2.3.2, the theoretical rationale of this method has not yet been sufficiently elaborated. This is an important issue. As indicated in subsection 2.3.1, arguments supporting the use of group consensus scoring initially referred to the role of consensus within the construct definition. The definitions in question included the assumption that emotionally and socially intelligent behavior is defined by what the group thinks is correct (Mayer et al., 2001). Later, discussion of the usage of consensus scoring concentrated on common standards of ability definition and measurement. Measurement of new ability constructs should meet common standards of intelligence tests (Roberts et al., 2001). To meet these standards, scoring rules ought to be based on trusted knowledge sources.

The question whether consensual judgment is a criterion of true knowledge, is a matter of discussion not only in EI and SI research: it is a general philosophical question, as well as a practical issue in politics, economics and the social sciences. In exploring possible theoretical rationales for consensus scoring, several important fields might be mentioned. On the one hand, consensus has been discussed extensively in philosophical literature, starting with the ancient Greek philosophers and continuing with present-day philosophers like Habermas. The philosophical consensus theory of truth assumes that under certain circumstances consensus might work as an indicator of truth (Habermas, 1984). On the other hand, consensus has been used as a predictor of future events, as an estimator of quantities, and as a source of information for solving complex problems in social science and economics. The WoC effect assumes that aggregated group knowledge provides reliable estimates of true knowledge and that judgments based on group knowledge work better than a single person's judgments (Surowiecki, 2004). In the following paragraph, the consensus theory of truth and the WoC effect are described and discussed with regard to the rationale of consensus scoring.

2.3.3.1 Consensus Theories of Truth. The term consensus refers to the agreement of a group of people concerning a particular proposition, opinion or statement (Rescher, 1993). The meaning of the term itself does not necessarily include the process of maintaining consensus. Consensus might be the result of a direct interaction which aimed to reach consensus, or it might be the agreement of independent individuals.

From the philosophical perspective, consensus is a concept of key importance that has a long history in philosophical debates. One key question in these discussions is: What conclusions can appropriately be drawn in a situation of agreement of individuals? Some philosophers postulate that under specific circumstances and conditions consensus might indicate truth. Consensus theories of truth were influenced by different philosophers (Rescher, 1993). The scope of this dissertation does not allow mentioning all of them; however, the most important contributions will be described and later discussed with regard to a theoretical foundation for consensus scoring.

Rescher (1993) mentioned several historically significant scholars who supported the idea of accordance between consensus and truth. Ancient theorists such as Aristotle viewed consensus as a supplement to deductive reasoning. These ideas will be addressed first. In contrast, other scholars, such as Jürgen Habermas, John Stuart Mill or Charles Sander Peirce, viewed consensus as the product and result of reason, as a future event that will exist in the long run (Rescher, 1993). The ideas of Jürgen Habermas will be outlined as exemplary for this line of thought.

Aristotle. Aristotle acknowledged the reasoning of a crowd in his Topics (Aristotle, 350 B.C.E), Nicomachean Ethics (Aristotle, 350 B.C.E) and Politics (Aristotle, 350 B.C.E). The *consensus omnium* is according to Aristotle a valid criterion of truth. Aristotle assigned consensus the status of substantial weight for a specific type of knowledge or reasoning. In his Topics he differentiated between demonstrative reasoning, dialectical reasoning and contentious reasoning. The first type of reasoning is inferred from true and primary premises and therefore is similar to deductive reasoning. Dialectical reasoning infers from premises grounded on generally accepted knowledge that is true for all or most individuals, or for philosophers. However, Aristotle also accepted the fact that generally accepted opinions might be wrong. In this case, reasoning is *contentious.* This type of reasoning starts with seemingly generally accepted opinions which prove, however, not to be generally accepted, and are therefore wrong. Hence, this type of reasoning should not be called reasoning, as it only appears to be such.

Habermas. Jürgen Habermas is the most prominent current exponent of the consensus theory of truth (Hesse, 1978). Habermas assumed that consensus is both the aim and the result of rational thinking and investigation (Habermas, 1984). Instead of claiming that truth can be found experimentally, he discussed the discursive foundation of truth (Healy, 1987). He rejected the idea behind the correspondence theory of truth, namely that truth is based on objective experience. He saw facts as linguistic terms and concluded that truth must be defined within discursive practice. The consensus theory of truth focuses on the conditions under which truth may be claimed as an outcome. In his 1972 chapter on theories of truth, Habermas states that the potential agreement of all is a condition of truth (Habermas, 1984). The truth of a proposition is defined by the fact that its discursive claim to validity is irredeemable. However, agreement among individuals does not indicate truth per se. Rational agreement as a result of discourse is suitable as a criterion of truth, whereas random agreement among a group is not.

Consensus is only an indicator of truth if it is the result of an ideal speech situation. The ideal speech situation has four conditions: 1) all potential participants ought to have the same chance of participating in the communication, 2) all participants ought to have the same chance of presenting their views, justifications or explanations, 3) all participants ought to have the same chance of using representative speech acts, that is, acts which refer to the intentions of the individual and 4) all participants ought to have the same chance of using regulative speech acts, such as to order, request, confirm, and so forth. The first and fourth conditions are necessary, though not sufficient for an ideal speech situation. The third condition refers to the truthfulness of the speaker and the second to the possibility to express one's thoughts freely in the ideal speech situation. These two conditions are sufficient for an ideal speech situation.

Habermas' consensus theory of truth has been extensively discussed in the philosophical literature. Rescher (1993) pointed out the distinction between the empirical reality of consensus and Habermas' designed hypothetical ideal consensus situation. He concluded that the view of consensus shifted away in Habermas' theory from de facto consensus to an idealized, non-feasible consensus. Healy (1987) criticized the lack of reasons for the link between consensus and truth and questioned the status of consensus as a definition (or even weaker criterion) of truth. He concluded that consensus cannot serve as a definition of truth. The circularity of the theory, namely that truth is defined by rational consensus which is itself defined by the ideal speech situation of competent individuals whose competence is defined by rational consensus, is one of the most widely discussed shortcomings of Habermas's theory (Beckermann, 1972; Ferrara, 1987). Other critiques concern its relation to the realist correspondence theory of truth. Habermas rejected the correspondence theory, but did not succeed in showing that his consensus theory of truth is non-realistic (Beckermann, 1972). Several assumptions of his theory are grounded in fundamental realist ideas. This inconsistency has also been criticized by other scholars (Ferrara, 1987). Another critical point refers to the theory's lack of applicability. As Ferrara (1987) pointed out, consensus in science, provided by scientific experts and most widely seen as true, has still been misleading and later proved untrue. Hence, it remains unclear how an ideal speech situation can be attained that guarantees truth in terms of consensus.

Philosophical Concerns about Consensus. According to Rescher (1993) skeptical philosophers like John Lock as well as other philosophers such as Plato and social scientists, have raised concerns about the meaning of consensus. Skeptics question the ability of the human mind on its own to make valid factual claims and reject the appropriateness of consensus as an indicator of truth. Rescher (1993) himself concluded that consensus is neither a definition nor a criterion of truth, but can at best serve as evidential support for the truth of a proposition. In particular, the consensus of competent experts can give evidence about truth; but, however, consensus in this view provides an epistemological-criteriological utility that is very different from the view that consensus is a functional equivalence of truth.

2.3.3.2 Wisdom of the Crowd Effect. Aggregated judgments of a large group of independent individuals have been hypothesized providing accurate predictions and suggested as an alternative to judgments of single, even highly able, individuals. The WoC effect has been commonly used for everyday judgments, and has been addressed in popular science literature (Surowiecki, 2004) as well as scientific literature (Mannes, Soll, & Larrick, 2014). Probably the most famous example of WoC in psychological literature was described by Francis Galton (1907) in his article vox populi. He described his visit to a fat stock exhibition where he collected ox' weights estimates given by farmers and casual visitors. The median of $N = 787$ data points was used as an estimate of the ox's true weight. As a matter of fact, this estimate was very close to the true weight, as the difference was only nine pounds. Galton expressed himself surprised that "democratic" judgments could be so trustworthy. Galton has been frequently cited as marking the point where the consensus idea entered social science research; the idea itself, however, is much older.

Applications of WoC Effect. WoC has been especially popular in economic or psychological applications (Gaissmaier & Marewski, 2011; Lee, Zhang, & Shi, 2011; Yi, Steyvers, Lee, & Dry, 2012). Beginning with Surowiecki's best-seller of 2004, the number of scientific studies of the WoC effect has increased. However, the aggregation of judgments or forecasts has a long history in psychological, economical, and statistical research. Clemen (1989) provided an overview of different methods, as well as applications of aggregated judgments. The WoC effect has been frequently known under terms like "swarm intelligence", "collective intelligence" or "aggregated group judgements" (Krause, James, Faria, Ruxton, & Krause, 2011; Nofer & Hinz, 2014; Surowiecki, 2004). However, the literature concerning the quality of group judgments in general covered various sorts of judgment. One very important classification attribute is whether judgments are given independently and aggregated mechanically, or result from some kind of group process that prohibits independence. WoC studies usually refer to the first case, whereas studies about group decision-making focus on interacting groups (Gigone & Hastie, 1997; Solomon, 2006). Crowds usually consist of different judges, but studies have also considered the WoC effect within individual respondents (Vul & Pashler, 2008). WoC effects have been investigated for judgments of group-related as well as group-unrelated events. For one thing, crowds were used to predict events in which members of the group in the future possibly could be part of. For instance, Gaissmaier and Marewski (2011) investigated the quality of election forecasts based on recognition of heuristic and WoC effects. As these authors showed, WoC generally performed better than, or at least as well as, prediction models based on recognition of parties and traditional polls. In economics, WoC has been used to make stock predictions. Hill and Ready-Campbell (2011) showed that WoC can outperform a famous stock index (S&P500). Predictions could be further improved by application of an algorithm reducing the crowd to specific experts. In a similar study, Nofer and Hinz (2014) investigated whether internet crowds are better in making stock predictions than professional experts. Based on analysis of data across several years, the authors concluded that crowd predictions provided higher returns than the predictions of experts. WoC has also been successfully used to predict events that are not group-related. For example, the WoC effect has been used to predict results of sports events (Herzog & Hertwig, 2011), to improve estimations of the correct price in the Price is Right (Lee et al., 2011), to solve complex problems (Yi et al., 2012), to estimate quantities (Krause et al., 2011) or to predict natural phenomena (Hueffer, Fonseca, Leiserowitz, & Taylor, 2013).

Most studies of the WoC effect present supporting results for the effect. However, few studies criticize the WoC effect or show limiting factors. Among these are Simmons, Nelson, Galak, and Frederick (2010) as well as Stephen and Lee (2010), who showed that crowds can make unwise decisions in sports betting when biased point spreads are available, even when the crowd knows that they are biased. In addition, Lorenz, Rauhut, Schweitzer, and Helbing (2011) showed that social influences provided as information by others could undermine the WoC effect.

Definition and Preconditions of WoC Effect. How did most studies come to the conclusion that the WoC effect exists? To define a group judgment as being *wise*, attributes of the judgment must be stated that allow precise evaluation of the wisdom. Firstly, it seems necessary that an objective criterion is available to compare the group judgment with. The group judgment needs to constitute a sufficient estimate of the target value. For instance, the ox weight measured with a calibrated scale is the target value in Galton's example. Secondly, group judgments are compared to other estimation procedures to make explicit in relation to which other procedures the crowd is considered comparatively wise (Mannes et al., 2014). Most commonly, the group judgment is compared to (1) a randomly selected person of the group or (2) an expert (i.e., an individual with high ability). Experts can also constitute estimators based on historical events or similar procedures (Hueffer et al., 2013). The third important question concerns the method of aggregating group judgments. Most often, central tendency measures are used, but other methods have been used, for example, in economics (Clemen, 1989; Merkle & Steyvers, 2011; Turner, Steyvers, Merkle, Budescu, & Wallsten, 2014).

According to Davis-Stober et al. (2014), the WoC effect can be defined as follows, incorporating the three aspects described above: "A crowd is wise if a linear aggregate, for example a mean, of its members' judgments is closer to the target value than a randomly, but not necessarily uniformly, sampled member of the crowd." (p. 1).

The WoC effect is usually bound to specific preconditions. However, the preconditions are seldomly mentioned in applications and the effect is hardly ever defined (Davis-Stober et al., 2014). Preconditions for the WoC effect are stated relatively consistently as (1) knowledge about the event, and (2) diversity and independence of judgments (Davis-Stober et al., 2014; Galton, 1907; Larrick et al., 2012; Nofer & Hinz, 2014; Surowiecki, 2004). Some authors additionally state motivation of respondents as a precondition (Nofer & Hinz, 2014):

- *Knowledge:* Larrick et al. (2012) state that judges should possess a level of expertise in the sense that they need to have experience with the event or that they are educated in the specific area. Although knowledge is a precondition of key importance, the precondition is not precisely defined insofar as cut-off levels of knowledge are elusive.

- *Diversity and independence:* Judgments need to be given independently from each other: that is, the judgment of one person should not influence judgments of other individuals (Nofer & Hinz, 2014). As a definitional characteristic, this aspect distinguishes the WoC effect from group decision making. However, the essential idea behind the independence precondition is a statistical assumption of error of measurement. Each measurement is distorted by error that can be either systematic or random. Averaging across different measurement (data) points has been shown to eliminate random error (Eysenck, 1939). However, systematic error is not eliminated by averaging. Non-independent judgment are prone to such systematic error, for instance error based on social influence.

Concerning diversity, Larrick et al. (2012) hypothesize that judges should have different perspectives on the event, should use different cues for judging, and therefore differ in errors. According to Nofer and Hinz (2014) diversity is advantageous because a diverse group provides more alternatives as sources of information and knowledge. According to Larrick et al. (2012) diversity and independence cannot be distinguished, as dependence results in less diversity. Davis-Stober et al. (2014) defined diversity as the highest possible negative correlation between respondents and conclude on the basis of their mathematical considerations that not independence, but maximal negative dependence will improve the WoC effect.

Although the aspect of diversity is mentioned frequently, the theoretical foundation is still questionable. It does not become evident why different perspectives of individuals ought to foster WoC effects. In addition, confusion of diversity and independence is misleading from a psychometric perspective. The definition of independence as zero correlation between respondents or of diversity as maximum negative correlation between respondents is rather unusual in psychometric theory. Usually, independence is an assumption in psychometric models; however this assumption does allow respondents to correlate positively. The assumption of local stochastic independence means no other latent variable than the assumed one should influence responses. Hence, responses of individuals show zero correlations only if the level of the latent variable is held constant. Although local stochastic independence constitutes a psychometrically reasonable assumption as a definitional aspect of the WoC effect, this assumption is seemingly unnecessary for group judgments to be wise (Gigone & Hastie, 1997). In addition, theories of expertise and agreement among individuals indicate that experts yield a common truth which causes their judgments to correlate (Batchelder & Romney, 1988; Legree et al., 2005). Hence, it remains unclear why diversity or zero correlations should possess any advantages.

2.3.4 Concluding Comments. Having introduced the measurement of SI and EI, addressed the challenge of the scoring for their measurement (see subsection 2.1) and referred to the philosophical view on measurement and the few scoring-related statements in intelligence literature (see subsection 2.2), the last subsection 2.3 presented hitherto popular empirical scoring methods as well as a potential extension of the theoretical foundation of consensus scoring. The major findings will be summarized and integrated in the next paragraph, followed by a discussion of the potential rationale of consensus scoring.

2.3.4.1 Summary and Integration of Literature Findings. Considering the huge amount of intelligence literature - thousands of articles and books in the last 100 years - it is somehow surprising how little attention has been drawn to scoring rules for *classical* intelligence measurement; in particular bearing in mind the critical discussion of scoring for the measurement of *new* ability constructs. Although the literature review presented in section 2.2.2 cannot claimed to be complete in the sense that there is no other article or book chapter mentioning scoring rules; however, the search for those articles and book chapters was tedious. This challenge reveals that scoring rules for *classical* intelligence measurement have been neglected as a part of theory and measurement of the construct. Moreover, even if they have been addressed, they were defined quite heterogeneously.

Although it is asserted quite frequently that correct responses of *classical* intelligence tests are unequivocally defined on the basis of veridical standards, the question why one specific response is the correct one in theoretical terms is most often dismissed for intelligence test. The evaluation standards for correct responses are rarely discussed from a theoretical perspective, although they can be hypothesized to a fundamental part of the definition of intelligence. It seems that every intelligence researcher knows what these evaluation standards are (as they are frequently used), but they have been explicated rarely. As a matter of course, for some abilities the scoring is less critical, for example, when scoring rules are defined in terms of speed, recognition, or deduction. For other abilities, however, the question of correctness is by far not as easily answered. The area of inductive reasoning gives an example of a (theoretical) scoring problem and of how literature in this area lacks of corresponding discussions.

The point of this subsection is not to question the stated correctness of response options in established *classical* intelligence assessments, but to raise the general question: On the basis of which veridical standards is correctness defined? This question has not been sufficiently answered in relevant literature. Nonetheless, these veridical standards have been set as a criterion which the measurement of *new* abilities has to meet.

The definitions and comments of scoring rules allow one to deduce that intelligence as a construct, as well as the evaluation of intelligent performance for its measurement, requires concepts of truth. Intelligent behavior is defined by correct behavior, which needs to be evaluated as correct or incorrect given specific (truth) standards. This idea has been addressed frequently, but, however, rather implicitly, referring to terms such as *objective* (rather than subjective) and *veridical*. However, the term objective on its own leaves plenty of room for interpretation and discussion. Convenient (objective) types of rules have not sufficiently delineated from non-convenient (non-objective) types of rules on a theoretical basis. For instance, cultural rules are not seen as appropriate for a definition of correctness, but the differentiations between semantic (dictionary) and cultural rules has not been further addressed (Guttman & Levy, 1991). Although the link between correctness and truth has been stated quite early by E. L. Thorndike et al. (1926), this idea has only once been further discussed (Nevo, 1993). It still remains open, which exact truth criteria have been used for intelligence test construction, and which should be used. It also has become obvious that truth theories do not provide clear criteria for correctness.

The truth criterion *consensus* is of key importance for this study. It has been one of the major reasons for the debate of the validity of EI measurements. Many authors questioning the validity of consensus-scoring methods have pointed at *standard* intelligence tests and claimed that EI measurement needs veridical scoring keys, just as these intelligence test. Therefore, it might be quite surprising that consensus was mentioned and used as a scoring criterion - as revealed by literature review - for these intelligence tests. Albeit some researchers justified consensus scoring with the reason that it has been used before in intelligence tests, however, the previous use of a criterion does not necessarily provide means for its justification. In fact, the respective theoretical rationale of consensus scoring has been only insufficiently elaborated and the empirical investigation of its quality is limited to few studies.

2.3.4.2 Towards a Rationale for Consensus Scoring. Proceeding from these findings, two areas which may contribute to a consolidated rationale of consensus scoring have been presented in the last section: Consensus theory of truth and the WoC effect. Do these areas help establish a rationale for consensus scoring that supports or goes beyond the reasons provided by Legree et al. (2005)? Is the consensus of individuals a sufficient source of knowledge to fulfill the definition of correctness? If so, under what conditions?

From a philosophical perspective, consensus as an indicator of truth is controversial. Even philosophical theories that support consensus as a criterion of truth set strict preconditions for that situation. From that perspective, consensus as a result of ideal group interaction, as opposed to random de-facto consensus, is viewed as an appropriate criterion of truth. The literature on the WoC effect, in particular popular-science literature, allocates consensus a variety of meanings and presents various examples of the effect. However, even here, limiting preconditions of the WoC effect are defined. These two areas partly underpin the rationale for consensus scoring as presented by Legree et al. (2005), however, both the WoC effect and the consensus theory of truth emphasize the boundaries to consensus in terms of preconditions.

One of the preconditions for consensus to indicate truth, ability, is a common element in philosophical discussions of consensus as a criterion of truth and in literature about the WoC effect. Both areas support Legree et al. (2005), who stated knowledge as an influencing factor for the quality of CBM. Hence, the role of ability/knowledge is not only supported by theoretical assumptions, but is highlighted as probably the key influence on the quality of consensus scoring.

Other mentioned theoretical preconditions are not supported throughout the literature. For consensus scoring to indicate unknown knowledge structures, diversity cannot be justified on a theoretical basis. The assumption that the variety of experience of journeyman might exceed the experience of experts is not directly supported when it comes to true knowledge structures (for scoring in ability measurement). There is no theoretical link between the variety or diversity of a group and their mean ability. In contrast, agreement among individuals, that is, less diversity, has been seen as a result of a high level of knowledge. Hence, it remains unclear why diversity should increase quality of consensus scoring when individual judgments are averaged. However, diversity might play a different role when it comes to interaction of groups, and also regarding the consensus of opinions and decisions that are not directly related to the concepts of correctness or true knowledge structures.

Moreover, independence is not required as a theoretical precondition for consensus to indicate facts. The ideal speech situation, fictional as it is, introduces a situation of highly dependent individuals. As an interacting group can predict objective target values (Gigone & Hastie, 1997), there is no empirical or theoretical reason why interacting groups per se should be less appropriate than mean tendencies of individual judgments. However, the consensus of independent and dependent groups might be different with regard to possible conclusions (Lorenz et al., 2011; Solomon, 2006). As studies from social psychology have shown, group interaction might lead to conformity under special conditions (Asch, 1955, 1956; Deutsch & Gerard, 1955). Conformity is a kind of consensus that might be different from independent group consensus, as group processes such as bias may shape it. On the other hand, consensus resulting from an ideal interaction need not necessarily differ from independent consensus, or might be even better in terms of possible conclusions (Gigone & Hastie, 1997). Independence has not been discussed as a major influencing variable for the quality of consensus scoring, yet consensus scoring methods require independent judgments.

Hence, transferring these thoughts to consensus scoring, the rationale of CBM (Legree et al., 2005) is only partly supported. Although consensus has been mentioned in intelligence literature as one evaluation standard for correctness, the theoretical rationale for consensus scoring, that is, consensus as an indicator of truth, cannot be justified stringently. From the theoretical perspective, only the consensus of highly able individuals may indicate correctness. If this condition is fulfilled, and consensus can indicate true knowledge structure, however, would it meet state-of-the-art standards for psychological measurement?

For investigation of consensus as a scoring criterion - which is the major aim here - one needs to canvass philosophical standards to measurement to evaluate this correctness criteria. From the philosophical perspective, the distinction of realist from non-realist theories of truth (and hence correctness) is important insofar as realist measurement ought to be based on a realist definition of scoring rules. The consensus theory of truth, however, is mostly seen as a non-realist criterion for truth, although some philosophers viewed parts of it as realist. If, on the other hand, correctness is defined on the basis of non-realist criteria, measurement is aligned to a non-realist perspective. Or can a correctness criterion, which is not independent of human thinking about it, be used for realist measurement? A more thorough philosophical discussion of this point is beyond the scope of this study, however, from a realist point of view, consensus should not define correctness, but rather - if at all - provide evidential support for (realist) correctness.

To conclude, consensus-based scoring should be seen critically from a theoretical perspective, as there is no unconditional theoretical support for consensus as an indicator of truth. However, because it might work under certain conditions, and because the idea has been used for scoring of intelligence tests, a systematic empirical investigation is required to evaluate empirical scoring methods which base on consensus. For this systematic evaluation, the true correctness of a response has to be known, as it has been partly implemented for EI research. More studies are needed which compare the consensus scoring keys to veridical scoring keys. Besides the pending evaluation of the CBM scoring methods, the investigation of potentially promising other scoring methods, which have not been used for EI or SI, is a research aim addressed in the present studies. Without further investigation of scoring methods, one can surmise that other challenges concerning the measurement of new ability constructs that were only touched upon in this chapter will remain unsolved.

3 Description of Empirical Scoring Methods

This chapter will describe different methods that have been used, or suggested for use, for scoring measurements of ability constructs. Because of the intense critique of CBM scoring, the alternative methods are introduced and investigated in the present studies. As empirical scoring methods, they supplement CBM described in section 2.3 and might represent promising alternatives to CBM scoring methods. The first of these methods is based on the assumption that agreement among respondents (consensus) indicates truth. Two other methods do not explicitly state this assumption, but use response information to estimate the ordering of response options with respect to the latent ability. These data-driven scoring methods are used on the assumption that an underlying true ordering of response options exists, although this may not yet be evident - due, for example, to lack of elaborated theory.

Note, however, that the selection of scoring methods presented below is not claimed to be complete. Other empirical scoring methods not investigated any further here may also be available (e.g., Clemen, 1989; Merkle & Steyvers, 2011; Turner et al., 2014).

3.1 Consensus Analysis

Consensus Analysis includes a class of methods, also known as Cultural Consensus Theory (CTT), that were developed in order to gain objective empirical evidence as a basis for *true* cultural knowledge - that is, the knowledge inherent in a culture (Romney, Weller, & Batchelder, 1986; Romney, Batchelder, & Weller, 1987). These methods emerged in a period when the objectivity of anthropological studies was questioned. According to Batchelder and Romney (1988), CTT provides a statistical model that can objectively describe and define cultural knowledge. However, as Karabatsos and Batchelder (2003) state, the knowledge in question is group-specific, so the terms *true* and *objective* are somehow misleading.

The CTT model is based on the idea that informants do not necessarily provide true information, but that knowledge statements of respondents are probabilistic, that is, as a function of knowledge, the probability of giving a correct response to a specific question will increase. Respondents with high ability tend to have higher probability of knowing the correct response to a question and thus give better information about the correctness of the endorsed response options. In CTT, the probability of a response option being correct is defined mainly by the competence of respondents endorsing that option. Additional processes such as difficulty of an item or a bias parameter may also be modeled. The CTT model is structurally related to the latent class model (Batchelder & Romney, 1988; Romney, 1999). However, it is not applied to respondents, but to items allocated to correctness classes.

As Batchelder and Romney (1988) state, CTT can be used for ability testing, in particular knowledge testing, whenever the researcher assumes that a common knowledge structure exists, but does not yet possess that knowledge. The crucial assumption of the model is that individuals who agree share a common knowledge, and that the level of agreement of independent respondents represents a measurable degree of true knowledge. Thus, like CBM scoring, CTT uses the consensus among respondents to estimate an unknown response key. However, in contrast to CBM, responses are weighted with the competencies of respondents. As Weller (1987) has stated, unweighted consensus among respondents also converges towards a true response; however using CTT models, smaller sample sizes might be possible and higher validity might be provided.

CTT has been suggested as an alternative scoring method for EI by Legree et al. (2005) and Schulze et al. (2007). However, no research has yet been published that uses CTT to score tests for EI, SI, or ability tests in psychological research in general. Possible reasons for this may be the complexity of the methods, strong model assumptions, or software limitations. In recent years, the development of CTT models has proceeded, and many versions of such models, as well as estimation procedures, are available (Anders & Batchelder, 2015; Aßfalg & Erdfelder, 2012a; Karabatsos & Batchelder, 2003; Oravecz, Anders, & Batchelder, 2015). The following paragraphs will first describe the statistical model and its variants with reference to different estimation procedures and then present a number of CTT applications.

3.1.1 Statistical Model and Estimation. In its central function, CTT aims to (1) estimate the (cultural) knowledge of respondents and (2) use this knowledge to estimate the true scoring key of items. Different estimation methods of the statistical model are described in the literature. Methods are available that can handle categorical (e.g., true-false, multiple choice, and fill-in items) (Romney et al., 1986) as well as ordinal items (Anders & Batchelder, 2015). Note, however, that estimation procedures are partly distinct for different response formats.

The model presented here applies to dichotomous data, and estimation procedures are described in particular for the dichotomous model. In line with the available literature, models and estimation procedures for data with more than two categories are outlined wherever this is necessary and possible. However, these models and estimation procedures are seldomly described precisely in the relevant literature.

In the basic CTT model, $i = 1, ..., N$ respondents respond to $k = 1, ..., M$ questions with categorical response format with $l = 1, ..., L$ categories. In the response profile data (raw data) for dichotomous items a value of one represents the endorsement of one response option (e.g., *yes* or *correct*) and a value of zero the endorsement of the other answer option (e.g., *no* or *incorrect*). This type of data does not provide any information per se about the correctness of the response options. The true response key of the data is unknown, but the existence of a true answer key is a crucial assumption. For each item the response key has a value of one if the true response to the item is *yes* and the response key has a value of zero if the true response to the item is *no*. Applying the scoring key to the response data gives the performance profile data (scored data). Here, a respondent earns a value of one if the correct response to an item was given and a value of zero if an incorrect response was given. The response profile data is unknown. Referring to Batchelder and

Romney (1988, p. 72) the three situations can be expressed as follows:

1. Raw data. $\mathbf{X} = (X_{ik})_{N \times M}$, where

   $X_{ik} = 1$ if respondent i answers *yes* to item k

   $X_{ik} = 0$ if respondent i answers *no* to item k

2. Answer key. $\mathbf{Z} = (Z_k)_{1 \times M}$, where

   $Z_k = 1$ if the correct response to item k is *yes*

   $Z_k = 0$ if the correct response to item k is *no*

3. Scored data. $\mathbf{Y} = (Y_{ik})_{N \times M}$, where

   $Y_{ik} = 1$ if respondent i gives the correct response to item k

   $Y_{ik} = 0$ if respondent i gives the incorrect response to item k

The CTT models aim to estimate the unknown response key based on the given raw data, while making some crucial assumptions. CTT includes a number of different models that can be classified as members of the General Condorcet Model (GCM) family (Aßfalg & Erdfelder, 2012a, 2012b; Batchelder & Romney, 1988; Karabatsos & Batchelder, 2003). GCM describes response probabilities of dichotomous items as hits- and false alarms. It makes the following two assumptions (Batchelder & Romney, 1988, p. 73).

1. Common truth. For each item there exists a true response, that is, $z_k = 1$ or $z_k = 0$, $k = 1, 2, ...M$. This assumption implies that the correct answer is true for all respondents, that is, all respondents come from the same culture.

2. Local independence. The informants respond independently, that is, the response to items is independent of other respondents, as well as being independent in the sense that no other response process except the necessary competence determines the answer.

   Local independence is formally stated as

$$P[(X_{ik})_{N\times M} = (x_{ik})|(Z_k)_{1\times M} = (z_k)] = \prod_{k=1}^{M}\prod_{i=1}^{N} P(X_{ik} = x_{ik}|Z_k = z_k) \tag{1}$$

The GCM model incorporates parts of the signal detection theory (Macmillan & Creelman, 2005). Hence, the probability of endorsing a response option, given that this option is the correct response, is described in the GCM as the hit rate (Batchelder & Romney, 1988, p. 72) with

$$H_{ik} = P(X_{ik} = 1|Z_k = 1) \tag{2}$$

The probability of endorsing a response option, given that it is the incorrect option, is described as the false alarm rate (Batchelder & Romney, 1988, p. 72) with

$$F_{ik} = P(X_{ik} = 1|Z_k = 0) \tag{3}$$

The probability of a correct response (Aßfalg & Erdfelder, 2012b, p. 7) is therefore

$$p_{ik} = H_{ik}^{Z_k}(1 - F_{ik})^{1-Z_k} \tag{4}$$

and the probability of an incorrect response is

$$p_{ik} = (1 - H_{ik})^{Z_k} F_{ik}^{1-Z_k} \tag{5}$$

Batchelder and Romney (1988) described a GCM sub-model, GCM1, with the additional restriction that hits and false alarms do not vary between items. Thus, the third assumption of the model is (Batchelder & Romney, 1988, p. 73):

3. Homogeneity of items. Each respondent has a fixed hit rate and a fixed false alarm rate with the additional constraint $0 < F_{ik} < H_{ik} < 1$, which is needed to identify the model. The probability of success for each respondent does not vary between items.

In the GCM1, without index k, the probability of a correct response (Aßfalg & Erdfelder, 2012b, p. 9) becomes

$$p_{ik} = p_z H_i + (1 - p_z)(1 - F_i) \tag{6}$$

whereas the probability of an incorrect response is

$$p_{ik} = p_z(1 - H_i) + (1 - p_z)F_i \tag{7}$$

where p_z is the prior probability that the correct response to item k is yes (Aßfalg & Erdfelder, 2012b) and accordingly that the item k falls into a specific correctness class (Batchelder & Romney, 1988).

GCM1 can be described as a multinomial processing tree model (Erdfelder et al., 2009) and therefore allows the modeling of response pattern on the item instead of single respondents. As Batchelder and Romney (1988) and Romney (1999) state, GCM is structurally isomorphic to a latent class model for two classes, where M items fall into two classes; that is, items with correct response option one and correct response option zero. GCM1 allows one to describe the probability of J response patterns given the correct response. The probability of a (correct) response pattern is according to Batchelder and Romney (1988, p. 75)

$$p_j = p_z \prod_{i=1}^{N} H_i^{x_{ij}}(1 - H_i)^{(1-x_{ij})} + (1 - p_z) \prod_{i=1}^{N} F_i^{x_{ij}}(1 - F_i)^{(1-x_{ij})} \tag{8}$$

As the key assumption of the model, hit and false alarm probabilities are dependent on different person as well as situation variables. The hit and false alarms are further modeled with additional parameters as

$$H_i = D_i + (1 - D_i)g_i \tag{9}$$

$$F_i = (1 - D_i)g_i \tag{10}$$

where $0 \leq D_i \leq 1$ is a competence parameter of a person and $0 \leq g_i \leq 1$ is a so-called bias (Batchelder & Romney, 1988, p. 73). In GCM1, the bias parameter is usually fixed for identification reasons as $g_i = 1/2$ (Karabatsos & Batchelder, 2003), in which case the additional restriction $F_i = 1 - H_i$ is needed (Batchelder & Romney, 1988). The probability of a response pattern then becomes

$$p_j = p_z \prod_{i=1}^{N} H_i^{x_{ij}} (1 - H_i)^{(1-x_{ij})} + (1 - p_z) \prod_{i=1}^{N} (1 - H_i)^{x_{ij}} H_i^{(1-x_{ij})} \tag{11}$$

where the hit probability for a response pattern is mainly determined by the competencies of the respondents.

Other GCM sub-models allow the assumptions of item homogeneity as well as fixed item bias to be relaxed. Releasing the constraints of item homogeneity, Batchelder and Romney (1988) suggested specifying the competency parameter in more detail as a version of a Rasch model (Karabatsos & Batchelder, 2003). Here, the competence parameter D_{ik} is defined as the probability of an individual i knowing the correct response to item k, as

$$D_{ik} = \frac{c_i(1 - d_k)}{c_i(1 - d_k) + (1 - c_i)d_k} \tag{12}$$

with c_i as the ability of a respondent i and d_k as the difficulty parameter of item k (Karabatsos & Batchelder, 2003, p. 375). With no knowledge of the correct response, the respondent guesses the correct answer with probability g_i (cf. Equations 9 and 10). In the event of knowledge about the correctness of a response, both the ability of a respondent and the difficulty of an item determine the hit probability. If item difficulties are homogenous (i.e., $d_k = .5, k = 1, ...M$) and response tendencies are homogenous (i.e., $g_i = .5, i = 1, ..., N$), the model is referred to as GCM1. Both restrictions can be weakened in other GCM sub-models, such as GCM2 or GCM3 (Karabatsos & Batchelder, 2003). GCM2 either assumes a neutral guessing parameter or item homogeneity; GCM3 allows free estimation of competence, guessing and difficulty.

GCM can be estimated using different methods. The most commonly used method, which uses factor analysis for estimation of competencies, allows parameter estimation for GCM1. Maximum Likelihood (ML) estimation is available for GCM1 as well as GCM2 assuming item homogeneity (Aßfalg & Erdfelder, 2012a; Batchelder & Romney, 1988). Markov Chain Monte Carlo Method (MCMC, Karabatsos & Batchelder, 2003) and Hierachical Bayesian Modeling (HBM, Oravecz, Vandekerckhove, & Batchelder, 2014; Oravecz et al., 2015) are the most flexible estimation methods allowing estimation of all GCM versions. However, these methods are quite time consuming. Referring to the so-called *informal* model, Weller (2007) presented a different estimation method, described as appropriate for rank order data. However, this estimation procedure simplifies the model radically and may be neglected for that reason.

The model and estimation procedures described so far only include two-categorical data. However, the factor analysis method allows consideration of more than two response options (Batchelder & Romney, 1988; Romney et al., 1986). For Hierarchical Bayesian Modeling other response formats can be estimated within different CTT models (Anders & Batchelder, 2015). For MCMC and ML estimation, so far, only the GCM model has been considered, and extension for more than two response options is only rarely mentioned in the literature.

Due to limitations both on the availability of methods for polytomous data and on processing speed of software for estimation procedures, this study only incorporates the restricted GCM1 model estimated with factor analysis estimation method (Batchelder & Romney, 1988). However, GCM1 is not necessarily the most reasonable model choice, as it is very restrictive. Other estimation procedures, such as MCMC, ML or HBM, have the advantage that the respective restrictions can be tested.

3.1.2 Consensus Analysis Based on Factor Analysis. The factor analysis estimation method can be subdivided into two steps. First, the competence parameters of respondents are estimated using minimum residual factor analysis (Comrey, 1962). Second, competencies of respondents are used to calculate the response key in accordance with the

Bayes theorem.

Based on the crucial assumption of the model that true knowledge is indicated by agreement, the matches of respondents are calculated in a first step of competence estimation as the proportion of matching m_{ij} of two informants i and j on all items, that is, in the case of two response categories (Aßfalg & Erdfelder, 2012b, p. 12):

$$m_{ij} = \frac{1}{M} \sum_{k=1}^{M} [x_{ik}x_{jk} + (1 - x_{ik})(1 - x_{jk})] \tag{13}$$

In the case of more than two response options the sum of the diagonal of the contingency table for a pair of respondents serves to express the matches. These matching scores are corrected for guessing as follows (Romney et al., 1986, p. 320):

$$\widehat{D_i D_j} = \frac{Lm_{ij} - 1}{L - 1} \tag{14}$$

The corrected matching scores are arranged in a matrix with N columns and with N rows which is used for a single factor minimum residual factor analysis based on Comrey (1962). This factor analysis method ignores the diagonal of the matrix and estimates the loadings on the factor, that is, the competence estimates (Weller, 2007). In order to apply the factor analysis method, one needs to assume that only one factor can account for the agreement between respondents, that is, that all respondents have a common truth. A somewhat loose criterion for judging if this assumption holds is the inspection of eigenvalues. The eigenvalue of the first factor should be roughly three times larger than that of the second factor (Batchelder & Romney, 1988; Romney et al., 1986; Romney, 1999; Weller, 2007). If this is not the case, the respondents may not be from the same culture. The CTT model also requires competence scores within a value space between zero and one; hence, negative scores are undefined by the model (Weller, 2007).

Based on the estimation of competencies, the posterior probability of a response option being correct, given the actual response pattern J_k, is

$$P(Z_k = l|J_k) = \frac{P(J_k|Z_k = l)P_l}{\sum_{z=1}^{L} P(J_k|Z_k = z)P_z} \tag{15}$$

where the a-priori probability of response l is set to $P_z = 1/L$ (Romney et al., 1986, p. 335). With this additional assumption, one only needs to estimate the conditional probability $P(J_k|Z_k = l)$. As stated by Romney et al. (1986, p. 335), the model implies the two following conditional probabilities

$$P(X_{ik} = l|Z_k = l) = D_i + (1 - D_i)/L \tag{16}$$

$$P(X_{ik} \neq l|Z_k = l) = (1 - D_i)(L - 1)/L \tag{17}$$

which, given the local independence assumption stated above, can be rearranged as follows

$$P(J_k|Z_k = l) = \prod_{i=1}^{N}[\hat{D}_i + (1 - \hat{D}_i)/L]^{X_{ik,l}}[(1 - \hat{D}_i)(L - 1)/L]^{1-X_{ik,l}} \tag{18}$$

$$= \prod_{i=1}^{N}\left[\frac{\hat{D}_i(L-1)+1}{(L-1)(1-\hat{D}_i)}\right]^{X_{ik,l}} \frac{(1-\hat{D}_i)(L-1)}{L} \tag{19}$$

where $X_{ik,l}$ is a random variable with $X_{ik,l} = 1$ if $X_{ik} = l$ and $X_{ik,l} = 0$ otherwise (Romney et al., 1986, p. 335). Based on this equation, the posterior probability for each response option can be calculated. The response option with the highest posterior probability is assumed to be correct. The CTT factor analysis method can easily be conducted using standard statistical software like SPSS or R.

3.1.3 Application. CTT models have mainly been applied in ethnographical or anthropological studies. A large number of studies focusses on cultural beliefs or knowledge about illnesses (Weller, 2007). To investigate whether emotion expression and emotion labeling show consensus among cultures, Alvarado and Jameson (1996, 2002) and Alvarado (1996) used cultural consensus models. Romney, Boyd, Moore, Batchelder, and Brazill

(1996) used CTT to investigate the semantic structure of kinship words for English speakers. Waubert de Puiseau, Aßfalg, Erdfelder, and Bernstein (2012) used the CTT model to improve estimation of true answers in an eye witness experiment. In personality psychology, CTT was used to assess the consensual understanding of responsibility and sociability (Webster, Iannucci, & Romney, 2002).

Because CTT was developed to indicate cultural truths, the methods seems to be particularly interesting for EI and SI. However, although suggested in the literature, no studies using CTT models for EI or SI have as yet been published.

3.2 Homogeneity Analysis

Homogeneity Analysis was developed independently by different scientific groups and is known under a variety of names. Tenenhaus and Young (1985) showed that methods like multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and nonlinear principal component analysis are closely related to each other and may be synthesized in a single underlying model. The following description of the method is based on homogeneity analysis with alternating least squares (HOMALS) algorithm (Gifi, 1989, 1990). Like principal component analysis for continuous variables, the HOMALS algorithm aims to reduce correlations between categorical variables to a few components that explain the maximum variance of the variables. In extracting these components, the algorithm quantifies the nominal categories of the variables. This quantification allows the categories to be treated as continuous and its result can be used as scoring weights.

3.2.1 Mathematical Model. The quantifications of $l = 1, ..., L$ categories of $k = 1, ..., M$ variables and the localization of $i = 1, ..., N$ respondents on the component (object scores) are estimated with an iterative algorithm based on the indicator matrix G of the data. The number of components has to be determined by the investigator. The indicator matrix contains a column for each category l of each variable m instead of one column for each variable. If a respondent has endorsed the category, the column has a value of one, if not it has a value of zero. For each variable the columns sum up to unity.

In case of nonlinear multivariate analysis, categories need to be optimally quantified if correlations between variables are to be reduced to components and respondents are to be mapped on these components (De Leeuw, 1984). In the HOMALS algorithm, categories are treated as purely nominal without any prior assumptions about their ordering. For the HOMALS algorithm, optimality is defined in terms of a loss function that is based on a proportionality assumption of category quantification and object scores. These form a perfectly consistent system if object scores are perfectly discriminating and category quantifications are perfectly homogenous (De Leeuw, 1984). Perfect discrimination of object scores is given if for all variables, objects in the same category receive the same score. Perfect homogeneity of category quantifications is given if a category endorsed by the same objects gets the same quantification.

Using an iterative algorithm, a loss function of consistency is minimized, which includes vectors of the object scores and category quantifications. Note however, that minimizing this loss function is equivalent to minimizing two different loss functions: the loss of discrimination and the loss of homogeneity (De Leeuw, 1984). The loss function includes that object scores and category quantification are proportional to each other (Gifi, 1990), that is, object scores are proportional to the averaged quantification of the categories endorsed by the respondents. Likewise, the quantification is proportional to the averaged object score of respondents who endorsed this category. To avoid trivial solutions of the algorithm, either the object scores or the quantification are normalized to unity (De Leeuw, 1984; Gifi, 1990). A trivial solution would be X = 0 and Y = 0.

The iterative algorithm (Gifi, 1990, p. 107) starts with random values (except for the value of zero) for object scores with a mean of zero. In the next step, the vector of random values is normalized. Based on these values, the first quantification of categories $\tilde{y}$ is calculated using the proportionality characteristic. Subsequently, the algorithm calculates the updated object scores, which are then normalized and again used to update quantifications. The fourth step of the HOMALS algorithm includes the convergence test, which is conducted on the basis of a fixed convergence criterion. The algorithm stops when the specified convergence value is reached.

As stated above, HOMALS treats categories as purely nominal. However, restrictions can be imposed on category quantifications using prior information (De Leeuw, 1984). For example, using ordinal prior information restrictions on quantification can ensure that ordering remains stable. However, these methods are not used in this dissertation and will not be described any further. See De Leeuw (1984) or Michailidis and de Leeuw (1998) for more detailed information.

Nominal category quantifications reflect the level to which the category accounts for the value on the component. A positive category quantification means that endorsement of this option indicates a numerically higher value on the component, whereas endorsement of a category with a negative quantification indicates a numerically lower value on the component. However, the polarity of the component is not determined theoretically or mathematically, meaning that a high positive value on the component, if used for ability items, does not necessarily indicate a high ability. As the algorithm aims to reduce the correlations of nominal categories without any information about rank order, the quantification itself is the first anchor on a continuum without theoretical polarity. Thus, the algorithm allows specification of rank order for categories; however, the object score does not necessarily allow inferences to be made about the polarity of the underlying ability.

3.2.2 Application. As described in subsection 2.3.2, MacCann et al. (2004) used reciprocal averaging and Keele and Bell (2009) used optimal scaling to improve characteristics of consensus scores. Reciprocal averaging and optimal scaling are very similar to HOMALS (Gifi, 1990; Tenenhaus & Young, 1985). Both of these studies used CBM scoring values as starting values and aimed to increase reliability of this scoring method. Results showed that CBM scores could be improved using these techniques;

however, the increase in reliability was not as substantial as hoped. No study has yet investigated whether the HOMALS algorithm is qualified as an empirical scoring method of purely nominal categories, and whether it succeeds in identifying the true response key.

3.3 Nominal Response Model

The NRM is an item response model providing means to model items with nominal response categories (Bock, 1972). The model allows one to estimate the relative weights, which reveal how much the respective category is indicating about the latent ability. The ordering of response options does not need to be specified in advance, except for the lowest and highest category. They should be fixed for identification reasons for some estimation procedures. Each response option is modeled as an indicator of an underlying latent variable (or latent variables). The psychological interpretation of the answering process is that each response option activates a tendency in each person. The tendency depends upon the trait or ability of the respondent. Each respondent endorses the response option for which the tendency is maximal. The NRM allows either specification of the order of response options with respect to the latent variable or testing a specific order hypothesized by the researcher. In many testing situations a theoretical rationale for the ordering of response options is available. However, in situations in which - due to lack of theoretical assumptions - the order is not evident, the NRM may allow empirical specification of the order.

The NRM was developed by Bock (1972) in order to extract information from seemingly wrong response options in multiple choice tests, or in any other item format with categories that might provide at least some information about the latent ability or trait. Missing data can also be included here as a nominal response alternative, thus modeling the missing data category with respect to the information it gives about the latent variable, in particular in comparison to wrong response options. The NRM was later formulated in a different way by Thissen, Cai, and Bock (2010) to enable modeling of more than one dimension.

The NRM is a generalization of the Generalized Partial Credit Model (GPCM, Muraki, 1992), where the ordering of the response options needs to be specified before modeling is conducted: that is, the GPCM can be re-formulated as the NRM with parameter restrictions. In addition, the rating scale model of Andrich (1978) can also be re-formulated as the NRM.

3.3.1 Statistical Model and Estimation. In the NRM the probability of individuals $i = 1, ..., N$ with ability θ_i endorsing category $l = 1, ..., L$ of item $k = 1, ..., M$ is modeled as

$$P(X_{ik} = l) = \frac{\exp(a_{kl_k}\theta_i + c_{kl_k})}{\sum_{h=1}^{L_k} \exp(a_{kh}\theta_i + c_{kh})} \tag{20}$$

where $h = 1, 2, ..., l_k, ..., L_k$; a_{kh} is the slope parameter and c_{kh} is the intercept for the h-th category of item k (Bock, 1972, p. 30). To identify the model, either the sum of slope and intercept parameter across each item is restricted to zero (Bock, 1972, p. 31), that is:

$$\sum_{h}^{L_k} c_{kh} = 0 \tag{21}$$

$$\sum_{h}^{L_k} a_{kh} = 0 \tag{22}$$

or the parameters of the first category a_0 and c_0 of each item are set to zero (Thissen et al., 2010, p. 45). In the more recent parametrization, identification is maintained setting - for every item - the slope value of one category to zero, the slope value of another category to $L - 1$, and the intercept parameter of the first category to zero (Chalmers, 2012).

The item parameters of the model cannot be interpreted analogously to the 2pl model item parameters difficulty and item discrimination, but they can be transformed into these parameters (Thissen et al., 2010). However, this is not necessary for scoring, as parameter a reveals the ordering of response options with respect to the latent variable. Thissen et al. (2010) explicitly state that the model "tells the data analyst the order of the item responses" (p. 52). The higher the value of a, the more information about the latent variable is given by endorsing the response options (Chalmers, 2012), that is, the more correct is the response option when measuring ability. The successful and numerically stable estimation of the model depends highly on a good choice in fixing the a values of the two response options. If these two response options actually represent the categories with lowest and highest correctness, then the estimation is likely to result in a true ordering; if the choice is bad, the chances are that the model will not allow an adequate ordering of response options. Chalmers (2012) recommends to either theoretically decide which options are high and low anchors, or to re-estimate the model with a different ordering if a values are too high (greater than 10).

Unconditional or conditional estimation of the NRM is described by Bock (1972). The R-package used in the present study (mirt) is specified with marginal ML with the expectation/maximization (EM) estimation for item parameters and the expected a posteriori (EAP) procedure for person parameter estimation (Chalmers, 2012). As De Ayala and Sava-Bolesta (1999) and DeMars (2003) state, like other polytomous item response models, estimation of the NRM requires large sample size. De Ayala and Sava-Bolesta (1999) studied the item parameter recovery of the NRM with marginal ML estimation. In a simulation study with 25 replications they investigated the effect of the ratio of examinees to item parameters, information of items, number of item categories and distribution of ability on the item parameter recovery of NRM. The number of items was held constant at 28, whereas sample size varied for the 3 and 4 option items between 420/560 and 3360/4480. Results showed that distribution of the latent variable and the ratio accounted for most of the variance in successfully estimating parameters. For a normally distributed ability a ratio of 10:1 indicated stable parameter recovery. For example, according to DeAyalas scheme, a twelve item inventory with five response options would require 1200 (960 with restrictions) respondents for stable estimation. In an expanding study, DeMars (2003) investigated the influence of the number of items as well as the number of item categories on item parameter estimation performance. It was found that the number of categories of items influenced the quality of scoring to a greater degree than the number of items itself. Again, the number of respondents had a great impact on recovery of parameter estimation. These two studies do not cover all possible influencing parameters, but indicate that for NRM item parameter estimation large sample sizes are needed.

3.3.2 Application. Although the NRM was developed over 40 years ago, it does not constitute a very popular item response model. Nevertheless, it has been used relatively frequently, for example, in adaptive testing (De Ayala, 1992; Lima Passos, Berger, & Tan, 2007), to detect answer copying (Kalender, 2012) or to test the ordering of response options of a Likert scale (Preston, Reise, Cai, & Hays, 2011).

Barchard et al. (2013) have suggested the NRM model for scoring of EI tests, however, did not further refer to any concrete applications. Guo et al. (2016) used the NRM model for scoring response options of non-cognitive measures (SJTs) for which no clear theoretical response key was available, and compared the method to several other scoring methods, that is, expert scoring, a variant of proportion as well as mode CBM, and scoring based on the GPCM. For the GPCM, they used expert scoring for response option ordering; however, it remains unclear how they specified the NRM model. It seems most likely, that they used keyed data as an input for NRM estimation. Based on the comparison of parametric as well as non-parametric item characteristic curves for the different scoring methods, they concluded that the scoring methods result in very similar keys. However, the reliability estimate was highest for NRM keyed data. Although the authors summarize that NRM seems to be very useful for scoring tests without clear keys, the possible conclusions of this study are limited. As some methodological proceedings remain unspecified, the evaluation of the method is still an open research question. Most importantly, their conclusions were based on data sets for which no correct response key was available and NRM scoring keys were only compared to expert and consensus keys, instead of objective keys; so the investigation of NRM as a scoring method still leaves room for improvement.

4 Aims and Objectives

The major aim of the present studies is the systematic investigation and evaluation of empirical scoring methods which might be used for the identification of a true scoring key in ability measurement. Empirical scoring methods most generally use the respondents' test behavior to infer information about the correctness of response options. From a philosophical and scientific perspective, criteria for the correctness of response options in ability tests should be based on trusted knowledge sources. However, whether empirical methods are able to discover knowledge structures is a crucial, and as yet unanswered, question, for which the comparison to objective criteria is a key requirement.

As the empirical scoring methods which have been used for the measurement of new ability constructs have been mainly based on group and expert consensus, the present studies aim to investigate and compare an enlarged pool of empirical scoring methods. Hence, alongside with mode and proportion CBM scoring methods, the three empirical scoring methods which have been introduced in Chapter 3 will be investigated. The key idea, that is, how the method uses the respondents behavior to infer information about correctness, however, differs with respect to these methods. The nature of the methods used is twofold: While some of the described empirical scoring methods are based on consensus (CBM and Consensus Analysis), others are not explicitly based on that theoretical background (HOMALS and NRM). Hence, the aims - and possible conclusions - of the investigations differ with respect to the diverse rationales.

Both mode/proportion CBM and Consensus Analysis incorporate the idea that consensus of respondents gives information about correctness, albeit different procedures underlie these methods. Previous studies which investigated consensus based scoring did not use objective criteria for comparison of CBM scoring keys (Barchard & Russell, 2006; MacCann et al., 2004) or used rather specific criteria or tests (Barchard et al., 2013; Mohoric et al., 2010). Hence, the empirical investigation of the research question remains fragmentary. Moreover, Consensus Analysis has not been applied in ability testing contexts of psychological research. For these methods, the present studies aim to complement the existing empirical evidence by evaluating the methods performance using even more objective criteria while controlling for possible confounding variables.

HOMALS, however, bases on the more technical idea that optimal weights are given to response options by maximizing homogeneity of variables and that these optimal weights might be used for scoring purposes. Similar methods have been used with CBM scores as starting values (Keele & Bell, 2009; MacCann et al., 2004), however, the method in its nominal version - that is, without any information about the true ordering of response options - has not been used before. Hence, the methods' quality with regard to empirical scoring is aimed to be explored in the present studies.

The NRM method provides estimates of each response options slope parameter, which inform about the information the respective option gives about the underlying ability. The use of the NRM for scoring is limited to one published study, which did not systematically evaluate this method. Hence, the present studies aimed to systematically evaluate the NRM model for the first time. Because the estimation procedure requires the slope values of two response options to be fixed, here, the aim was to compare different fixation methods, from those which are *purely* empirical, that is, no information about the true ordering of response options is provided for estimation, to those which incorporate some (uncertain) information about the true ordering. Hence, different scenarios that might emerge in measurement situations are compared as follows: (1) the researcher has certain assumptions about the most and least correct response options, (2) the researcher has no theoretical assumptions about the correctness of response options, and (3) the researcher has uncertain, but reasonable assumptions about the most and least correct response options.

Major parts of the investigation of consensual and empirical scoring procedures will be of exploratory nature, without specific hypotheses about the performance of the methods in every realized data situation. However, a second main part of the investigation is of confirmatory nature and different potentially influencing variables are considered. The hypotheses concerning factors that influence the performance of consensus-based scoring procedures are based on the theoretical assumptions about the validity of consensus, described in Chapter 2. In addition, the characteristics of the other empirical scoring procedures described in Chapter 3 motivate the investigation of certain other influencing factors.

- *Ability and difficulty.* Ability is hypothesized to be the main determinant of performance in consensus-based scoring methods. It is assumed that with increasing ability, performance in consensus-based scoring methods also is enhanced. Therefore, different ability levels will be investigated in the present studies, that is, samples with a low mean ability, an average mean ability as well as a high mean ability. Along with individual ability, the difficulty of items determines the probability of answering an item correctly. Difficulty is consequently hypothesized to be a second major determinant of performance in consensus-based scoring methods. Moreover, difficulty has been empirically identified to be a determinant of the quality of consensus scoring methods (Barchard et al., 2013). With increasing difficulty, performance in consensus-based scoring methods is hypothesized to decrease. To investigate the hypothesized influence, the difficulty of items will be varied, where the different levels represent low, medium and high item difficulty. These item difficulties are combined in varying ways to cause situations with equal item difficulties as well as differing item difficulties. One reason for the investigation of the influence of varying item difficulties is the assumption of equal item difficulties in the factor-analysis based estimation method of Consensus Analysis.

- *General data characteristics.* Other potentially influencing general data characteristics are sample size, number of items and number of response categories.

  - *Sample size.* In comparison with CBM, the advantage of Consensus Analysis is hypothesized to be its quality even for small samples: Here, Consensus Analysis is supposed to work better than CBM scoring methods (see subsection 3.1.1). However, difference between Consensus Analysis and CBM is hypothesized to vanish with larger sample size. In addition, sample size is hypothesized to be important for NRM scoring, as the quality of item parameter estimation depends on the sample size (see subsection 3.3.1). Here, a better estimation based on larger sample size might result in advantages. The sample size levels are geared to mirror typical sample sizes in psychological research, from very small samples of $N = 20$ to samples of average size (e.g., $N = 100$) to larger sample sizes from $N = 1000$. The smallest sample size is included to investigate the quality of Consensus Analysis for smaller and larger samples.

- *Number of items.* As Consensus Analysis estimates the competence of individuals in terms of agreement among respondents, the number of items to potentially agree upon might be important for this method. As the competence might be estimated more certain with higher number of items, higher number of items might result in better performance of the scoring method. For CBM, a higher number of items has been shown to be advantageous, too (Barchard et al., 2013). For NRM, the number of items influences item parameter estimation quality (see subsection 3.3.1) and, hence, the scoring method might function better with less variables. The different number of variables aim to mirror reasonable number of items of tests.

- *Number of response categories.* As Consensus Analysis estimation procedure slightly differs for number of response categories (dichotomous versus multi-categorical), this data characteristic might also have certain influence (see subsection 3.1.1), however, without a hypothesis about the direction of a possible effect.

Other data characteristics can be hypothesized to be important influencing variables, for example general assumptions about the relation between ability, difficulty and probability of success, variance in ability, assumptions about guessing processes, different response scales or dimensionality. These potentially influencing data characteristics will not be investigated here, but will be discussed later.

In order to systematically examine the performance of consensus-based and empirical scoring methods, objective rules need to be available in a situation that is closely comparable to that of intelligence measurement (Barchard et al., 2013; Mohoric et al., 2010; Schulze et al., 2007). In the following sections, two situations, both providing objective scoring rules, will be presented to investigate the comparative performance of scoring methods. First, simulation studies provide the possibility to systematically investigate the influence of different data characteristics on the quality of scoring using models that are common for ability testing. Second, the real-world data of ability tests allows evaluation of the methods in a real life situation.

4.1 Simulation Studies

Statistical research questions that cannot be adequately solved using analytical-mathematical analysis can be investigated using simulation studies. For instance, simulations can be used to determine the unknown distribution of a parameter or to test the robustness of methods when the preconditions are not met (X. Fan, 2012). With growing computer capacities, the number of simulation studies has increased in the last decades. In some fields of psychological assessment, such as adaptive testing based on Item Response Theory (IRT), simulation has become an indispensable tool.

Simulation studies are experimental research studies in which independent variables are varied to investigate their influence on dependent variables (Harwell, Stone, Hsu, & Kirisci, 1996; Skrondal, 2000). In line with this perspective, simulation studies should be planned and conducted in the same way as psychological experiments. Instances of independent variables in simulation experiments are data attributes that can be varied in accordance to theoretical considerations. The combination of factor levels represents the conditions of the simulation experiment. For every condition, several simulations should be conducted. For every combination of levels of independent variables, for example, 10 000 data sets (iterations) are generated and dependent variables are computed.

Simulation studies should be geared to fulfill several criteria. According to Harwell et al. (1996), many simulation studies based on IRT between 1981 and 1991 did not meet specific quality criteria, which were defined by the journal Psychometrika (Psychometric Society, 1979). This policy states that simulation studies should only be considered as a research method if no other way of providing the required results is available. The comparison of different algorithms for the same purpose, where simulation seems in most cases appropriate, constitutes an exception to this restriction. Harwell et al. (1996) provided additional criteria for simulation studies. According to these authors, the independent variables used to simulate data should be as realistic as possible, the quality of the random number generator should be good, and the number of replications should be as large as possible.

Of great importance for simulation studies is the external validity and the question of generalizability of the results. The range of possible parameter values of independent variables is usually very large and only some specifications of the independent variables can be used, so the values of independent variables should be chosen carefully. The conclusions of simulation studies can only be generalized to the degree that simulated data reflects reality (Davey, Nering, & Thompson, 1997). However, simulated data never fully reflects reality, since the model may not include all relevant processes. On the other hand, the internal validity of simulation studies is a great advantage, as influencing variables can be completely controlled. Simulation studies allow the control of possible confounding variables, as well as systematic variation of relevant characteristics. The influence of relevant data features on the performance of scoring methods can be validly attributed to the variation of these features.

The aim of the present simulation studies was to evaluate the empirical and consensus-based scoring methods systematically and to judge their performance using objective criteria. An advantage of simulated data is that, given the model as input, the true scoring key is known. Comparison of the true and reconstructed scoring keys - as well as the comparison of true abilities and ability estimates based on the scoring keys - on different levels allows impartial judgment of the respective scoring method. Moreover, simulation studies aim to estimate the influence of independent variables on the performance of these scoring methods. Systematic variation of data characteristics yields the opportunity to identify situations in which scoring methods may lead to success or show massive shortcomings. Alongside with these aims, the simulation studies targeted to identify possible effects of the scoring methods on the structure of the data.

4.2 Real-world-data Study

Simulation studies suffer from a bad reputation with regard to external validity. To what extent results are generalizable, depends largely on the simulation model and the design of the study. But the design can only include a small sample of the range of all parameters and the identified effects may not be exhaustive. The models used for simulation may neglect a process that is relevant for response genesis and therefore may not represent real-world data. For this reason, real-world data studies were conducted to investigate performance under realistic conditions of ability measurement. However, real-world data may suffer from a lot of situation-specific influences and therefore include confounding variables

Real-world data studies aimed to mirror a situation in ability measurement where the correct response is unequivocally defined by objective standards and to compare the methods scoring quality to these objective standards. Hence, data from an international educational comparison study has been used. These studies meet high standards, in particular with regard to quality of test construction, sampling and administration. Data from the TIMSS 2011 study was chosen to investigate scoring methods in a realistic situation satisfying such high standards.

5 Evaluation of Empirical Scoring Methods in Simulation Studies

5.1 Methods

5.1.1 Models for Simulation Studies. To generate realistic data for simulation studies, models were selected which are commonly used for ability measurement. Ability measurement data is often modeled with nonlinear mixed models, in particular item response models (Rijmen, Tuerlinckx, DeBoeck, & Kuppens, 2003). Item response models regress the probability of choosing a response option on item and person parameters. In the present study, models for two-categorical as well as five-categorical one-dimensional data were selected.

5.1.1.1 Model for Two-categorical Data. For simulation of one-dimensional, two-categorical data, the three parameter logistic (3-pl) Birnbaum model was used (Birnbaum, 1968). The 3-pl model is a highly flexible and widely used model (Rijmen et al., 2003) in which the probability of choosing one, usually the correct, of two response options is regressed on the ability θ of a person j, the difficulty β of the item i, the discrimination a of the item i and a pseudo-guessing parameter c.

$$P(x_{ij} = 1|\theta_j) = c + (1-c)\frac{exp[a_i(\theta_j - \beta_i)]}{1 + exp[a_i(\theta_j - \beta_i)]} \tag{23}$$

Ability and difficulty are defined on the same metric scale, with possible values between $-\infty$ and $+\infty$. The ability parameter represents the one-dimensional latent person variable. A value of zero indicates medium ability, whereas positive values represent higher ability and negative values represent lower ability. Difficulty is defined as the point on the ability scale where the probability of endorsing the correct response option of an item equals .50 if no pseudo-guessing is expected. The discrimination of an item is defined as the gradient of the function and gives information about the degree to which the item can discriminate between different ability levels. Values of the discrimination parameter range between 0 and $+\infty$. The pseudo-guessing parameter represents the probability of a correct answer being given by a person with ability reaching $-\infty$. Thus, in the 3-pl model two different processes are used to predict the answer to an item. With an ability close to $-\infty$, other (mental) processes than ability (e.g., guessing) may cause response behavior. With higher ability, however, person and item characteristics are modeled as predictors of response behavior.

5.1.1.2 Model for Five-categorical Data.

For the simulation of one-dimensional data with five categories, the GPCM (Muraki, 1992) was used. Here, the probability of a person j endorsing an answer category $k = 1, 2, ...m_i$ of item i is dependent upon the ability θ of the person j, the discrimination a of the item i as well as the difficulty β of the item i and the response option v.

$$P(x_{ij} = k|\theta_j) = \frac{exp[\sum_{v=1}^{k} a_i(\theta_j - \beta_{iv})]}{\sum_{c=1}^{m_i} exp[\sum_{v=1}^{c} a_i(\theta_j - \beta_{iv})]} \tag{24}$$

For five response categories, the model can be stated as follows (compare Tuerlinckx & Wang, 2004).

$$\begin{gathered} P(x_{ij} = 0|\theta_j) = \frac{1}{1 + exp(a_i(1\theta_j - \beta_{i1})) + exp(a_i(2\theta_j - \beta_{i1} - \beta_{i2})) +} \\ \frac{}{exp(a_i(3\theta_j - \beta_{i1} - \beta_{i2} - \beta_{i3})) + exp(a_i(4\theta_j - \beta_{i1} - \beta_{i2} - \beta_{i3} - \beta_{i4}))} \end{gathered} \tag{25}$$

$$\begin{gathered} P(x_{ij} = 1|\theta_j) = \frac{a_i(1\theta_j - \beta_{i1})}{1 + exp(a_i(1\theta_j - \beta_{i1})) + exp(a_i(2\theta_j - \beta_{i1} - \beta_{i2})) +} \\ \frac{}{exp(a_i(3\theta_j - \beta_{i1} - \beta_{i2} - \beta_{i3})) + exp(a_i(4\theta_j - \beta_{i1} - \beta_{i2} - \beta_{i3} - \beta_{i4}))} \end{gathered} \tag{26}$$

$$\begin{gathered} P(x_{ij} = 2|\theta_j) = \frac{a_i(2\theta_j - \beta_{i1} - \beta_{i2})}{1 + exp(a_i(1\theta_j - \beta_{i1})) + exp(a_i(2\theta_j - \beta_{i1} - \beta_{i2})) +} \\ \frac{}{exp(a_i(3\theta_j - \beta_{i1} - \beta_{i2} - \beta_{i3})) + exp(a_i(4\theta_j - \beta_{i1} - \beta_{i2} - \beta_{i3} - \beta_{i4}))} \end{gathered} \tag{27}$$

$$\begin{gathered}P(x_{ij}=3|\theta_j) = \frac{a_i(3\theta_j-\beta_{i1}-\beta_{i2}-\beta_{i3})}{1+exp(a_i(1\theta_j-\beta_{i1}))+exp(a_i(2\theta_j-\beta_{i1}-\beta_{i2}))+} \\ \overline{exp(a_i(3\theta_j-\beta_{i1}-\beta_{i2}-\beta_{i3}))+exp(a_i(4\theta_j-\beta_{i1}-\beta_{i2}-\beta_{i3}-\beta_{i4}))}\end{gathered} \tag{28}$$

$$\begin{gathered}P(x_{ij}=4|\theta_j) = \frac{a_i(4\theta_j-\beta_{i1}-\beta_{i2}-\beta_{i3}-\beta_{i4})}{1+exp(a_i(1\theta_j-\beta_{i1}))+exp(a_i(2\theta_j-\beta_{i1}-\beta_{i2}))+} \\ \overline{exp(a_i(3\theta_j-\beta_{i1}-\beta_{i2}-\beta_{i3}))+exp(a_i(4\theta_j-\beta_{i1}-\beta_{i2}-\beta_{i3}-\beta_{i4}))}\end{gathered} \tag{29}$$

The ability θ_j of person j and discrimination a of item i have the same interpretation as in the 3-pl model. The difficulty parameter β_{iv} now has a slightly different interpretation, as there is a difficulty parameter for $k-1$ of k categories. It can also be denominated as a threshold parameter. The threshold parameter β_{iv} for item i and category v is defined as the point on the ability scale where the categories k and $k-1$ have the same probability of endorsement. For the first category $k=0$ of each item i, the numerator in Equation 24 is replaced by 1 (Muraki, 1992).

5.1.2 Design of Simulation Studies. The simulation studies are designed to investigate the scoring methods in various, systematically differing situations that constitute a broad selection of real-world situations. However, the number of conditions of the experimental study is restricted by available resources. The selection of values was in the present instance geared to ensure adequate variation of values in a reasonable parameter space, and at the same time to provide a design that was manageable in terms of available time and computer resources. In the following paragraphs, the design of the simulation studies of the present research project will be described. Note that the number of categories is also an element of the design (independent variable). As for use of different models, the design will be described separately for two-categorical and five-categorical data. Henceforth, the number of categories will not be mentioned as an independent variable. This separation continues in the results sections, as differences in the results of dichotomous and polytomous data cannot be clearly attributed to the number of categories, but to the differing statistical models used for data generation.

5.1.2.1 Independent Variables. As stated in Table 5, a total of four independent variables was systematically varied for two-categorical data. The number of variables was fixed at the levels of 12, 24 and 48 items. For the independent variable number of respondents (sample size), five factor levels were selected, varying from very small sample of 20 respondents to a maximum sample of 1000.

Item difficulty was varied on six factor levels. For each level, the item sample was subdivided into three parts. Each subgroup of items was assigned to have either a difficulty of -1 (easy), 0 (medium difficult) or 1 (difficult). With repetition, ten combinations of difficulty were possible. Out of the ten possible combinations, six were chosen that were considered to be of most interest. Hence difficulty combinations with equal difficulties across all items were selected, as well as different combinations with unequal difficulties. For the combinations with equal difficulties, all of the items were easy [e,e,e], medium difficult [m,m,m] and difficult [d,d,d]. In addition, three mixed difficulty combinations were selected: (1) The first with one-third of the items each having either low, medium or high difficulty, respectively, for instance, four out of twelve items with low, four out of twelve items with medium and four out of twelve items with high difficulty [e,m,d]; (2) the second with two-thirds low and one-third medium difficulty, for instance, eight out of twelve items with low difficulty and four out of twelve with medium difficulty [e,e,m]; and (3) a third with two-thirds high and one-third with low difficulty, for instance, eight out of twelve items with high difficulty and four out of twelve with low difficulty [d,d,e]. These combinations were considered to be of most interest, as they reflect important situations. Mixed difficulty levels are more common and are favored in ability testing, especially the combination [e,m,d]. The combination [e,e,m] might be most common in EI research, where easy items are generally hypothesized to cause a high correlation between expert and group consensus scoring. The combination [d,d,e] reflects a situation with large differences in item difficulty. Mixed item difficulty combinations are especially interesting for Consensus Analysis, as equal item difficulties represent a model assumption.

Table 5

Independent Variables for Simulation Studies with Two-categorical Data

Parameter		Values	Number of values
m	Items	12, 24, 48	3
N	Respondents	20, 50, 100, 200, 1000	5
a_j	Discrimination	1	1
β_j	Difficulty	6 combinations of easy (-1), medium difficult (0) and difficult (1) items[a]	6
c_j	Guessing parameter	0	1
θ_i	Ability	N(-1,1); N(0,1); N(1,1)	3

Note. N(M, SD) = normal distribution with specified mean and standard deviation. [a]The combinations of difficulties are: [e,e,e], [e,e,m], [e,m,d], [m,m,m], [d,d,e], [d,d,d].

Respondent abilities were drawn from a normal distribution with mean of -1 (low ability), 0 (medium ability) and 1 (high ability), and a standard deviation of 1. The discrimination of items was held constant at one and the guessing parameter was held constant at zero, as time and computer resources did not allow variation of that parameter. As presented in Table 6, independent variables and factor levels are similar for five-categorical data. However, slightly different values were chosen for the threshold parameters. The combination of difficulty values was approximately comparable to the case of two-categorical data.

In the full design, for each two-categorical and five-categorical data, 270 conditions were selected for consideration. For each of the conditions, 10,000 data sets were simulated.

The number of iterations was chosen in order to provide sufficient repetitions to ensure reliable results.

Table 6

Independent Variables for Simulation Studies with Five-categorical Data

Parameter		Values	Number of values
m	Items	12, 24, 48	3
N	Respondents	20, 50, 100, 200, 1000	5
a_j	Discrimination	1	1
β_j	Difficulty	6 combinations of easy (-1.75; -1.25; -.75; -.25), medium difficult (-.75; -.25; .25; .75) and difficult (.25; .75; 1.25; 1.75) items[a]	6
θ_i	Ability	N(-1,1); N(0,1); N(1,1)	3

Note. N(M, SD) = normal distribution with specified mean and standard deviation.

[a]Combinations of difficulties are: [e,e,e], [e,e,m], [e,m,d], [m,m,m], [d,d,e], [d,d,d].

Table 7 shows an overview of simulation studies. The full design as described above was conducted for three scoring methods, namely proportion CBM, mode CBM and HOMALS. For Consensus Analysis, the design had to be reduced by number of respondents = 1000 due to a very time-consuming algorithm (216 conditions). Due to much larger sample size requirements, not all conditions seemed feasible for NRM scoring (see subsection 3.3.1). Thus, only conditions with sample size of 200 or more (108 conditions) were used for NRM scoring. In addition, NRM scoring was only conducted for five-categorical data. This was inevitable, because the software used for model estimation fixed the slope parameters of the first and last answer category in order to ensure identification. In addition, three versions of NRM scoring were investigated (see section

4.1): (1) with certain fixing, that is, the slope parameter of the most correct and the least correct response option were fixed with 100% probability, (2) with random fixing of slope parameters of two response options, and (3) with semi-random fixing, that is, the slope parameter of the most correct and least correct response option were fixed with a specific probability. Regarding version (3), in the first step, the slope parameter of the least correct response option was fixed with a probability of .60, while every other option was fixed with a probability of .10 as least correct. For the remaining response options, the highest option value was fixed with a probability of .70 as the most correct option, while the other response options were fixed as the most correct option with a probability of .10. These probabilities were selected to mirror a situation in which reasonable assumptions about the construct are available allowing the positing of likely hypotheses about which behavior indicates high ability and which behavior indicates low ability.

5.1.2.2 Dependent Variables and Data Analysis. The performance of each scoring method was evaluated using several dependent variables. The main focus was on the comparison of the true and reconstructed scoring keys of the items as well as on the comparison of the true and re-estimated abilities of the individuals. Moreover, the dimensionality of the data was tested after scoring to identify possible effects of the scoring methods on the one-dimensional structure of the items. For two- and five-categorical data, $5,400,000$ different data sets were simulated. After the application of scoring methods, and across types of data, conditions and replications, $23,760,000$ data sets were analyzed. For each analysis, the dependent variables were stored, and after $10,000$ replications the mean and standard deviation for each condition was calculated.[2] Overall, similar analyses were sought across different scoring methods to ensure comparability of results. However, for some scoring methods, dependent variables differed slightly. Dependent variables are described together for all scoring methods whenever this is possible. Distinctive features of

[2] For some dependent variables, different descriptive statistics were stored, as described with the respective dependent variable.

Table 7

Overview of Simulation Studies

No.	Scoring method		Data	Conditions	Replications
I.	CBM	Mode	two-categorical	270	10,000
II.	CBM	Proportion	two-categorical	270	10,000
III.	CBM	Mode	five-categorical	270	10,000
IV.	CBM	Proportion	five-categorical	270	10,000
V.	CA		two-categorical	216	10,000
VI.	CA		five-categorical	216	10,000
VII.	HOMALS		two-categorical	270	10,000
VIII.	HOMALS		five-categorical	270	10,000
IX.	NRM	a-random	five-categorical	108	10,000
X.	NRM	a-semi-fixed	five-categorical	108	10,000
XI.	NRM	a-fixed	five-categorical	108	10,000

Note. CBM = Mode Consensus Based Measurement; CA = Consensus Analysis; HOMALS = Homogeneity with Alternating Least Squares; NRM = Nominal Response Model.

the analysis for some scoring methods are mentioned where necessary.

Analysis on the Level of the Scoring Key. First, data was analyzed on the level of the scoring key. For each scoring method, the relative proximity or distance between the true scoring key and the respective reconstructed scoring key was calculated. To present the different measures used on this level, first, the different types of keys are introduced. Two types of scoring keys are used in the following: (1) Single vectors, with length equal to number of items, which contain the (according to the method) correct response options, and (2) matrices which contain scoring values for each response category of each item.

In general, different response options of simulated data are designated by different numerical values. In the original simulated data, the numerical values of response options incorporate an ordinal information. The true correct responses - that is, the original scoring key used for the comparison with the reconstructed scoring key - is based on that information: For two-categorical data, the possible response categories of the simulated data were designated with numerical values of one or zero. According to the simulation model, a value of one indicated a correct response option and a value of zero an incorrect response option. Thus, for two-categorical data the (1) vector-type true scoring key of correct responses consisted of values of one. For five-categorical data, possible response categories consisted of values of one to five. According to the simulation model, a higher value indicated a more correct response option. Hence, for five-categorical data the (1) vector-type true answer key of correct responses consisted of values of five, whereas in the (2) matrix-type scoring key the most correct response option had a value of five, the least correct response option had a value of one, and the values in between represented partly correct responses.

Depending on the scoring method, the type of the reconstructed keys - that is, keys as a result of the respective scoring method - differed. For mode CBM and Consensus Analysis, reconstructed (vector-type) scoring keys consisted of the numerical values of the response options chosen as correct according to the scoring method. Both scoring methods allowed only one correct response option for both two- and five-categorical data. Hence, reconstructed scoring keys consisted of values of one and zero for two-categorical data, whereas reconstructed scoring keys consisted of values of one to five for five-categorical data. However, the numerical values of options only designate the response option chosen as correct according to the method, without predicating any ordinal or metric meaning of numerical value that could be interpreted in terms of the scoring method. Still, the numerical values have their original meaning based on the simulation model. That is, if a method has selected the response option with value one as correct (for two-categorical data), it consequently has chosen the true correct response option, as response options with values of one designate the true correct options according to the simulation model.

For proportion CBM, HOMALS and NRM, the (matrix-type) scoring keys consisted of weights for each response option of each item. For proportion CBM, values were defined as the relative frequency with which the response option was chosen by the sample. Accordingly, higher values indicated more correct responses. For HOMALS, interpretation of the weights was difficult, because the polarity of the component was not defined beforehand (see section 3.2). Hence, positive values may indicate a correct or an incorrect response. For NRM, higher values indicated that response options more closely represented ability and were therefore more correct. This was only true if the choice for fixing response options of lowest and highest rank was good. If chosen randomly, the polarity of the latent variable was unknown, and the interpretation of slope values became obscure. Additionally, in HOMALS and NRM weights were not bound in parameter space. High values were possible, in particular for NRM, indicating a bad choice in fixation (Chalmers, 2012). These matrix-type scoring keys were in some cases (only for two-categorical data) changed to vector-type scoring keys, which only contained the method scoring values for the true correct response options.

For the comparison of true the scoring key with the scoring keys based on CBM, HOMALS, Consensus Analysis and NRM, the Euclidian distance between each pair of original and reconstructed scoring keys was calculated as follows. In the case of two-categorical data and for mode CBM, as well as Consensus Analysis, the distance was calculated between the vector of correct responses according to the simulation model - that is, a vector of ones - and the vector that contained the correct answers according to the method. For proportion CBM and HOMALS, the distance calculated was based on the vector of values of one and the scoring values that were assigned to the originally correct response options. Thus, if all answers that were correct by model definition were also identified as being correct by mode CBM or Consensus Analysis, the Euclidian distance was zero. For proportion CBM and HOMALS, the Euclidian distance was not necessarily zero if the methods successfully identified the correct answer as being correct, because these scoring methods allow values other than zero and one.

In the case of five-categorical data and mode CBM or Consensus Analysis, the distance was again calculated between the correct response options that were true by definition of the model - that is, in this case a vector with values of five - and the values of the response options that were identified as being correct by the scoring methods. Thus, if the original response option five was correctly identified as being correct, the distance equaled zero. If response option four was identified as being correct, the distance as part of the Euclidian distance was one, and so forth. Thus the minimum Euclidian distance here was again zero, but in general values were expected to be higher, since greater distances were possible. For proportion CBM, HOMALS and NRM, Euclidian distance was calculated between vectors of values of the original level of correctness (with values of one, two, and so forth) and vectors of the values that were assigned to the corresponding response options. Because Euclidian distance here was based on five comparisons for each item, Euclidian distances were hypothesized to be highest.

Because Euclidean distances were calculated differently for the scoring techniques, minimum and maximum possible values of distances varied. For mode CBM and Consensus Analysis, minimum Euclidean distance was zero for both two- and five-categorical data. In these cases, the scoring method succeeded in identifying the correct response option for each item. Maximum Euclidean distance varied with number of variables. For two-categorical data the maximum was 3.46 for 12 items, 4.89 for 24 items and 6.93 for 48 items. For five-categorical data maximum values were 13.86 for 12 items, 19.60 for 24 items and 27.71 for 48 items. These values indicated that for all items the (most) incorrect response option was identified as being correct by the CBM or CA method.

Concerning proportion CBM and two-categorical data, minimum Euclidean distance was zero when all respondents chose the correct response option for all presented items. However, this case was very unlikely. Maximum Euclidean distance again varied with the number of respondents. If all respondents chose the most incorrect response option, maximum values were 3.46 for 12 items, 4.89 for 24 items and 6.93 for 48 items. In the case of five-categorical data, the minimum distance was greater than zero. If respondents chose the most correct response option (option 5) for all items, distances greater than zero still resulted, because the possible value space of proportion consensus scores is between zero and one, whereas for the true correctness it is between one and five. Minimum values were in this case 23.49 for 12 items, 33.23 for 24 items and 46.99 for 48 items. Maximum values resulted if all respondents endorsed the most incorrect response option (option 1) in 25.46 for 12 items, 36 for 24 items and 50.91 for 48 items. For HOMALS and NRM, which also provide weights as scores, values of Euclidean distance were expected to be different, because the value spaces of both methods differ. For NRM scoring, value space was not specified, which made it impossible to theoretically define lower and upper bounds for Euclidean distance. For HOMALS scoring, weights were expected to be low positive and negative values, because of the normalizing procedure of quantification and component scores. Thus, for HOMALS and NRM no possible minimum or maximum values are stated.

In the case of five-categorical data and matrix-type scoring keys (Proportion CBM, HOMALS and NRM), Euclidian distance could barely account for ranks. In these cases, Kendall's τb was calculated to indicate the agreement of true and reconstructed ordering of response options. Kendall's τb values of each item were then averaged across all items of the respective data set. High positive values of Kendall's τb indicate that the original rank order of the response options and the reconstructed order of response options converged, whereas high negative values indicate that ranking was inverted.

Analysis on the Level of the Scored Data. On the second level, for each scoring method, the scored data was analyzed in two major steps. First, the one-dimensional structure of the scored data was tested with Confirmatory Factor Analysis (CFA). Second, the scoring methods ability estimates were compared with the true abilities of respondents.

For the confirmation of the hypothesized structure, one-dimensional CFAs with two different estimators were conducted. For some data sets the CFA could not be conducted due to a zero variance of one of more variables in the data set. These cases were counted. The different estimators were needed because data scored with proportion CBM, HOMALS and NRM did not provide integers, but decimal numbers. Thus, CFAs for these data sets could not use the weighted least squares means and variance adjusted (WLSMV) estimator as used for mode CBM and Consensus Analysis scored data. For proportion CBM, HOMALS and NRM, the ML estimator was used. In general, CFA was only conducted with sample sizes of $N = 50$ or higher. Although a sample size of $N = 50$ or even $N = 100$ might be too small, especially with higher numbers of variables, CFAs were still conducted in these cases in order to reflect (possible, but bad-choice) reality. When CFA could not be successfully conducted, the software output file was stored to identify the reason for estimation failure. For all CFAs, important information about model fit was stored, namely, the value of the χ^2-test statistic, the respective p-value and the degrees of freedom of the χ^2-test, the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI). For the χ^2-test statistic, the RMSEA, and the CFI, mean and variance over all $10,000$ iterations were stored. For p-values, the p-values smaller than the chosen alpha level of $\alpha = .05$ were counted. The CFA results are mainly evaluated based on the most common standards for fit indices: According to Hu and Bentler (1999), CFI values of .96 or greater, as well as RMSEA values of .06 or smaller, indicate an at least satisfactory fit.

In order to compare different scoring methods on the level of the scored data, the Pearson correlation of true abilities and reconstructed abilities was calculated. This dependent variable allowed one to judge the relative proximity of true and observed abilities. A (high) negative correlation indicates that the ability estimates based on the scoring method identified respondents as able who are in fact not able. A (high) positive correlation indicates that highly able respondents were identified as being able, and a zero correlation indicates that true abilities and ability estimates shared nothing in common. Reconstructed abilities were either a sum score of the scored data (CBM and Consensus

Analysis) or an ability estimate provided by the scoring method (NRM and HOMALS). For mode CBM and Consensus Analysis, the number of correctly answered items (according to the method) was used as an ability estimate. The sum of scoring weights of the endorsed response options served as the ability estimate for proportion CBM. For HOMALS, the component scores, and for NRM, factor scores (person parameter estimates) provided ability estimates.

5.1.3 Programming of Simulation Studies. The present simulation studies were programmed and implemented using R versions 2.14 till 3.1.3 (R Core Team, 2013). The following subsection gives a short overview about the programming structure of the simulation studies, referring to the used software and R-packages. The simulation studies can be found in the digital supplemental material.

For each condition, the parameters described in the previous subsection were used to calculate the probability of endorsing each response option of each fictional respondent and each fictional item. The calculated probabilities were used as an input for either a binomial random number generator (two-categorical data) or a multinomial random number generator (five-categorical data). Random numbers were generated with the Mersenne-Twister random number generator (Matsumoto & Nishimura, 1998). Seeds for the random number generator were stored and can be used to re-simulate data. For most parameters, values were fixed by design, as described in the paragraph 5.1.2.1. Only for ability of respondents were parameters drawn from a normal distribution, and thus not individually specified before simulation. The values of ability can be re-simulated using the stored seeds. The simulation codes were programmed by the author of this dissertation. For calculation reasons, the psych package (Revelle, 2013) was used to add matrices. Apart from that, only R base packages were used for simulating data.

Manipulation and analysis of data were conducted for each data set individually. Firstly, each simulated data set was scored with the different scoring methods. Secondly, the dependent variables were calculated. For the scoring of data sets, two different

R-packages were used. For HOMALS scoring of the data, the function homals of the homals package was used (De Leeuw & Mair, 2009). For the NRM, the mirt package was used (Chalmers, 2012). CBM as well as CA scoring scripts were programmed by the author and are available in the digital appendix.

For data analysis, the following R-packages and software were used: For test of one-dimensionality Mplus (Muthén & Muthén, 1998-2010) was used through R with the MplusAutomation package (Hallquist & Wiley, 2013). For calculation of Kendall's τb, the R packages Kendall was used (McLeod, 2011). Euclidean distance as well as correlation of the true abilities and ability estimates were calculated using R base packages. The results for each data set were stored intermediately and the simulated data was deleted immediately after analysis. After simulating, manipulating and analyzing $10,000$ data sets for each condition, summary statistics for each dependent variable were calculated and stored.

5.2 Results

As described in the methods section, several dependent variables were used to evaluate the influence of independent variables on the performance of scoring methods. Results of the analyses of dependent variables are described later in this chapter following a specific structure. In the first part, results are outlined for each scoring method individually. Firstly, the true scoring key is compared to the reconstructed scoring key for each scoring method using distance or proximity measures of the two scoring key vectors. For each scoring method, measures of distance or proximity differ slightly (see paragraph 5.1.2.2). Due to the resulting different metrics of the dependent variables, results are used to individually evaluate the influences of data characteristics on the performance of methods without intending to compare them. However, whenever possible results of the different scoring procedures are compared. Secondly, data was scored with the resultant scoring key of each method and structure analysis was conducted. CFA was used to test whether the one-dimensional structure remains stable after scoring. In the second part of results, the correlation of true abilities of respondents is compared to the re-estimated abilities after scoring. This dependent variable allows comparison of different scoring methods, as the metric of the measure is the same for all methods. Hence, this dependent variable is used to directly judge performance of methods in comparison to other methods, as well as to infer conclusions and recommendations for use of this scoring methods. Results of dependent variables are presented simultaneously for two- and five-categorical data in the following order: 1) mode CBM, 2) proportion CBM, 3) Consensus Analysis 4) HOMALS, and 5) NRM.

Prior to the presentation of results, the structure of the analysis is described to ensure that the results are readily comprehensible. As described in the methods section, for every condition of the simulation studies, means and standard deviations of the distribution of 10 000 values of the dependent variables were stored. Based on the stored summary statistics for each condition, further analytic steps were conducted. Firstly, results for the dependent variables included descriptives statistics (arithmetic mean, standard deviation, minimum and maximum) across all 270 conditions of both the dependent variables means over all iterations and the standard deviation over all iterations.

Secondly, consistent with analytic tradition for experimental research, descriptive analysis of variance was used to highlight which independent variables account for variance in the dependent variable. Hence, a $3 \times 5 \times 3 \times 6$ (Number of variables $[12, 24, 48]$ × number of respondents $[20, 50, 100, 200, 1000]$ × Ability [low, medium, high] × Difficulty [[e,e,e], [e,e,m], [m,m,m], [e,m,d], [d,d,e], [d,d,d]]) analysis of variance (ANOVA) was conducted. However, a reduced factorial design was used for Consensus Analysis, as the factor level number of variables equal to 1000 was omitted and used for NRM, as conditions with number of respondents smaller than 200 were excluded. In addition, the factorial design had to be adapted for HOMALS. Here, the number of iterations in the conditions differed, because the algorithm could not be used when one variable of the simulated data set had a zero variance. Due to differing sample sizes in conditions, the balanced ANOVA was not appropriate here. However, different sample sizes only occurred for two factor levels of the factor sample size, which is why these factor levels were excluded for ANOVA. ANOVAs are only presented for the dependent variables Euclidean distance, Kendall's τb and Pearson correlation of true and reconstructed scoring keys, because these are the most important results with regard to the hypotheses.

To give an impression of the relative amount of variance accounted for by the single effects, the effect size η^2 was calculated. That is, the sum of squares of the single effects was divided by the total sum of squares. This effect size is evaluated by data-driven standards which have been established based not on common standards, but on the results of all present simulation studies. Across the different dependent variables, a total of 270 different effect sizes for first-order interactions and main effects was calculated. 63% (170) of these effect sizes were smaller than $\eta^2 = .01$. The .25-, .50- and .75-fractiles of the distribution of the remaining effect sizes equal to or greater than $\eta^2 = .01$ were used to classify effect sizes into *small* ($\eta^2 \geq .02$ and $\eta^2 < .07$), *medium* ($\eta^2 \geq .07$ and $\eta^2 < .24$), and *large* ($\eta^2 \geq .24$) effects. For convenience, only main effects and first order interactions were computed. If main and first order interaction effects do not account for at least 95% of the variance in the dependent variable, this will be explicitly mentioned. Graphical presentation of main and first order interactions effects allows identification of the direction of effects. The most important graphics are included in this chapter, whereas the majority of figures are presented in the appendices.[3] Graphical presentation of main effects include the mean values as bars and the standard deviation as arrows.[4] In interaction plots, the means of each factor level combination are presented as points linked by lines.

[3] Graphical presentation of interaction effects is limited to the effects which have effect sizes of $\eta^2 \leq .01$. When no effect sizes were calculated or effect sizes were based on a reduced design (because of varying samples sizes in conditions), all interaction effects are presented.

[4] Arrows do not represent standard errors or confidence intervals, as the presentation of results is not aligned with statistical inference.

Graphical presentation as well as text presentation of results often includes the means (and standard deviations) of the values on each level of one or two independent variables. Note that the means of the levels of the independent variable were collapsed over all other independent variables and their levels. For example, for the independent variable number of variables the mean of the Euclidian distance for the level $nvar = 12$ is $M = 2.03$. This marginal mean is calculated across all other levels of independent variables such as number of respondents, ability and difficulty. The graphical and text presentation in the case of differing numbers of iterations (i.e., for HOMALS) is based on non-weighted means and standard deviations.

Note that the results section is aligned to present results in a condensed way highlighting most important results in graphical form, whereas less important results are only indicated in the text and are presented in graphical form in the appendix. The presented results are selected to provide the reader with enough information to adequately judge performance of scoring methods without presenting an overwhelming amount of information.

5.2.1 Individual Analysis of Scoring Methods. For each method the individual analysis is structured as follows: The first part focuses on the comparison of the reconstructed and the true scoring keys. In the second part, the results of CFAs are presented for the scored data sets.

5.2.1.1 Mode Consensus Based Measurement.

Euclidian Distance. The Euclidean distance between the true and reconstructed answer keys was on average $M = 3.01$ ($SD = 2.06$) across all 270 conditions for mode CBM scoring of dichotomous data. The Euclidean distance ranged from a minimum of zero to a maximum of 6.93 between conditions. The standard deviation of the Euclidean distance within conditions had a total mean of $M = .33$ ($SD = .30$). The amount of variation ranged from a minimum of zero to a maximum of .95 for different conditions. For five-categorical data, the Euclidean distance ranged from a minimum of zero to maximum of 27.71 between conditions with a total mean of $M = 11.31$ ($SD = 7.98$). The standard deviation within condition was on average $M = 1.26$ ($SD = 1.18$), varying from zero to a maximum of 4.31 for different conditions.

Table 8

Descriptive Statistics for Euclidian Distance for Factor Levels of the Independent Variable Number of Variables (Mode CBM)

Data	Number of variables	*M*	*SD*	Minimum	Maximum
two-categorical	12	2.03	1.24	0	3.46
	24	2.90	1.74	0	4.90
	48	4.11	2.45	0	6.93
five-categorical	12	7.66	4.83	0	13.86
	24	10.87	6.80	0	19.60
	48	15.40	9.60	0	27.71

Note. The descriptive statistics are collapsed over all factor levels of the other independent variables.

Because Euclidean distances values vary with number of variables, Table 8 presents the descriptive statistics for each factor level of the independent variable number of items for both two- and five-categorical data. The possible minimum and maximum values of Euclidean distance (see paragraph 5.1.2.2) were reached on each respective factor level. Hence, the study included conditions where mode CBM fully succeeded as well as conditions in which the key was completely incorrectly specified. Here, this interpretation is equivalent to the following: There are conditions in which the majority of respondents endorse the correct responses for all items, and there are conditions in which the majority of respondents endorse the incorrect response option for all items.

Table 9

Four-way ANOVA Results for Euclidian Distance (Mode CBM)

	2-categorical data			5-categorical data		
Source	SS	*df*	η^2	SS	*df*	η^2
M: No. of variables	1953489.00	2	.16	27257230.00	2	.15
N: No. of respondents	2790.72	4	.00	28096.05	4	.00
D: Difficulty	3351898.00	5	.28	52539850.00	5	.29
A: Ability	4296799.00	2	.36	68777540.00	2	.38
M × N	473.14	8	.00	3570.26	8	.00
M × D	256529.60	10	.02	4042300.00	10	.02
M × A	323223.10	4	.03	5245057.00	4	.03
N × D	4014.07	20	.00	74244.73	20	.00
N × A	2688.52	8	.00	65942.71	8	.00
D × A	1082220.00	10	.09	11971220.00	10	.07
Within	535300.4		.05	8001692.00		.04
Total	11907520.00			179124400.00		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions.

ANOVA sum of squares and effect sizes for main effects and first order interactions are presented in Table 9. As the results indicate, the variation of the independent variables number of variables, difficulty, and ability had substantial effects on the distance between scoring keys. Ability and difficulty both have high effect sizes, explaining 36% - 38% and 28% - 29% of the variance. The factor number of variables ($\eta^2 = .16/\eta^2 = .15$) had a medium effect of substantially lower size.

The directions of variation of Euclidean distance for the manipulated factor levels are presented in Figure 6 for dichotomous data. As the main effects are very similar for

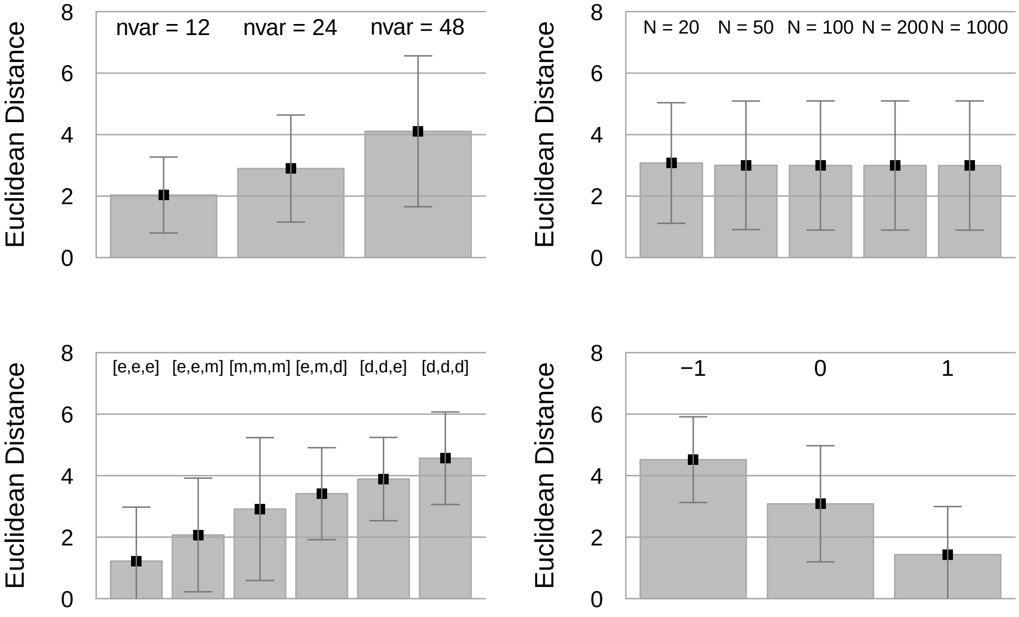


Figure 6. Two-categorical data: Main effects of independent variables on Euclidian distance (Mode CBM). Top left panel: Number of variables; top right panel: Number of respondents; bottom left panel: Difficulty; bottom right panel: Ability.

five-categorical and two-categorical data, the graphical presentation of results is only aligned to two-categorical data, however, the absolute values of Euclidean distance are higher for five-categorical data (see Figure 23 in Appendix A). For number of variables, the reason for the increasing Euclidian distance is trivial, as it grows with the number of variables used to calculate it. As presented in Figure 6, the distance between the true and reconstructed answer keys increases with the difficulty of the items. With low difficulty, the average euclidian distance was $M = 1.22$ for two-categorical data and $M = 4.26$ for five-categorical data, whereas with high difficulty the average was $M = 4.57$ and $M = 17.50$, respectively. As with difficulty, the distance declined with increasing ability: The average distance was $M = 4.52$ for dichotomous data and $M = 17.32$ for five-categorical data with low ability, whereas with high ability it was $M = 1.43$ and $M = 4.97$, respectively.

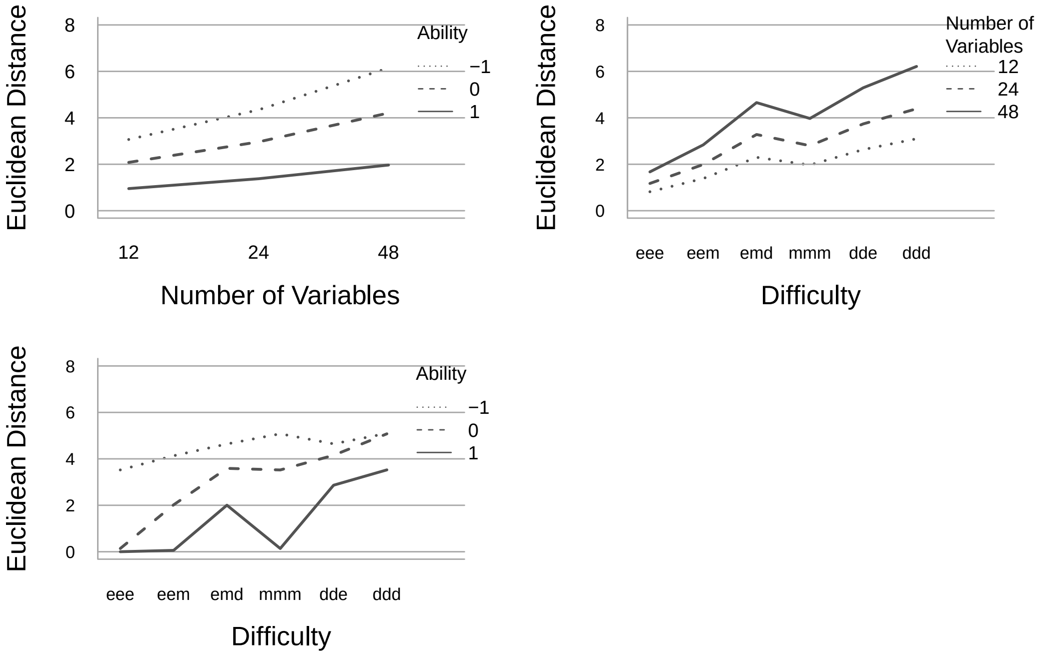


Figure 7. Two-categorical data: Interaction effects of independent variables on Euclidean distance (Mode CBM). Top left panel: Interaction of number of variables and ability; top right panel: Interaction of number of variables and difficulty; bottom left panel: Interaction of ability and difficulty.

In addition to the main effects, some independent variables were observed to interact. The interaction of ability and difficulty ($\eta^2 = .09/\eta^2 = .07$) was of medium effect size. Small effects occurred for interactions of number of variables and difficulty ($\eta^2 = .02/\eta^2 = .02$) and number of variables and ability ($\eta^2 = .03/\eta^2 = .03$). The interaction effects are presented in Figure 7 for dichotomous data (see Figure 23 in Appendix A for five-categorical data). With high ability and easy as well as medium difficult items (i.e., difficulty combinations [e,e,e], [e,e,m] and [m,m,m]), Euclidian distance approached zero, whereas with difficulty combination [e,m,d], [d,d,e] and [d,d,d] the distance was substantially different from zero. For medium ability, distance was zero for only easy items and it increased quite steadily with combinations of increasing difficulty. For low ability the distance was highest, however, the differences in the distances for each difficulty level were comparably smaller. Interaction effects of number of variables with both ability and difficulty can be explained by a characteristic of the Euclidean distance measure. As larger distances are taken into greater account by Euclidean distance, the distance increased more steeply for low abilities than it did for high abilities. Comparably, higher differences in Euclidean distance between the different levels of number of variables were observed for data sets with difficult items compared to data sets with easier items because of taking squares of the differences (see Figure 7).

Confirmatory Factor Analyses. CFAs of scored data were successfully conducted in almost every run. Results are based on a mean of $M = 9996.15$ ($SD = 14.71$) WLSMV CFAs for two-categorical data and $M = 9999.90$ ($SD = .56$) WLSMV CFAs for five-categorical data. For two-categorical data, the confirmation of the hypothesized structure was not successfully conducted in 823 runs (all for factor level $N = 50$) due to zero variance of at least one variable in the data set. The number of exclusions due to zero variance for two-categorical data varied with factor levels of the independent variables ability and difficulty. That is, with high and low ability and with easier and more difficult items, zero variances were observed more often compared to medium ability and medium difficulty combinations. Eight additional CFAs did not succeed because the software package reported error messages.[5] For five-categorical data, 21 CFAs could again not be conducted because of similar error and warning messages. However, no data sets had to be excluded due to zero variance of a variable for five-categorical data.

Table 10 contains descriptive statistics for CFA results across all 216 conditions. The maximum RMSEA values indicate excellent fit, as they do not exceed the cut-off values of .05 in any condition, whereas CFI values and number of p-values of the χ^2-Test that were smaller than the alpha level of $\alpha = .05$ indicate that fit varies between conditions. The CFI ranges from .69 to 1.00 for two-categorical data and from .87 to 1.00 for five-categorical

[5] Error and warning messages addressed non positive definite residual covariance matrices, identification problems, and empty cells in bivariate tables.

data, indicating that at least for some conditions the one-dimensional structure was not supported.

Table 10

Descriptive Statistics for CFA Results (Mode CBM)

	2-categorical data				5-categorical data			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Chi Square			54.26	1174.01			54.16	1767.72
df			54	1080			54	1080
No. p-val[a]	679.53	1125.59	0	6533	1773.37	2786.83	0	10000.00
CFI	.95	.07	.69	1.00	.97	.03	.87	1.00
RMSEA	.02	.01	.00	.05	.02	.01	.00	.05

Note. Estimator: Weighted Least Squares Mean and Variance Adjusted. Min = Minimum; Max = Maximum. [a]Number of p-values p < .05

The fit indices varied with the manipulated data characteristics. The graphical presentation of main effects is displayed in Appendix A. Non-satisfactory fit values (i.e. low CFI and higher RMSEA values) were mainly caused by sample sizes of 100 or lower for two-categorical data. In addition, fit indices varied with other factors: Fit slightly decreased with higher number of variables, however, not according to RMSEA; it was worse for high, as well as low, ability compared to medium ability, and it varied somewhat with difficulty combinations. Here, it was best for the combination [e,e,m], whereas it was less good for all other difficulty combinations.

Although fit was overall acceptable for five-categorical data, the fit also slightly decreased (i.e. lower CFI and higher RMSEA values) for smaller sample sizes. However, number of significant χ^2-tests increased abruptly with sample size of 1000 and was almost equal for all other levels of this factor. In addition, number of significant χ^2-tests was higher for a high number of variables, while CFI and RMSEA were not substantially affected by number of variables. Again, fit was slightly better for medium ability in comparison to high and low ability and it slightly varied with different difficulty combinations for five categorical data. Here, differences between difficulty factor levels were not consistent across the different fit indices.

Overall, sample size seemed to have the most substantial effect on the fit of CFAs, whereas number of variables, ability and difficulty had less influence. With a sample size of at least $N = 200$ for two-categorical data and $N = 100$ for five-categorical data, fit was satisfactory and the one-dimensional structure was supported. The influence of independent variables on the fit indices and the Euclidean distance was not similar, as Euclidean distance was mainly affected by ability and difficulty and was independent of sample size, whereas fit was mainly affected by sample size. The results indicate that the one-dimensional structure remained stable independent of the proximity of scoring keys.

5.2.1.2 Proportion Consensus Based Measurement.

Euclidean Distance. The Euclidian distance ranged from a minimum of .54 to a maximum of 5.87 with a mean of $M = 2.57$ ($SD = 1.22$) for two-categorical data. For five-categorical data, the mean Euclidian distance was $M = 35.76$ ($SD = 10.01$), ranging from a minimum of 23.61 to a maximum of 50.24. As Euclidean distance was expected to vary with levels of the factor number of variables, Table 11 presents the descriptive statistics of Euclidean distance for each of these levels. On the basis of the possible minimum and maximum values of Euclidean distance (see paragraph 5.1.2.2), it can be concluded that proportion CBM succeeded in some condition where distance was close to the possible minimum. However, the scoring method failed in other conditions, where distance was close to the possible maximum.

Variation within conditions ranged from a minimum of .02 to a maximum of .33 with a mean of $M = .11$ ($SD = .07$) for two-categorical data and from a minimum of .01 to a maximum of .23 with a mean of $M = .07$ ($SD = .05$) for five-categorical data. As ANOVA results in Table 12 reveal, within variation accounted for 1% of the variance in the dependent variable for two-categorical data. However, Euclidean distance substantially varied with the manipulation of independent variables for two categorical data. Here, the variation of the factors ability (η^2 = .35), number of variables (η^2 = .34) and difficulty (η^2 = .24) yielded large effect sizes. The interactions of number of variables with both ability and difficulty had very small to small effects. In contrast to mode CBM, no interaction of ability and difficulty was observed.

Table 11

Descriptive Statistics for Euclidian Distance for Factor Levels of the Independent Variable Number of Variables (Proportion CBM)

Data	Number of variables	*M*	*SD*	Minimum	Maximum
two-categorical	12	1.74	.65	0.54	2.94
	24	2.47	.92	.76	4.16
	48	3.49	1.30	1.08	5.87
five-categorical	12	24.30	.42	23.61	25.12
	24	34.37	.59	33.39	35.53
	48	48.61	.84	47.23	50.24

Note. The descriptive statistics are collapsed over all factor levels of the other independent variables.

The direction of main effects for two-categorical data is presented in Figure 8. The Euclidian distance was lowest for high ability ($M = 1.67$), whereas it was substantially higher for low ability ($M = 3.46$). In addition, the distance between scoring keys increased with high difficulty levels, with a mean of $M = 1.65$ for the difficulty combination [e,e,e] and a mean of $M = 3.48$ for the difficulty combination [d,d,d]. Again, number of variables had a trivial main effect on Euclidian distance, increasing with the length of the item vector. The interaction effects for proportion CBM are presented in Appendix B, Figure

Table 12

Four-way ANOVA Results for Euclidian Distance (Proportion CBM)

	2-categorical data			5-categorical data		
Source	SS	*df*	η^2	SS	*df*	η^2
M: No. of variables	1382805.83	2	.34	268404600.00	2	1.00
N: No. of respondents	823.65	4	.00	47.42	4	.00
D: Difficulty	982969.77	5	.24	400271.60	5	.00
A: Ability	1433930.63	2	.35	609255.30	2	.00
M × N	72.07	8	.00	3.45	8	.00
M × D	76323.01	10	.02	31141.03	10	.00
M × A	111324.86	4	.03	47336.23	4	.00
N × D	59.09	20	.00	.13	20	.00
N × A	57.16	8	.00	.17	8	.00
D × A	8610.88	10	.00	7341.01	10	.00
Within	47095.37		.01	18102.36		.00
Total	4044365.69			269491800.00		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions.

33. Interactions for number of variables with both ability and difficulty are very similar to those for mode CBM.

For five-categorical data the effect sizes for Euclidean distances are not meaningful. Because Euclidean distance was calculated in this case on the basis of differences between the scoring values of the five response options, number of variables yielded a huge effect, which masked other effects (see Table 12). Therefore, for five-categorical data, the dependent variable Kendall's τb is analyzed in the following, as it is expected to provide a better measure of proximity between the true and reconstructed scoring key.

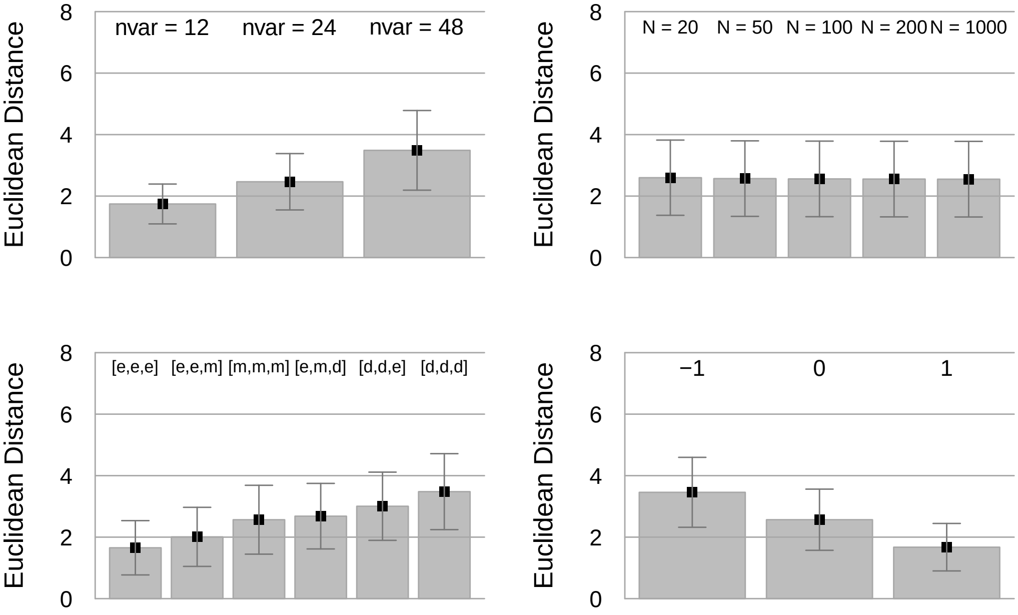


Figure 8. Two-categorical data: Main effects of independent variables on Euclidian distance (Proportion CBM). Top left panel: Number of variables; top right panel: Number of respondents; bottom left panel: Difficulty; bottom right panel: Ability.

Kendall's τb. For five-categorical data, Kendall's τb ranged from $\tau = -1$ to $\tau = 1$ with a mean of $M = .04$ ($SD = .66$). Thus, over all conditions the true and estimated scoring keys based on proportion CBM did not share anything in common. However, minimum and maximum values indicate that proximity was dependent upon the varying data characteristics. At least in some conditions, the reconstructed and true scoring keys fully converged. Variation within conditions ranged from a minimum of $\tau = .00$ to a maximum of $\tau = .29$ ($M = .11$, $SD = .09$) and accounted for 4% of total variation of the dependent variable (see Table 13).

The proximity of true and estimated scoring keys was highly influenced by the independent variables ability ($\eta^2 = .52$) and difficulty ($\eta^2 = .36$). In addition, a medium effect was observed for the interaction of ability and difficulty, accounting for 7% of the variance. As presented in Figure 9, Kendall's τb was high and positive for easy items

($\tau = .62$), whereas it was high and negative for difficult items ($\tau = -.62$). For high ability, the mean Kendall's τb was $\tau = .62$, whereas for low ability it was $\tau = -.57$.

Table 13

Four-way ANOVA Results for Kendall's τb (Proportion CBM)

	5-categorical data		
Source	SS	*df*	η^2
M: No. of variables	.00	2	.00
N: No. of respondents	13.24	4	.00
D: Difficulty	431896.67	5	.36
A: Ability	628941.78	2	.52
M × N	.08	8	.00
M × D	.11	10	.00
M × A	.02	4	.00
N × D	2548.35	20	.00
N × A	3497.54	8	.00
D × A	90465.28	10	.07
Within	53743.58		.04
Total	1213008.11		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions.

The interaction effect of difficulty and ability is plotted in Appendix B Figure 34. With high ability and difficulty combinations [e,e,e], [e,e,m] and [m,m,m], the rank orders of true and reconstructed scoring keys were very close, whereas for difficulty combinations of [e,m,d], [d,d,e], and [d,d,d] the correlation dropped. With medium ability, Kendall's τb was only high for easy items. Here, it almost continuously decreased with combinations of increasing difficulty. For low ability, Kendall's τb never exceeded a value of zero and was negative for most difficulty combinations.

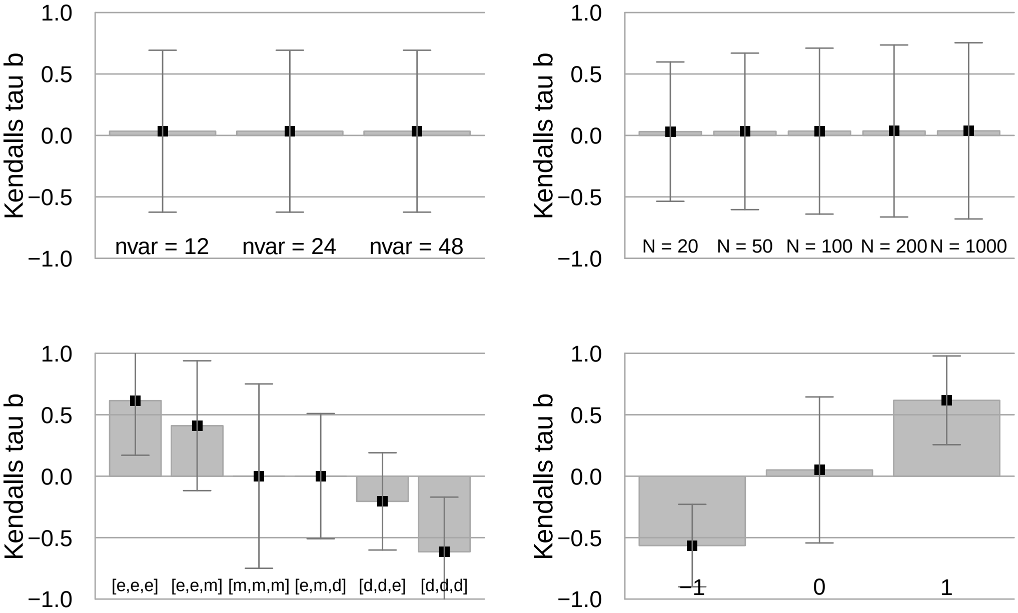


Figure 9. Five-categorical data: Main effects of independent variables on Kendall's τb (Proportion CBM). Top left panel: Number of variables; top right panel: Number of respondents; bottom left panel: Difficulty; bottom right panel: Ability.

Confirmatory Factor Analyses. One-dimensional CFA (ML) was conducted for every run of the 216 conditions. For data scored with proportion CBM, the software package reported problems with CFA estimation. The reasons for the failure of CFAs were mainly convergence or non-identification problems, which indicated that the choice of estimator might have been problematic.

For two-categorical data, a mean of $M = 5641.11$ ($SD = 4139.47$) factor analyses per condition were successfully conducted. The number of successful CFAs varied with factor levels of the independent variables (see Figure 35 in Appendix B). For five conditions no

CFA could be conducted (all for $N = 1000$), and for 103 conditions 5000 or fewer CFAs were successfully estimated, whereas for 50 conditions no problems were observed at all. With a lower number of variables, more CFAs were successfully calculated. However, with growing sample size, as well as with sample size of $N = 50$, fewer CFAs converged. Most CFAs converged with equal item difficulties across all items, whereas with mixed item difficulties the number of successful CFAs decreased. Most CFAs converged with medium ability, followed by high ability, whereas the fewest CFAs converged with low ability. In addition, interactions of ability and difficulty, difficulty and number of variables, as well as difficulty and number of respondents, had an effect on the number of successful CFAs. These interaction effects are presented graphically in Appendix B, Figure 36 and 37.

For five-categorical data a mean of $M = 8625.39$ ($SD = 2909.53$) CFAs was successfully conducted for every condition. For one condition only 78 CFA could be conducted and 30 conditions provided 5000 or fewer successful CFAs, whereas for 94 conditions a maximum of 10000 CFAs were successful. Variation with levels of independent variables was less distinctive in the case of five-categorical data (see Figure 38 in Appendix B). Nevertheless, a great drop in the number of successful CFA occurred with sample size of $N = 1000$. Thus, for $N = 50$, $N = 100$ and $N = 200$, on average almost 10,000 CFA were successful, whereas for $N = 1000$ on average only $M = 5340.50$ ($SD = 4304.15$) CFAs were successful. For the factor level $N = 1000$ the variation among the factor levels of the other independent variables appeared to be more pronounced (see Figure 39 in Appendix B). Thus, for $N = 1000$ the number of converged CFAs decreased with the number of variables, was lower for items with mixed difficulty and was lower for low ability. Moreover, an interaction of ability and difficulty was observed (Figure 40 in Appendix B).

Unfortunately, estimation problems indicate that results of the CFAs might not be as meaningful as hoped. The ML estimator was most likely not a good choice and does not work under these data conditions, especially for not normal data. Although this assumption is just a potential post-hoc explanation, the results for the confirmation of the one-dimensional structure of proportion CBM scored data should be interpreted with caution.

Across all conditions, fit indices were nonsatisfactory for both two- and five categorical data (Table 14). Although mean RMSEA values were satisfactory, mean CFI values were not. In addition, minimum values of CFI and maximum values of RMSEA indicated bad fit in some conditions, whereas in other conditions fit appeared to be good. The graphical presentation of main effects is presented in Appendix B. For two-categorical as well as five-categorical data, sample size was observed to have a substantial influence: CFI values dropped dramatically and RMSEA increased steeply for small sample sizes, whereas fit indices were distinctly better for $N = 1000$. In addition, fit was worse for a higher number of variables. Smaller variations were observed for ability and difficulty: Fit was slightly better for medium ability and worst for difficulty combinations [e,e,e] and [m,m,m]. However, differences were not as pronounced as they were for the other independent variables, in particular for five-categorical data.

Table 14

Descriptive Statistics for CFA Results (Proportion CBM)

	2-categorical data				5-categorical data			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Chi Square			38.43	3085.56			55.04	2610.23
df			54	1080			54	1080
No. p-val[a]	3531	3746.45	0	10000	5836.37	3562.00	24	10000
CFI	.79	.24	.11	1	.85	.19	.17	1.00
RMSEA	.05	.04	.00	.19	.05	.04	.00	.17

Note. Estimator: Maximum Likelihood. Min = Minimum; Max = Maximum. [a]Number of p-values p < .05.

To summarize, the one-dimensional structure was not supported after scoring. However, this might be caused by the choice of estimator more than the scoring method itself. Moreover, the model fit did not vary with independent variables in a way similar to Euclidean distance, but was more dependent upon sample size and number of variables.

5.2.1.3 Consensus Analysis.

Euclidean Distance. The total mean Euclidean distance between the true scoring key and the scoring key resulting from Consensus Analysis was $M = 2.59$ ($SD = 2.32$), ranging from a minimum of zero to a maximum of 6.93 for two-categorical data. For five-categorical data, this dependent variable ranged from a minimum of zero to a maximum of 27.71 with a total mean of $M = 10.59$ ($SD = 9.20$).

Table 15

Descriptive Statistics for Euclidian Distance for Factor Levels of the Independent Variable Number of Variables (Consensus Analysis)

Data	Number of variables	*M*	*SD*	Minimum	Maximum
two-categorical	12	1.76	1.45	0	3.46
	24	2.49	2.05	0	4.90
	48	3.52	2.91	0	6.93
five-categorical	12	7.12	5.74	0	13.86
	24	10.17	8.10	0	19.60
	48	14.48	11.40	0	27.71

Note. The descriptive statistics are collapsed over all factor levels of the other independent variables.

Within conditions, mean variation was $M = .69$ ($SD = 1.00$) with a range from zero to 3.46 for two-categorical data and $M = 2.29$ ($SD = 3.47$) ranging from zero to 13.77 for five-categorical data. Table 15 presents the descriptive statistics for each factor level of the dependent variable number of variables for both two- and five-categorical data. Again, the possible minimum as well as maximum values (see paragraph 5.1.2.2) for the Euclidean distance were observed in some conditions. Hence, the scoring method performed very well in some experimental conditions, whereas in other conditions scoring keys did not converge at all.

Table 16

Four-way ANOVA Results for Euclidian Distance (Consensus Analysis)

	2-categorical data			5-categorical data		
Source	SS	*df*	η^2	SS	*df*	η^2
M: No. of variables	1121732.69	2	.08	19676059.63	2	.09
N: No. of respondents	17540.77	3	.00	101215.00	3	.00
D: Difficulty	2537545.42	5	.17	41521697.21	5	.19
A: Ability	5840602.22	2	.39	88929858.31	2	.41
M × N	1589.39	6	.00	18099.20	6	.00
M × D	196638.52	10	.01	3361844.13	10	.02
M × A	464056.74	4	.03	6651046.16	4	.03
N × D	7859.99	15	.00	54746.54	15	.00
N × A	19293.95	6	.00	268494.40	6	.00
D × A	1265027.72	10	.09	19163983.76	10	.09
Within	3177769.58		.21	37240008.92		.17
Total	14794711.97			219217767.98		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions.

Results of analysis of variance in Table 16 reveal that Euclidean distance varied mostly with ability of respondents ($\eta^2 = .39$ and $\eta^2 = .41$) for both two- and five-categorical data. In addition, difficulty of items ($\eta^2 = .17$ and $\eta^2 = .19$), number of items ($\eta^2 = .08$ η^2

= .09) and interaction of difficulty and ability (η^2 = .09 and η^2 = .09) yielded medium effects and accounted for variation in the dependent variable. Very small and small effects were also present for interactions of number of variables with both ability and difficulty.

The direction of effects is identical for two- and five-categorical data, which is why results are presented graphically only for five-categorical data. Main and interaction effects for two-categorical data are presented in Appendix C, Figures 49 and 50. As the graphical presentation of main effects in Figure 10 shows, Euclidean distance increased with number of variables, was not (mainly) affected by number of respondents, increased with difficulty and decreased with higher ability. That is, distance was comparably low for difficulty combination [e,e,e] with a mean of $M = 3.74$, whereas it was somewhat higher for difficulty combination [d,d,d] ($M = 17.13$) for five-categorical data. Similarly, for high ability mean Euclidean distance was lower ($M = 3.65$) than for low ability ($M = 18.36$).

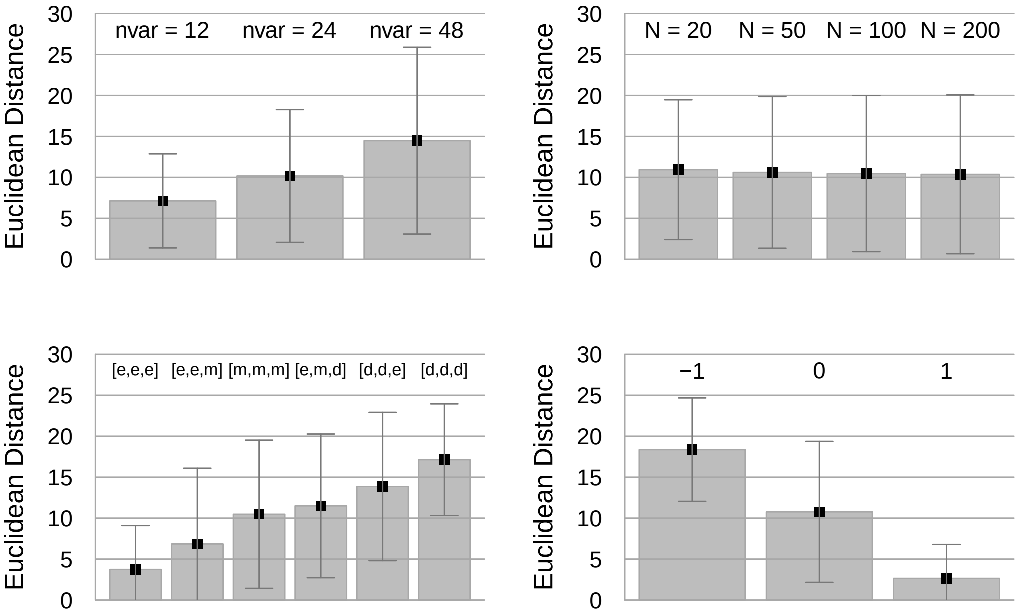


Figure 10. Five-categorical data: Main effects of independent variables on Euclidean distance (Consensus Analysis). Top left panel: Number of variables; top right panel: Number of respondents; bottom left panel: Difficulty; bottom right panel: Ability.

The interaction effects of ability and difficulty (Figure 11) were pronounced. With high ability, Euclidean distance was low for easy and medium difficulty combinations, slightly increased with difficulty combination [d,d,e] and strongly increased with only difficult items. For medium ability, only easy difficulty combinations (i.e., [e,e,e] and [e,e,m]) resulted in low Euclidean distance. For low ability, however, Euclidean distance never approached zero, although it was lower for difficulty combination [e,e,e] compared to the other difficulty combinations. The interaction effects observed for number of variables with both ability and difficulty were again caused by the way Euclidean distance is calculated: Because of squaring, with increasing number of variables, values increase more steeply when the distances are greater.

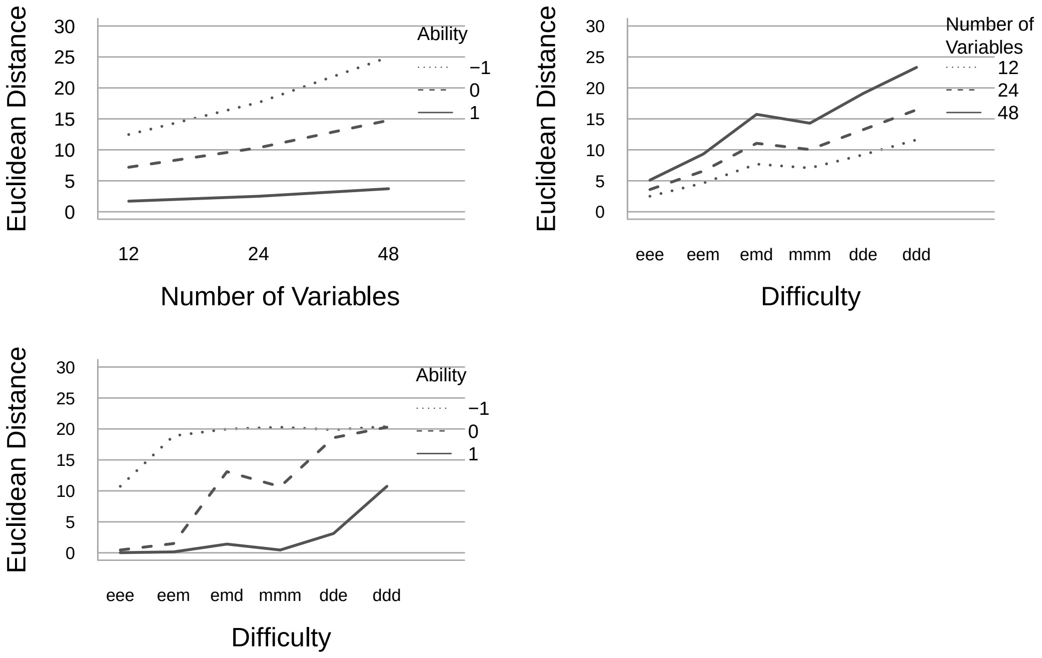


Figure 11. Five-categorical data: Interaction effect of independent variables (Consensus Analysis). Top left panel: Interaction of number of variables and ability; top right panel: Interaction of number of variables and difficulty; bottom left panel: Interaction of ability and difficulty.

Confirmatory Factor Analyses. After the application of Consensus Analysis as a scoring method to score the simulated data, some items showed zero variances; hence these data sets could not be used for the confirmation of the one-dimensional structure. For two-categorical data, the mean number of exclusions due to zero variance was $M = 5.08$ ($SD = 16.79$), ranging from a minimum of 0 to a maximum of 115 between conditions. A total of 823 data sets had to be excluded because of zero variance after scoring. These data sets were the same ones that were excluded for CFAs of mode CBM scored data because of zero variance (see paragraph 5.2.1.1). The data sets had zero variances before mode CBM or Consensus Analysis were used, which was apparently most likely for low sample size and certain combinations of difficulty and ability.

For five-categorical data, although no zero variances before scoring were observed, zero variances were observed after application of Consensus Analysis[6]. The number of excluded data sets was fairly high, with a mean of $M = 277.67$ ($SD = 1038.91$), ranging from 0 to 6854 between conditions. Here, a total of 44982 data sets had to be excluded because of zero variance after scoring. Zero variances after scoring could only occur if these response options - which were not chosen by any respondent (or by every respondent, which did not occur as zero variance before scoring was tested) - were selected as correct. Obviously, the algorithm would fail in this case. This point will be readdressed in the discussion section. The number of exclusions because of zero variance varied with the manipulated data characteristics for five-categorical data. Figure 51 in Appendix C present the number of exclusions after scoring for each factor level. The number of exclusions increased with increasing number of variables, and decreased with increasing sample size. In addition, observations of zero variance were dependent upon item difficulty and ability. The number of exclusions was very low for difficulty combinations [e,e,e], [m,m,m], and [d,d,d], that is, for combinations with uniformly distributed difficulty. However, for mixed

[6] This was only registered for $N > 20$, as CFAs were not conducted for data sets with very low sample size.

difficulty combinations, the number of excluded data sets was fairly high. In addition, the number of exclusions slightly increased with higher ability. Moreover, the number of exclusions was highest for the combination low ability and [e,e,m], medium ability and [e,m,d], as well as high ability and [d,d,e] (see Figure 53 in Appendix C). However, for the remaining data sets most CFAs were successfully conducted. Only for two-categorical data, nine CFAs showed difficulties, as the residual covariance matrix was not positive definite. For five-categorical data, no estimation difficulties were observed.

Table 17

Descriptive Statistics for CFA Results (Consensus Analysis)

	2-categorical data				5-categorical data			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Chi Square			55.23	1174.01			54.88	1977.43
df			54	1080			54	1080
No. p-val[a]	1021.90	1339.51	2	6533	994.40	1810.98	0	10000
CFI	.93	.07	.69	.99	.97	.05	.68	1.00
RMSEA	.02	.01	.01	.05	.02	.01	.01	.09

Note. Estimator: Weighted Least Squares Mean and Variance Adjusted. Min = Minimum; Max = Maximum. [a]Number of p-values $p < .05$.

As stated in Table 17, mean results for WLSMV CFAs are satisfactory for five-categorical data, whereas for two-categorical data CFI values indicate non-satisfactory fit, but RMSEA values indicate good fit. Minimum values of CFI reveal for both two- and five-categorical data that fit was non-satisfactory in some conditions. In fact, the fit indices varied with data characteristics, number of respondents having the biggest influence on fit indices for two categorical data (see Appendix C for the graphical presentation). In addition, fit was slightly worse for data sets with a higher number of variables (only with respect to CFI and χ^2-test), better for medium ability compared to high and low ability, and comparably better for difficulty combination [m,m,m] (compared to other difficulty combinations) for two categorical data. The effects of sample size and number of variables were similar, although less pronounced, for five categorical data. However, fit was comparably worse for medium ability compared to high and low ability. Moreover, fit was substantially worse for the difficulty combination [e,m,d]. The effects of ability and difficulty seemed to be more pronounced for five-categorical data compared to two-categorical data.

To summarize, the one-dimensional structure remains relatively stable after scoring, most substantially influenced by sample size and number of variables. Although fit varied with manipulated data characteristics, the pattern of variation was different compared to Euclidean distance. The influence of the independent variables ability and difficulty seems to be somewhat higher compared to CBM methods, however, different for two- and five-categorical data and also different compared to the way both independent variables influenced Euclidean distance. Moreover, for the interpretation of the influence of ability and difficulty, one should bear in mind the CFA drop-out for five-categorical data, which also mainly varied with different factor levels of ability and difficulty.

5.2.1.4 HOMALS. Unlike the other scoring methods, for the application of HOMALS a necessary requirement had to be met: Data sets could only be scored with the algorithm if no variable had a zero variance. Hence, data sets with items with zero variance had to be excluded from scoring. The presentation of results therefore differs in one point from the results of other methods: Different group sample sizes (successful iterations) were possible for all dependent variables, resulting in an unbalanced design.[7]

For two-categorical data, over all conditions a mean of $M = 301.37$ ($SD = 1088.43$) data sets were excluded from analysis, ranging from a minimum of zero to a maximum of

[7] Despite of differing sample sizes in conditions, for the calculation of descriptive statistics (e.g. total mean) the differences in sample size are not taken into account, that is, unweighted between condition means and variances are presented.

6934. With a sample size of $N = 100$ or higher, no item showed zero variance and hence no data set had to be excluded. However with $N = 20$ and $N = 50$ a mean of $M = 1491.61$ and $M = 15.24$ data sets had to be excluded.[8] In addition, the number of excluded data sets varied with levels of other factors. More data sets showed zero variances when number of variables were higher, ability was more extreme (either low or high ability) or difficulty was more extreme. In addition, for five-categorical data variables showed zero variances with sample size $N = 20$. Here, 223 data sets were excluded for HOMALS scoring. Again, exclusions were more pronounced when more variables were simulated, as well as with higher or lower ability or more extreme difficulty combinations.

Euclidean Distance. For two-categorical data, Euclidean distance had a mean of $M = 5.09$ ($SD = 1.42$) ranging from a minimum of 3.40 to a maximum of 6.95. Within variation was rather low, with a mean of $M = .05$ ($SD = .04$) with a range from .01 to .19. For five-categorical data, the mean of Euclidean distance between conditions was $M = 37.78$ ($SD = 10.56$), ranging from a minimum of 25.54 to a maximum of 51.39. Within variation had a mean of $M = .10$ ($SD = .08$) with a range from zero to .30. As can be seen in Figures 62 and 65 in Appendix D, variation of Euclidean distance was, for both two- and five-categorical data, mainly explained by the number of variables. This conclusion is supported by ANOVA results for two- and five-categorical data. Because of varying sample size for factor levels $N = 20$ and $N = 50$ for two-categorical data and $N = 20$ for five-categorical data, these factor levels were excluded from ANOVA analysis. Both the $3 \times 6 \times 3 \times 3$ ANOVA for two-categorical data, as well as the $3 \times 6 \times 3 \times 4$ ANOVA for five-categorical, revealed that 100% of the variance could be explained by the factor number of variables, whereas other independent variables did not yield meaningful effect sizes.

[8] For the factor level $N = 50$, these data sets were the same as those for which no CFAs were possible for mode CBM and Consensus Analysis.

Kendall's τb. For five-categorical data, mean Kendall's τb was $M = -.01$ ($SD = .19$), ranging from a minimum of $-.39$ to a maximum of .40. Within conditions, variation was fairly high with a mean of $M = .83$ ($SD = .19$). The amount of variation ranged between conditions from .29 to 1.00.

Table 18

Four-way ANOVA Results for Kendall's τb (HOMALS)

	5-categorical data		
Source	SS	*df*	η^2
M: No. of variables	2.22	2	.00
N: No. of respondents	2.34	3	.00
D: Difficulty	32114.10	5	.02
A: Ability	43823.52	2	.02
M × N	1.46	4	.00
M × D	11.46	10	.00
M × A	4.20	4	.00
N × D	233.29	10	.00
N × A	277.88	4	.00
D × A	2595.90	8	.00
Within	1733520.27		.96
Total	1813145.78		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions of a reduced design, excluding the factor level $N = 50$.

The influence of manipulated factors was analyzed with the reduced ANOVA design, excluding the factor level $N = 20$. As Table 18 shows, the independent variables ability and difficulty had small effects. However, most of the variance (96%) of Kendall's τb was observed within conditions. The main effects based on the full design are displayed in Figure 12. The independent variables number of variables and sample size did not effect Kendall's τb, whereas difficulty and ability had effects on this dependent variable. Kendall's τb was higher (albeit still of small positive size) for low ability, and was small (again small negative size) for high ability. Regarding the main effect of difficulty, the rank order correlation was highest for difficult items and continuously dropped for easier items. Graphical presentation of interaction effects for the full design (see Figures 68 and 69 in Appendix D) indicate no substantial effect, supporting ANOVA results for the reduced design.

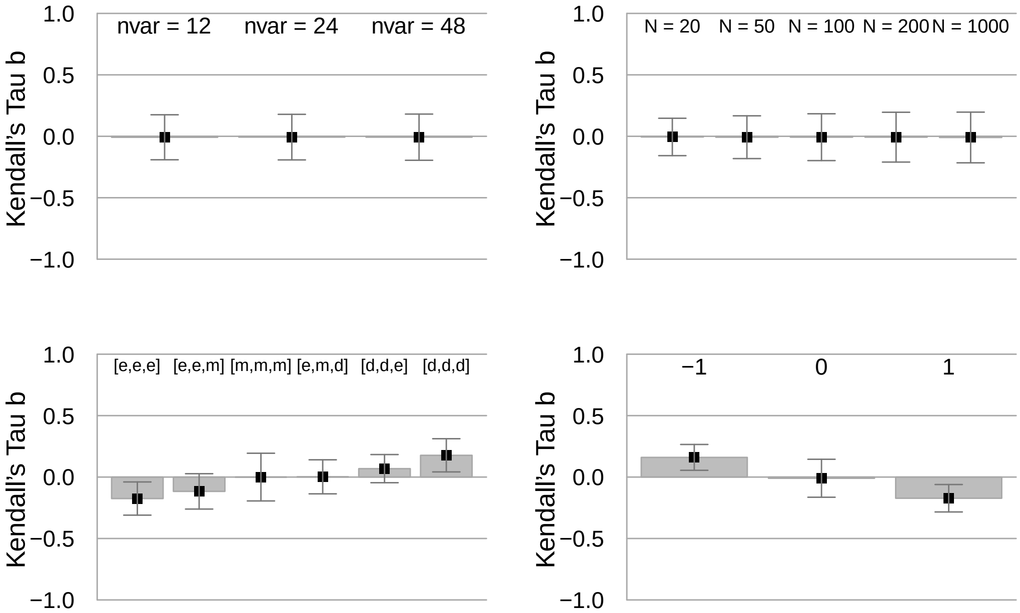


Figure 12. Five-categorical data: Main effects on Kendall's τb (HOMALS). Top left panel: Number of variables; top right panel: Number of respondents; bottom left panel: Difficulty; bottom right panel: Ability.

Confirmatory Factor Analyses. Descriptive statistics of ML CFAs are presented in Table 19. Results of the CFAs indicate a non-satisfactory fit across conditions. The mean value of CFI for one-dimensional factor analysis was $M = .78$ ($SD = .24$) for two-categorical and $M = .86$ ($SD = .16$) for five-categorical data. Moreover, the mean of RMSEA was just satisfactory with $M = .05$ ($SD = .04$) for two-categorical data and just non-satisfactory with $M = .08$ ($SD = .06$) for five-categorical data. Minimum and maximum values of fit indices indicate that in some conditions fit was good, whereas in other condition fit was non-satisfactory. Note, that results are comparably bad for proportion CBM, where the same estimator was used.

Table 19

Descriptive Statistics for CFA Results (HOMALS)

	2-categorical data				5-categorical data			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Chi Square			56.85	3454.10			61.41	4566.33
df			54	1080			54	1080
No. p-val[a]	4925.68	3958.11	0	10000	7046.13	3999.51	0	10000
CFI	.78	.24	.10	.99	.86	.16	.35	1.00
RMSEA	.05	.04	.01	.21	.08	.06	.01	.25

Note. Min = Minimum; Max = Maximum. Estimator: Maximum Likelihood. [a]Number of p-values p < .05.

The manipulated independent variables affected the fit indices of factor analysis. Although fit for five-categorical data was better, the direction of effects was similar for two- and five-categorical data. The independent variable number of respondents had a substantial effect on fit indices, as CFI increased and RMSEA dropped, and the number of significant χ^2-tests decreased with higher sample size. For $N = 50$, fit indices were very bad, whereas for $N = 1000$ fit indices were very good (Figures 70 and 74, Appendix D). In addition, the model fit was influenced by number of variables, as fit indices were better for fewer variables (Figures 71 and 75, Appendix D). A small variation with different ability levels was also observed, as fit indices were slightly better for medium ability compared to high and low ability. In addition, fit was slightly affected by different difficulty combinations. However, effects were different for two- and five-categorical data. Again, although effects of independent variables - most importantly sample size and number of items - were observed, the pattern of effects did not correspond to the effect pattern for Euclidean distance or Kendall's τb. Hence, scoring itself did not influence the structure of the data, but other data characteristics did.

5.2.1.5 Nominal Response Model.

Euclidean Distance. The Euclidean distance between the true and the NRM scoring key varied with the fixation of the highest and lowest response category used for model estimation. As expected, the mean of Euclidean distance between conditions was lowest for NRM V1 with $M = 12.75$ ($SD = 4.21$) varying from a minimum of 7.79 to a maximum of 25.879. In addition, the variation within conditions was lowest with a mean variation of $M = 1.60$ ($SD = 2.46$), ranging from .19 to 10.93. For NRM V1, fixation was based on true ordering of response options. When fixation was based on random ordering of response options (NRM V2), the mean distance was highest with $M = 141.07$ ($SD = 16.73$), ranging from 103.19 to 173.72. Within variation was moderately higher for this fixation method, with a mean variation of $M = 28.08$ ($SD = 7.18$) ranging from 16.85 to 39.74. With the plausible but rather uncertain fixation method (NRM V3), the distance was still high, with a mean of $M = 86.74$ ($SD = 19.50$), ranging from 49.86 to 136.36, but lower compared to NRM V2. Here, within variation was highest with $M = 30.09$ ($SD = 8.68$) varying from 9.95 to 46.15.

Variation of distances can also be explained by the variation of some factors in the design. However, explained variance differs for NRM studies and fixation method (see Table 20). For NRM V1, 33% of the variance is due to within condition variation and 48%

of the variance is explained by the factor number of variables. Here, Euclidean distance increased with more variables, as it did before for the other scoring methods (see Figure 78 in Appendix E). In addition, the remaining variance is explained by other independent variables, in particular, interaction of ability and difficulty ($\eta^2 = .05$) and the main effect of number of respondents ($\eta^2 = .04$). These small effects are also presented graphically in Appendix E. Euclidean distance was slightly higher for $N = 200$ compared to $N = 1000$. The independent variables ability and difficulty were observed to interact in a way that with high ability and easy items (i.e., [e,e,e]), as well as with low ability and difficult items (i.e., [d,d,d]), the Euclidean distance was slightly higher compared to the other factor level combinations.

In NRM V2, the variation in the distance measure was mainly explained by within condition variation ($\eta^2 = .75$). Moreover, the variation in the dependent variable was explained by the interaction of ability and difficulty ($\eta^2 = .15$), the main effect of number of variables ($\eta^2 = .05$) and the main effect of number of respondents ($\eta^2 = .03$). Here, Euclidean distance was higher with more variables as well as with lower sample size (Figure 80 in Appendix E). Moreover, in conditions with equal item difficulties and a mean ability which equals difficulty (e.g., [e,e,e] with low ability), Euclidean distance was higher compared to the other factor level combinations (see Figure 81 in Appendix E).

Again for NRM V3, the variation in the dependent variable was mainly based on error ($\eta^2 = .72$). In addition, number of variables had a substantial influence ($\eta^2 = .20$), whereas number of respondents ($\eta^2 = .03$) and interaction of ability and difficulty ($\eta^2 = .03$) had small impact. The direction of effect was similar to the other NRM studies: Aside from the increase with number of variables, Euclidean distance was lower for $N = 1000$ compared to the factor level $N = 200$ (Figure 82, Appendix E). The interaction effect of ability and difficulty was less pronounced, however the differences between Euclidean distance for each ability level were larger for easy items than they were for medium or difficult items (Figure 83 in Appendix E).

Table 20

Four-way ANOVA Results for Euclidean Distance (NRM)

	NRM V1			NRM V2			NRM V3		
Source	SS	*df*	η^2	SS	*df*	η^2	SS	*df*	η^2
M: No. of variables	13489956.83	2	.48	56258130.00	2	.05	290974166.30	2	.20
N: No. of respondents	1215821.91	1	.04	40458060.00	1	.03	41881559.80	1	.03
D: Difficulty	292547.35	5	.01	6684014.00	5	.01	1186653.20	5	.00
A: Ability	563235.88	2	.02	2695825.00	2	.00	3168160.20	2	.00
M × N	72466.74	2	.00	2688657.00	2	.00	667022.60	2	.00
M × D	39596.52	10	.00	338115.40	10	.00	5446997.70	10	.00
M × A	52007.62	4	.00	294543.20	4	.00	9441253.80	4	.01
N × D	217652.37	5	.01	298744.10	5	.00	786395.00	5	.00
N × A	363768.00	2	.01	53898.92	2	.00	522189.70	2	.00
D × A	1388355.02	10	.05	179404100.00	10	.15	38803630.40	10	.03
Within	9242174.52		.33	906413400.00		.75	1058213650.80		.72
Total	28191207.31			1205766000.00			1464991166.80		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions.

Kendall's τb. Relative proximity of the true and NRM-based scoring key (Kendall's τb) also mainly varied with the fixation method. For the fixation based on true ordering of response options (NRM V1), the rank order of true and reconstructed scoring keys was very close, with a mean correlation of $M = .98$ ($SD = .04$) ranging from .83 to a perfect correlation. Within conditions, the correlation hardly varied ($M = .01$, $SD = .02$), with a range from zero to a maximum of .01. For NRM V2, where highest and lowest response options were randomly fixed, the mean value of Kendall's τb between conditions was $M = -.01$ ($SD = .13$) ranging from $-.26$ to .26. Within variation was moderately higher for this fixation method, with a mean of $M = .50$ ($SD = .11$) ranging from, .34 to .75. When fixation was based on plausible but still uncertain assumptions about the highest and lowest category (NRM V3), the rank correlation of true and reconstructed scoring key was fairly high with a mean of $M = .67$ ($SD = .09$), ranging from .49 to .78. Within variation was moderate with a mean of $M = .16$ ($SD = .06$) with a range of .09 to .33.

The effect sizes of the four-way ANOVA are presented in Table 21. For NRM V1[9], a main part of the variation of Kendall's τb is due to within variation (28%). In addition, the main effects number of respondents ($\eta^2 = .16$), ability ($\eta^2 = .07$), and difficulty ($\eta^2 = .05$) had small to medium impact. Moreover, several interaction effects explain variance: interaction of ability and difficulty ($\eta^2 = .21$), interaction of number of respondents and ability ($\eta^2 = .05$), and interaction of number of respondents and difficulty ($\eta^2 = .04$). However, the graphical presentation of effects (see Figures 84 and 85 in Appendix E) reveals that absolute differences in Kendall's τb were very small, as the correlation was on a high level in all conditions. Therefore, the direction of effect is not further described here: it is presented in the Appendix.

When fixation was randomly assigned (NRM V2), values of Kendall's τb mainly varied within conditions ($\eta^2 = .94$). Only ability ($\eta^2 = .03$) and difficulty ($\eta^2 = .02$) had

[9] Note, that only 86% of the total variance was explained here by main and first order effects as well as within variance. Hence, higher-order interactions account for a substantial amount of variation.

additional small effects. Here, correlation was lower for easy items and higher for difficult items (see Figure 86 in Appendix E). Moreover, correlation was of low positive size for low ability, whereas it was of low negative size for high ability.

For NRM V3, again most variation was due to within condition variation ($\eta^2 = .79$), however, small effects of ability ($\eta^2 = .06$) and number of variables ($\eta^2 = .03$) were observed. In addition, the interaction of ability and difficulty explained 6% of variance. Kendall's τb was slightly higher for 12 items compared to 48 items. Regarding ability, higher rank-order correlations were observed for low ability, and lower correlations for high ability. Again, the independent variables ability and difficulty interacted in the following way: The difference in rank-order correlations for different ability levels was higher for easy items than for difficult items. For easy items, low ability samples provided the highest correlation.

Confirmatory Factor Analyses. As presented in Table 22, for all NRM studies the confirmation of the one-dimensional structure showed on average good values for RMSEA and more or less satisfactory values for CFI. However, the fit did not substantially differ between NRM studies. Hence, even if the scoring according to NRM does not work very well - as results above indicate for NRM V2 - the structure is not influenced. When scoring works well, as indicated by results for NRM V1, fit indices show the same support for one-dimensionality.

The directions of effects are very similar for all NRM studies, as presented in Appendix E. CFI slightly decreased and number of p-values smaller than alpha distinctly increased with number of variables. However, the number of variables revealed no influence on RMSEA values. Sample size had a positive effect on fit indices: CFI increased and RMSEA decreased with $N = 1000$ compared to $N = 200$, whereas the number of significant χ^2-tests was not affected. The independent variable ability had a small effect on RMSEA and CFI: The fit was better for medium ability, but worse for high and low ability. Model fit was best for difficulty combination [e,e,m] and worst for [e,m,d].

Table 21

Four-way ANOVA Results for Kendall's τb (NRM)

	NRM V1			NRM V2			NRM V3		
Source	SS	*df*	η^2	SS	*df*	η^2	SS	*df*	η^2
M: No. of variables	1.41	2	.00	.11	2	.00	1033.16	2	.03
N: No. of respondents	286.11	1	.16	3.78	1	.00	38.01	1	.00
D: Difficulty	83.95	5	.05	6057.17	5	.02	934.78	5	.02
A: Ability	125.52	2	.07	9434.71	2	.03	2194.80	2	.06
M × N	0.91	2	.00	.72	2	.00	2.27	2	.00
M × D	0.12	10	.00	17.86	10	.00	325.21	10	.01
M × A	0.11	4	.00	28.71	4	.00	560.29	4	.01
N × D	66.46	5	.04	226.09	5	.00	18.45	5	.00
N × A	94.69	2	.05	296.70	2	.00	25.83	2	.00
D × A	379.00	10	.21	1257.01	10	.00	2189.44	10	.06
Within	512.92		.28	280040.14		.94	30484.24		.79
Total	1835.06			297595.96			38413.67		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions.

Table 22

Descriptive Statistics for CFA Results (NRM)

	NRM V1				NRM V2				NRM V3			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Chi Square			60.92	3370.05			62.39	2663.26			61.19	3221.00
df			54	1080			54	1080			54	1080
No. p-val[a]	8576.90	2436.10	1684	10000	8477.63	2433.33	2057	10000	8546.713	2425.36	1709	10000
CFI	.94	.07	.65	1.00	.95	.05	.75	1.00	.95	.06	.66	1.00
RMSEA	.04	.03	.01	.11	.04	.02	.01	.09	.04	.02	.01	.10

Note. Min = Minimum; Max = Maximum. Estimator: Weighted Least Squares mean and variance adjusted. [a]Number of p-values p < .05.

5.2.2 Comparison of Scoring Methods. The second part of the analysis focuses on the comparison of different scoring methods. The ability estimates of each scoring method based either on scored data (i.e., sum scores for CBM and Consensus Analysis) or on parameter estimates of the scoring method (i.e., person parameter estimates for HOMALS and NRM) (see paragraph 5.1.2.2) were correlated with true abilities to determine the extent to which ability estimates after scoring reflect true ability. The Pearson correlation between true abilities and ability estimates allows direct comparison of the results for the different methods. For purposes of comparison, the mean of the correlation between true ability and sum scores for every respondent before using any scoring method was $M = .87$ ($SD = .06$) for two-categorical data and $M = .93$ ($SD = .04$) for five-categorical data (for the full design of 270 conditions).

5.2.2.1 Pearson Correlation Between True Abilities and Estimated Abilities. Table 23 gives an overview about the results for the Pearson correlation between true and estimated abilities for all scoring methods. For consensus-based scoring methods - that is, mode and proportion CBM and Consensus Analysis - results are very similar: Total mean values of correlations do not substantially differ from zero; hence, across all conditions, the sum scores based on these scoring methods have nothing in common with true abilities. However, the variation between conditions is quite high, indicating that some conditions yielded high positive correlations between the true and estimated abilities. In fact, the maximum values of correlations are very high for all methods. The within condition variation was comparably low for these methods; however, it was slightly higher for Consensus Analysis.

In contrast to the consensus-based methods, the correlations did not vary greatly between conditions for HOMALS scoring (see Table 23). The mean values, standard deviations as well as maximum values of the Pearson correlation indicated that HOMALS did not provide valid person parameter estimates in any condition. The positive correlation did not exceed $r = .29$ and $r = .38$. However, within conditions the variation of correlations was substantially higher compared to consensus-based methods.

Table 23

Descriptive Statistics for the Dependent Variable Pearson Correlation Between True and Estimated Abilities

		Total mean (between)				Total variance (within)			
Method	Data	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Mode CBM	two-categorical	.02	.65	−.93	.93	.21	.20	.00	.63
Mode CBM	five-categorical	.01	.73	−.96	.96	.18	.21	.00	.67
Prop. CBM	two-categorical	.02	.69	−.92	.92	.16	.21	.00	.67
Prop. CBM	five-categorical	.02	.77	−.96	.96	.16	.23	.00	.70
CA	two-categorical	.01	.74	−.94	.94	.26	.35	.01	.95
CA	five-categorical	.01	.80	−.96	.96	.23	.34	.01	.90
HOMALS	two-categorical	−.01	.11	−.29	.29	.84	.08	.56	.95
HOMALS	five-categorical	−.01	.19	−.38	.38	.87	.11	.54	.98
NRM V1	five-categorical	.96	.03	.86	.99	.01	.00	.00	.02
NRM V2	five-categorical	.00	.03	−.08	.07	.95	.03	.86	.99
NRM V3	five-categorical	.93	.08	.46	.99	.11	.16	.00	.74

Note. Mode CBM = Mode Consensus Based Measurement; Prop. CBM = Proportion Consensus Based Measurement; CA = Consensus Analysis; HOMALS = Homogeneity with Alternating Least Squares; NRM V1 = Nominal Response Model with fixation based on true ordering; NRM V2 = Nominal Response Model with random fixation; NRM V3 = Nominal Response Model with plausible fixation.

For the NRM studies, the correlation between true abilities and NRM ability estimates varied with the fixation of the highest and the lowest response category. For NRM V1 - where fixation was based on true ordering of response options - the mean correlation between conditions was very high and the variation of the dependent variable within condition was rather small. Hence, with the optimal fixation choice, the NRM estimation method completely succeeded, albeit, with some variation between conditions. For NRM V2 - where fixation was based on random choice - the mean correlation between conditions was not substantially different from zero. As the low minimum and maximum values indicate, this method did not succeed in any of the specified conditions. However, within conditions the variation of correlations was very large. Thus, as fixation was random for NRM V2, it can be concluded that the correlation of true abilities and NRM ability estimates was dependent upon the (random) quality of fixation. When fixation was based on a good choice of highest and lowest category (NRM V3), the total mean correlation was quite high. The variation between conditions was higher for NRM V3 compared to the two other NRM studies, indicating that the correlation for NRM V3 is somehow dependent upon the varied factors. The variation within condition was higher compared to NRM V1, but not as high as in NRM V2, mapping the partly random process of fixation.

As Table 24 shows for mode CBM, the independent variables difficulty and ability had the biggest influence on the correlation of abilities for both two-categorical and five-categorical data; ability explained 44% and 50%, difficulty explained 28% and 27% of the variance of the dependent variable. In addition, effect sizes for the interaction of ability and difficulty (η^2 = .10 and η^2 = .11) and within-condition variance (η^2 = .17 and η^2 = .12) were of medium size. The effects of difficulty and ability on the dependent variable were very similar for proportion CBM scoring (see Table 25). Here, ability had a large effect (η^2 = .51 and η^2 = .52), whereas the effects of difficulty of items (η^2 = .24 and η^2 = .23) and the interaction effect of ability and difficulty (η^2 = .12 and η^2 = .13) were of medium size.

The direction of effects was very similar for mode and proportion CBM for both two- and five-categorical data. To illustrate this, Figure 13 shows main effects for five-categorical data and proportion CBM. Graphics for both mode CBM and two-categorical data are included in Appendix F. As Figure 13 indicates, correlation was highest ($r > .50$) with

Table 24

Four-way ANOVA Results for Pearson Correlation (Mode CBM)

	2-categorical data			5-categorical data		
Source	SS	*df*	η^2	SS	*df*	η^2
M: No. of variables	21.68	2	.00	49.16	2	.00
N: No. of respondents	.74	4	.00	.34	4	.00
D: Difficulty	379604.22	5	.28	442866.18	5	.27
A: Ability	597971.70	2	.44	821571.36	2	.50
M × N	1.09	8	.00	.84	8	.00
M × D	2089.45	10	.00	675.78	10	.00
M × A	5396.39	4	.00	2753.00	4	.00
N × D	37.91	20	.00	75.07	20	.00
N × A	49.20	8	.00	376.94	8	.00
D × A	137963.97	10	.10	178712.42	10	.11
Within	227114.82		.17	203993.55		.12
Total	1352599.87			1654126.72		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions.

only easy items, it decreased with difficulty combination [e,e,m], and was zero for combinations [m,m,m] and [e,m,d]. With difficulty combination [d,d,e] correlation was negative and decreased further ($r < -.50$) for combinations of only difficult items. With high ability, the correlation was highly positive, that is, the mean correlation for all conditions with high ability was greater than $r = .50$. With medium ability the correlation was around zero and with low ability it was highly negative, that is, smaller than $r = -.50$.

Interaction effects were also very similar for proportion and mode CBM for both two- and five-categorical data. To illustrate the directions, Figure 14 presents the interaction

Table 25

Four-way ANOVA Results for Pearson Correlation (Proportion CBM)

	2-categorical data			5-categorical data		
Source	SS	*df*	η^2	SS	*df*	η^2
M: No. of variables	4.33	2	.00	7.25	2	.00
N: No. of respondents	15.86	4	.00	80.07	4	.00
D: Difficulty	341611.30	5	.24	407006.55	5	.23
A: Ability	741717.62	2	.51	942415.62	2	.52
M × N	1.20	8	.00	1.55	8	.00
M × D	1469.31	10	.00	268.61	10	.00
M × A	5290.82	4	.00	1394.35	4	.00
N × D	121.93	20	.00	216.65	20	.00
N × A	589.12	8	.00	612.13	8	.00
D × A	170664.17	10	.12	231392.16	10	.13
Within	186749.96		.13	213823.41		.12
Total	1451592.09			1799806.97		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions.

effect of ability and difficulty for proportion CBM for five-categorical data (for more interaction plots for CBM see Appendix F). As can be seen from the interaction plot, high ability in combination with easy and medium difficult items (i.e., combinations [e,e,e],[e,e,m],[e,m,d] and [m,m,m]) yielded high positive correlations, whereas correlation dropped when item difficulty increased. However, it was still reasonably high for the difficulty combination [d,d,e] (and high ability), but dropped to zero for the combination [d,d,d]. For medium ability, correlations were highly positive for easy items (i.e., [e,e,e] and [e,e,m]), around zero for medium difficult items (i.e., [e,m,d] and [m,m,m]) and highly

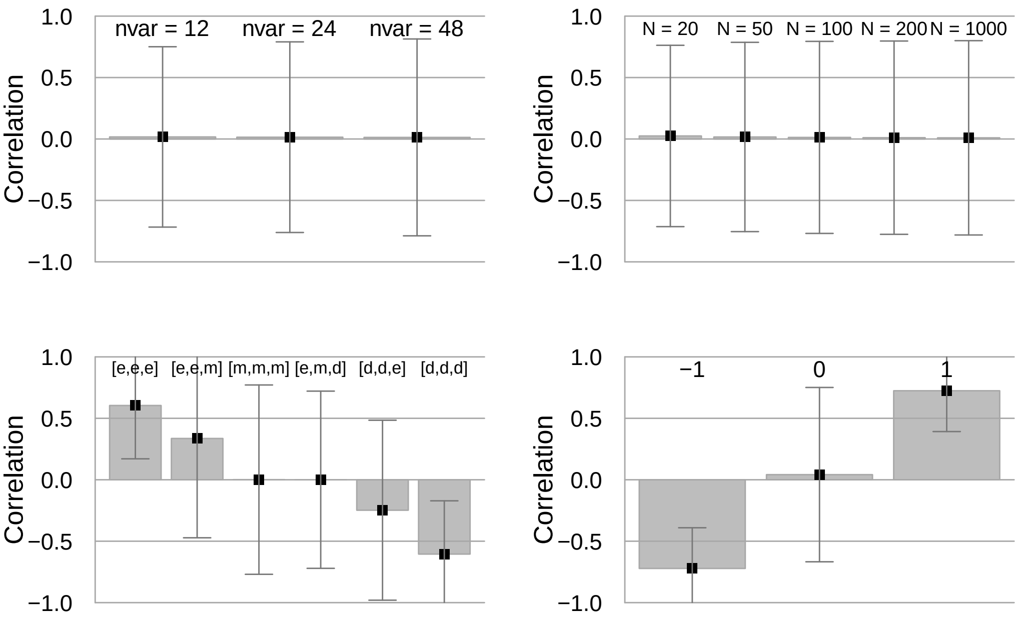


Figure 13. Five-categorical data: Main effects of independent variables on correlation (Proportion CBM). Top left panel: Number of variables; top right panel: Number of respondents; bottom left panel: Difficulty; bottom right panel: Ability.

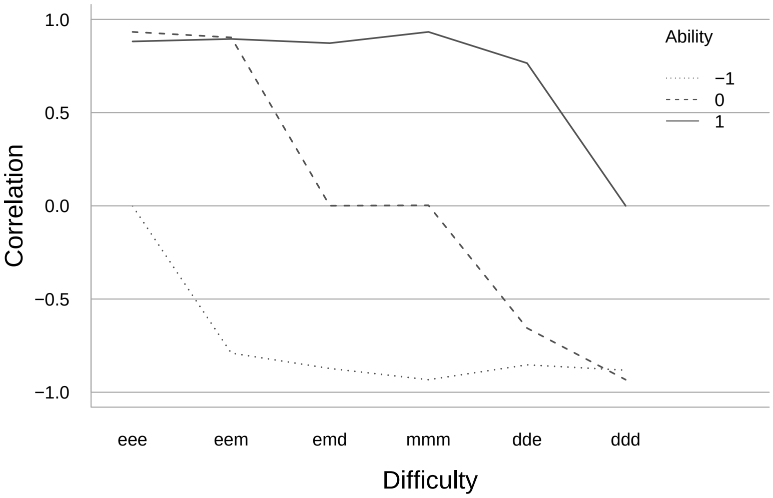


Figure 14. Five-categorical data: Interaction effect of ability and difficulty on correlation (Proportion CBM).

negative for difficult items (i.e., [d,d,e] and [d,d,d]). For low ability, however, correlations never exceeded a value of zero.

For Consensus Analysis, as Table 26 shows, the dependent variable mainly varied with ability ($\eta^2 = .44$ and $\eta^2 = .47$), difficulty ($\eta^2 = .17$ and $\eta^2 = .19$), and interaction of both ($\eta^2 = .12$ and $\eta^2 = .13$). The within variation accounted for 26% and 21% of the variance. Compared to the CBM methods, the effect of within variation was higher and the effects of ability and difficulty were slightly smaller. However, these main and interaction effects are the most important influencing variables for all consensus-based scoring methods.

The direction of effects of Consensus Analysis for both two- and five-categorical data was very similar to the CBM effects presented above. As Figure 15 illustrates for Consensus Analysis for five-categorical data, correlation was highly positive ($r > .50$) for high ability, around zero for medium ability and highly negative ($r < -.50$) for low ability. In addition, correlation was highly positive ($r > .50$) for easy items, whereas it decreased with medium difficult items and resulted in a highly negative mean correlation for difficult items. Interaction effects in Figure 16 were again similar to CBM, although small differences appeared: With high ability, correlation was high for almost every difficulty combination, except for only difficult items. Thus, correlations were higher for difficulty combination of [d,d,e] and high ability for Consensus Analysis compared to CBM methods.

Calculation of the effects of the independent variables on the correlation between true abilities and component scores of the HOMALS algorithm was based on the reduced ANOVA design. That is, the factor levels $N = 20$ and $N = 50$ for two-categorical data and the level $N = 20$ for five categorical data were excluded from analysis due to missing values. Results in Table 27 indicate that a huge amount of variance for both two- and five categorical data remains unexplained. Only ability and difficulty were observed to have very small effects on the dependent variable. Graphical inspection of effects for the full design for HOMALS indicated that the quality of the algorithm seemed to vary randomly

Table 26

Four-way ANOVA Results for Pearson Correlation (Consensus Analysis)

	2-categorical data			5-categorical data		
Source	SS	*df*	η^2	SS	*df*	η^2
M: No. of variables	5.87	2	.00	.50	2	.00
N: No. of respondents	5.87	3	.00	67.80	3	.00
D: Difficulty	270975.84	5	.17	333095.71	5	.19
A: Ability	694680.40	2	.44	812727.91	2	.47
M × N	2.61	6	.00	.23	6	.00
M × D	1286.65	10	.00	422.49	10	.00
M × A	3465.30	4	.00	1098.02	4	.00
N × D	360.97	15	.00	151.31	15	.00
N × A	723.70	6	.00	183.68	6	.00
D × A	188402.25	10	.12	225754.86	10	.13
Within	400839.81		.26	362574.76		.21
Total	1563463.25			1737292.69		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions.

within conditions. The small main effects that were observed were similar to the effects on the other dependent variables for HOMALS (see Figures 109 and 112 in Appendix F): The correlations increased with low ability ($M = .10$ and $M = .18$ for two- and five categorical data), were very close to zero with medium ability ($M = -.01$), and were negative for high ability ($M = -.11$ and $M = -.19$). In addition, the correlation was highest ($M = -.11$ and $M = .19$) for the difficulty combination [d,d,d], and lowest ($M = -.11$ and $M = -.19$) for the combination [e,e,e].

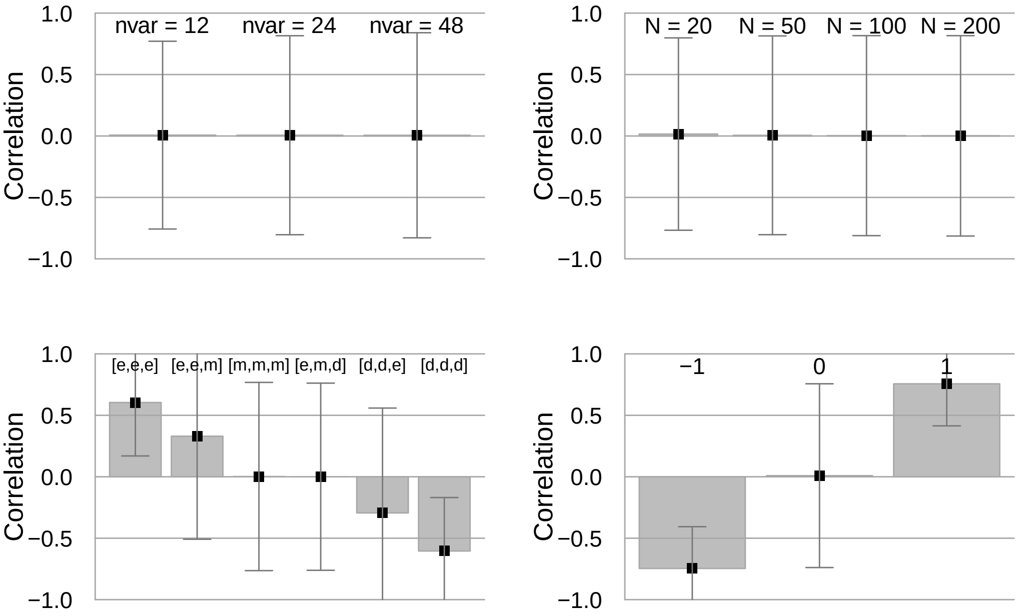


Figure 15. Five-categorical data: Main effects of independent variables on correlation (Consensus Analysis). Top left panel: Number of variables; top right panel: Number of respondents; bottom left panel: Difficulty; bottom right panel: Ability.

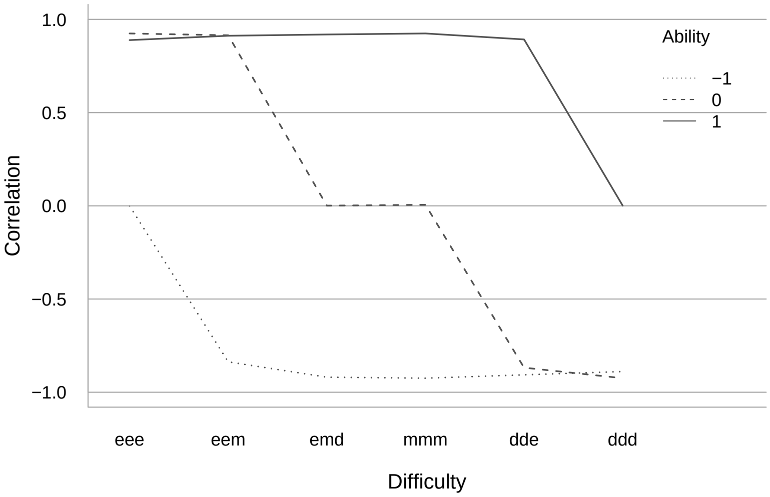


Figure 16. Five-categorical data: Interaction effect of ability and difficulty on correlation (Consensus Analysis).

Table 27

Four-way ANOVA Results for Pearson Correlation (HOMALS)

	2-categorical data			5-categorical data		
Source	SS	*df*	η^2	SS	*df*	η^2
M: No. of variables	2.14	2	.00	1.53	2	.00
N: No. of respondents	.63	2	.00	1.92	5	.00
D: Difficulty	6525.32	5	.01	30883.09	3	.02
A: Ability	9193.21	2	.01	43685.04	2	.02
M × N	5.56	4	.00	1.53	6	.00
M × D	98.98	10	.00	24.12	10	.00
M × A	108.70	4	.00	24.50	4	.00
N × D	36.31	10	.00	86.70	15	.00
N × A	75.97	4	.00	90.40	6	.00
D × A	211.79	10	.00	1848.20	10	.00
Within	1201490.00		.99	1681339.78		.96
Total	1217935.00			1758143.79		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions.

For NRM studies, the value of the correlation between true and estimated abilities, as well as the influences of independent variables on the correlation, differed between the three fixation methods. For NRM V1, the variation of the dependent variable was mainly accounted for by number of variables (η^2 = .50) and interaction of difficulty and ability (η^2 = .29). Small effects also occurred for (amongst other things) difficulty and ability (see Table 28). The main effect graphics for NRM V1 can be found in Appendix F, Figure 115. The effect of the independent variable number of variables has to be interpreted with respect to the low variation between conditions: The correlation was high on every level of the independent variable; for 12 items it was $M = .93$, for 24 items it was $M = .96$ and for 48 items it was $M = .98$. Because of low between condition variation, the differences in correlations between different combinations of factor levels of ability and difficulty were also rather low (see Figure F, Appendix 116). With high ability and easy items, as well as with low ability and difficult items, the correlations slightly dropped compared to the other factor level combinations. In addition, the very small effects of the factors ability and difficulty presented themselves in a way that for medium ability correlation was higher compared to low and high ability, and that for difficulty combination [d,d,d] the correlations dropped.

For NRM V2, as Table 28 states, the variation of correlations is almost completely accounted for by within variance, which masked other effects. Hence, the correlation depended on the random good choice of fixation, but did not systematically vary with the levels of independent variables (see Figure 117 in Appendix F).

The high correlations between true ability and NRM ability estimates mainly varied within conditions ($\eta^2 = .85$) for NRM V3. However, other rather small influences became apparent, namely, the interaction of ability and difficulty ($\eta^2 = .05$) and the main effects of difficulty ($\eta^2 = .03$), number of variables ($\eta^2 = .03$) and ability ($\eta^2 = .03$). The main effects are presented in Figure 118 in Appendix F. With increasing number of variables, the correlation increases: From $M = .88$ for 12 items, increasing to $M = .95$ for 24 items, and $M = .97$ for 48 items. Moreover, the correlation decreased when only difficult items were used ($M = .86$), whereas for all other difficulty combinations the correlations were higher and did not vary much ($M = .94 - .95$). In addition, with low ability correlation decreased ($M = .90$), whereas with medium and high ability it was roughly equal ($M = .96$ and $M = .95$). The interaction of ability and difficulty is interesting: For the combination of low ability and only difficult items, the correlation substantially dropped to a mean of $M = .68$ ($SD = .17$) (see Figure 119 in Appendix F).

Table 28

Four-way ANOVA Results for Pearson Correlation (NRM)

	NRM V1			NRM V2			NRM V3		
Source	SS	*df*	η^2	SS	*df*	η^2	SS	*df*	η^2
M: No. of variables	424.27	2	.50	.41	2	.00	1469.99	2	.03
N: No. of respondents	2.24	1	.00	2.73	1	.00	2.42	1	.00
D: Difficulty	56.55	5	.07	36.95	5	.00	1189.74	5	.03
A: Ability	39.44	2	.05	151.04	2	.00	715.38	2	.02
M × N	.11	2	.00	4.14	2	.00	.05	2	.00
M × D	4.98	10	.01	15.35	10	.00	294.48	10	.01
M × A	4.55	4	.01	4.41	4	.00	274.52	4	.01
N × D	.15	5	.00	55.49	5	.00	31.84	5	.00
N × A	.10	2	.00	26.27	2	.00	20.32	2	.00
D × A	249.45	10	.29	722.28	10	.00	2347.30	10	.05
Within	35.68		.04	982548.10		1.00	38529.65		.85
Total	846.65			983727.10			45356.99		

Note. Table includes sum of squares, degrees of freedom, and effect sizes for main effects and first order interactions.

5.2.2.2 Cases With High Correlation Between Abilities. In order to draw conclusions for the application of the scoring methods (see section 7.1 in the Discussion), a criterion is required to define good performance of these methods. As stated before, a high positive correlation indicates that a method performs well, but the size of correlation that designates the cut-off point of good performance is somehow arbitrary. As reliability of the data is defined as the squared correlation between true and observed scores, the conventions for reliability may be used as such a criterion. Consistent with current standards in psychological research, good reliability is usually defined as at least .80. With a correlation of $r = .90$ or above, a reliability of at least .81 is provided. In addition to this rather strict criterion, satisfactory values for reliability (.70) were also considered. In the following, cases with correlations of $r \geq .84$ and $r \geq .90$ are considered in more detail for each scoring method. However, HOMALS scoring is not investigated any further in this paragraph as positive correlations for this method did not exceed $r = .29$ or $r = .38$. In addition, NRM V2 is not considered any further for the same reason. The following paragraph explicates cases for consensus-based scoring methods with regard to the most important influencing variables. Case analysis for NRM V1 and V3 studies is described separately afterward, as different influencing variables were important here.

Consensus-based Scoring Methods. Table 29 and 30 provide an overview of the number of conditions in which the correlations between true and estimated abilities were equal to or higher than the criterion values of .84 and .90. More conditions with these values were observed for five-categorical data compared to two-categorical data. In addition, more conditions with high correlations were observed for Consensus Analysis (in comparison to CBM methods) for two-categorical data, whereas for five-categorical data Consensus Analysis and (proportion) CBM methods did not show big differences.

Table 29

Number of Conditions with Correlations ≥ .84 for Consensus-based Scoring Methods

	2-categorical data		5-categorical data		
Method	Absolute Frequency	%	Absolute Frequency	%	Total
Mode CBM	38	14.07	78	28.89	270
Proportion CBM	39	14.44	87	32.22	270
Consensus Analysis	48	22.22	71	32.87	216

Because ability and difficulty were identified to have non-trivial influence on the dependent variable, Table 31 contains the relative frequencies with which correlations of least .84 were observed in each respective factor level combination. Factor level combinations for low ability are omitted because no high correlations were observed here. Table 32 presents these frequencies for a correlation of at least .90. The frequencies indicate how many high correlations were observed in combinations of factor levels of ability and difficulty, and hence, indicate in which situations consensus-based scoring methods worked best.

As can be seen from Table 31 and Table 32, five-categorical data in general yielded more cases with high correlations between true abilities and sum scores based on scored data. Comparing the relative frequencies for high and medium ability conditions, the main effect of ability shows its effect for all consensus-based scoring methods; that is, high ability conditions yielded more cases with high correlations compared to medium ability conditions. For medium ability, only conditions with easy difficulty combinations [e,e,e] as well as [e,e,m] yielded high correlations. In contrast, high correlations were also observed for high ability with difficult items, apart from the combination [d,d,d]. The main effect of difficulty is apparent by comparison of different difficulty combinations: For combinations with easy items, more high correlations were observed compared to items with high difficulty.

Table 30

Number of Conditions with Correlations ≥ .90 *for Consensus-based Scoring Methods*

	2-categorical data		5-categorical data		
Method	Absolute Frequency	%	Absolute Frequency	%	Total
Mode CBM	10	3.70	43	15.93	270
Proportion CBM	11	4.07	54	20.00	270
Consensus Analysis	20	9.26	51	23.61	216

Comparing the relative frequencies for mode and proportion CBM reveals some differences between these methods: proportion CBM provided more cases which met the criteria. However, a clear pattern of differences with regard to the factor level combinations was not identifiable. In comparison to CBM scoring, Consensus Analysis scoring appeared to perform slightly better as more conditions yielded high correlations, in particular with the cut-off of $r = .84$. Differences were more pronounced for difficulty combinations with mixed difficulties compared to conditions with equal difficulties (with high ability only). In other words, the number of conditions with high correlations were very similar for all consensus-based scoring methods for difficulty combinations [e,e,e], [m,m,m] and [d,d,d]. However, differences appeared for difficulty combinations [e,e,m], [e,m,d] and [d,d,e], in which Consensus Analysis performed better overall. Moreover, Consensus Analysis yielded high correlations in conditions with high ability and mostly difficult items [d,d,e] - a condition, in which no high correlation were observed for CBM methods. Similarly, for the difficulty condition [e,m,d], Consensus Analysis provided more cases in which correlations were above the cut-off point of $r = .84$ and $r = .90$. Hence, Consensus Analysis scoring showed advantages in conditions with difficult items. However, for only difficult items [d,d,d], no scoring method yielded high positive correlations.

To summarize, high correlations between true and re-estimated abilities were only observed in specific conditions characterized by combinations of ability and difficulty where

Table 31

Relative Frequencies of Observed Correlation $\geq .84$ for Factor Level Combinations of Ability and Difficulty

Two-categorical data		Ability				
	Consensus Analysis		Mode CBM		Proportion CBM	
Difficulty	0	1	0	1	0	1
$[e, e, e]$	.67	.33	.67	.33	.67	.33
$[e, e, m]$	.67	.67	.27	.60	.33	.33
$[e, m, d]$	.00	.58	.00	.00	.00	.27
$[m, m, m]$	.00	.67	.00	.67	.00	.67
$[d, d, e]$	.00	.42	.00	.00	.00	.00
$[d, d, d]$	.00	.00	.00	.00	.00	.00
Five-categorical data		Ability				
	Consensus Analysis		Mode CBM		Proportion CBM	
Difficulty	0	1	0	1	0	1
$[e, e, e]$	1.00	1.00	1.00	1.00	1.00	.93
$[e, e, m]$	1.00	1.00	.60	1.00	.93	1.00
$[e, m, d]$	.00	1.00	.00	.60	.00	.73
$[m, m, m]$	.00	1.00	.00	1.00	.00	1.00
$[d, d, e]$	.00	.92	.00	.00	.00	.20
$[d, d, d]$	.00	.00	.00	.00	.00	.00

Table 32

Relative Frequencies of Observed Correlation $\geq .90$ *for Factor Level Combinations of Ability and Difficulty*

Two-categorical data	Ability					
	Consensus Analysis		Mode CBM		Proportion CBM	
Difficulty	0	1	0	1	0	1
$[e, e, e]$	.33	.00	.33	.00	.33	.00
$[e, e, m]$	.33	.08	.00	.00	.07	.00
$[e, m, d]$	.00	.33	.00	.00	.00	.00
$[m, m, m]$	.00	.33	.00	.33	.00	.33
$[d, d, e]$	.00	.25	.00	.00	.00	.00
$[d, d, d]$	.00	.00	.00	.00	.00	.00
Five-categorical data	Ability					
	Consensus Analysis		Mode CBM		Proportion CBM	
Difficulty	0	1	0	1	0	1
$[e, e, e]$	.67	.33	.67	.33	.93	.33
$[e, e, m]$	.67	.67	.27	.67	.53	.60
$[e, m, d]$	.00	.67	.00	.27	.00	.27
$[m, m, m]$	.00	.67	.00	.67	.00	.93
$[d, d, e]$	.00	.58	.00	.00	.00	.00
$[d, d, d]$	.00	.00	.00	.00	.00	.00

mean person ability exceeded mean item difficulty. More specifically, best results were observed for the high ability group with easy or medium difficult items.

Nominal Response Model. For NRM V1, 100% of the conditions revealed a correlation between true abilities and NRM ability estimates ≥ .84. In 95.4% of the conditions, a correlation ≥ .90 was observed. The five cases, in which the correlation was lower than .90, were observed for the two conditions with low ability and difficulty combination [d,d,d], the two conditions with high ability and difficulty combination [e,e,e] as well as high ability and difficulty combination [e,e,m] (here only one of two conditions).

For the NRM V3, 101 of 108 (93.5%) conditions revealed a correlation between true abilities and NRM ability estimates of ≥ .84, and 90 of 108 (83.3%) revealed a correlation of ≥ .90. Again, the observed small main effects of number of variables, ability and difficulty, as well as the interaction of ability and difficulty showed their effects. Correlations lower than .84 were only observed for low ability in combination with high difficulty combinations, that is [d,d,e] and [d,d,d]. However, with an increasing number of variables, the number of cases with correlations smaller than .84 decreased for these conditions. While with twelve variables, no correlation of ≥ .84 was observed for low ability and difficulty combinations [d,d,e] and [d,d,d], with a higher number of variables more cases reached the criterion. For number of variables at 24, only the combination of low ability and difficulty condition [d,d,d] revealed correlations lower than .84, whereas the combination of low ability and difficulty combination [d,d,e] yielded correlations equal to or above the cut-off point. For number of variables at 48, one out of two conditions revealed correlation smaller than .84 for low ability and difficulty combination [d,d,d].

Results are similar for stricter criteria: For each level of the independent variable number of variables - that is twelve, 24 or 48 variables - the combination of low ability and difficult items [d,d,d] revealed correlations smaller than .90. In addition, for number of variables at twelve, correlations were lower than .90 in the combinations low ability with [e,m,d], [m,m,m] and [d,d,e] as well as with medium ability and only difficult items ([d,d,d]). Moreover, with high ability and difficulty combinations [e,e,e] and [e,e,m], the correlations were smaller than .90. In all other combinations, the criterion was reached and correlations were higher than .90.

5.2.3 Summary of Results. The results for the HOMALS scoring method are poor throughout all dependent variables. The relative proximity of the true and HOMALS scoring was low and the correlations of component scores and true abilities did not exceed a low to medium high correlation in any condition. The dependent variables did not mainly vary with the manipulated data characteristics, but were dependent upon unspecified random processes. Moreover, the one-dimensional structure of the scored data set was not supported, but this could be due to the choice of the ML estimator.

Consensus-based empirical scoring methods (CBM and Consensus Analysis) mainly depended on the manipulated variables ability and difficulty and their interaction. In total, these methods only performed well in situations in which the ability of the individuals exceeded the difficulty of items. The two CBM scoring methods did not show substantial differences; however, proportion CBM showed small advantages as this method provided more high correlations. Results for Consensus Analysis were slightly better, in particular because with highly able individuals and mainly difficult items [d,d,e] high correlations were still observed. In addition, Consensus Analysis performed better (in terms of high correlations) compared to CBM in situations with mixed item difficulties. However, the mixed difficulty combinations were also the cases in which zero variances were observed for five-categorical data, indicating a failure of the algorithm. The structure remained stable after scoring for mode CBM and Consensus Analysis, independent of whether scoring worked or not. The CFA results were more dependent on other data characteristics, such as sample size. For proportion CBM scoring, results of CFAs were non-satisfactory, however, the number of convergence problems indicates that the estimator choice might have been problematic, and the difference in fit between proportion CBM and mode CBM scored data may be due to the estimator.

The NRM scoring method depended less on manipulated factors, than on fixation methods. With fixed most correct as well as least correct response options, the NRM fully succeeded in re-estimating the abilities of individuals, as the correlation between true and estimated abilities was highly positive for most conditions. However, the success of re-estimation was somehow dependent upon varied factors, most importantly number of variables (correlation was higher for more variables) and the interaction of ability and difficulty (correlation was slightly lower for easy items and high ability as well as for difficult items and low ability). The relative proximity of scoring keys was high and the distance between scoring keys was relatively low compared to other NRM studies.

With random ordering of response options, the NRM did not succeed in estimating the true ordering of response options as well as the true abilities of individuals: That is, the correlation between true and estimated abilities as well as relative proximity of scoring keys was low, whereas distance between scoring keys was high. For all dependent variables, the within-variance masked effects of manipulated factors and explained an huge amount of variance in dependent variables.

For fixation based on uncertain, yet plausible assumptions, the correlation between true and estimated abilities was highly positive, however mainly dependent upon random processes and partly dependent upon manipulated data characteristics (small effects), such as the interaction of ability and difficulty (lowest for the combination of low ability and difficult items) as well as the main effects of number of variables (higher with more variables), difficulty (lower for difficult items), and ability (lower for low ability). The relative proximity of the two scoring keys was also quite high for NRM V3, and small effects of the independent variables were also observed. However, these effects partly differ from that of the dependent variable correlation of abilities. The scoring keys were closer for low ability, fewer variables and difficult items. In addition, the CFA results for NRM methods revealed that structure was stable independent of the performance of scoring methods, although fit indices slightly varied with all other manipulated data characteristics.

6 Evaluation of Empirical Scoring Methods with TIMSS 2011 Data

6.1 Methods

The eighth grade mathematics items of TIMSS 2011, conducted by the International Association for the Evaluation of Educational Achievement (IEA), were used for evaluating scoring methods with real-world data. TIMSS is an international study of 63 countries and fourteen benchmark countries assessing performance in both mathematics and science, comparing educational systems and collecting information about learning climate, resources, teachers, et cetera (Mullis, Martin, Foy, & Arora, 2012). TIMSS seeks to provide input for political and educational policy makers in order to improve teaching and learning environments in science and mathematics. Planned, developed, designed and conducted with high quality standards - for instance concerning item translations, item pre-testing, and sampling (Mullis, Drucker, Preuschoff, Arora, & Stanco, 2012; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009) - the study has been conducted at four year intervals: 1995, 1999, 2003, 2007, 2011, and 2015. The TIMSS target population include students with four and eight years of schooling starting with the first school years (Mullis et al., 2009). For the present study, the eighth-grade data was used, which was collected in 45 countries and fourteen benchmark countries. Some countries decided to administer TIMSS in sixth and ninth grade due to distinctive features in their education system. These countries, as well as the benchmark countries, will not be used in the present study. A total of $600,000$ students participated in TIMSS 2011, approximately 300,000 for each fourth and eighth-grade. Each country collected data from approximately 4000 students in 150-200 different schools. In 2011, a total of $N = 307,038$ eighth-grade students took part in the study, without ninth grade and benchmarking the sample size reduces to $N = 239,960$ students.

6.1.1 TIMSS 2011: Assessment information. For this dissertation project, items measuring mathematical competence were selected. The mathematical competence assessed in TIMSS is structured in sub-abilities of four contents (number, algebra, geometry, and data & chance) and three cognitive domains (knowing, applying, and reasoning). The expected ideal performance of eighth-grade students in each domain - that is, what students should know according to the curricula - is described in greater detail in Mullis et al. (2009). For the TIMSS 2011 study, the hypothesized measurement model has been partly investigated in a study of the Turkish sub sample (Arikan, 2015).

Regarding the abilities in the content domain of number, eighth-grade students should provide an understanding of numbers, ways of representing numbers, relationship between numbers, and number systems (Mullis et al., 2009), including whole numbers, fractions, decimals, integers, ratios, proportion, and percent. For example, they should be able to convert between fractions and decimals and to approximate with whole numbers (Mullis et al., 2009, p. 31). Regarding algebra, eighth-grade students should recognize and extend patterns, use algebraic symbols to represent a mathematical situation, and fluently solve linear equations including patterns, algebraic expression, formulas and functions (Mullis et al., 2009, p. 32). For example, they should be able to generalize pattern relationships in a sequence and indicate whether a value satisfies a given equation. In geometry, eighth-grade students should "be able to analyze the properties and characteristics of a variety of two and three-dimensional geometric figures, including lengths of sides and sizes of angles, and to provide explanations based on geometric relationships. They should be able to apply the Pythagorean Theorem to solve problems." (Mullis et al., 2009, p. 34). Geometric competence includes geometric shapes, geometric measurement, location, and movement. For example, students should be able to measure and estimate the size of a given angle or use geometric transformations such as reflection or rotation (Mullis et al., 2009, p. 36). Regarding data and chance, eighth-grade students should be able to organize data, display data in graphs and charts and understand issues of misinterpretation of data. For example, they should be able to organize data in tables and compare data using mean, modus and median (Mullis et al., 2009, p. 37).

Content domains are cross-classified with cognitive domains as described by Mullis et al. (2009). The domain knowledge includes the recall of definitions or terminology, recognition, computation, retrieval, use of instruments, and the ability to classify as well as to order. The domain applying includes solving routine problems, forming representations, selecting a strategy, generating appropriate models, and implementing instructions. Reasoning includes solving non-routine problems, analyzing data, integrating and synthesizing knowledge, and justifying statements by known mathematical results.

The number of items in each domain was determined by experts in the respective areas and curricula (Mullis et al., 2009). For eighth-grade, 30% each of the items were allocated to the content domains number and algebra, and 20% each to geometry and data and chance. In addition, 35% of the items represented the cognitive domain knowing, 40% applying, and 25% reasoning. For eighth-grade, a total of 217 mathematics items was administered. Approximately half of them were multiple choice items and the other half were constructed response items. 60% of the items were retained from previous TIMSS studies, which allowed linkage of studies across time. The other 40% of items were newly developed by trained item writers. If countries used calculators in everyday schooling, students were permitted to use calculators for solving the tasks. However, the items were developed and tested to function equivalently without calculator use.

As in many international educational comparison studies, a particular survey matrix design was used in TIMSS 2011 to provide reliable measurement with an adequate number of items while at the same time ensuring that each student only had to answer a manageable number of items. Hence students had to respond to items from one out of 14 student achievement booklets. The booklets were distributed so that each booklet's data had approximately the same mean ability. Each eighth-grade booklet contained two science and mathematics item blocks with twelve to 18 items. To link the different booklets, each item appeared in two different booklets. Of the 28 blocks (14 each for mathematics and science), twelve blocks were newly developed, whereas 16 blocks included trend (linking) items which had been administered before. Each block took 22.5 minutes testing time;

hence total time was limited to 90 minutes performance testing and 30 minutes questionnaires. For half of the booklets, the mathematics blocks came first and for most booklets half of the items were linking items. A total of six blocks of mathematics items was released, while eight were kept secure. The released items of TIMSS 2011 were administered in blocks M01, M02, M03, M05, M06, or M07; these were replaced in the 2015 TIMSS round.

For TIMSS 2011, the multiple choice items were analyzed with a 3-pl IRT model (Foy, Brossman, & Galia, 2012) for which items were calibrated concurrently - that is, data from 2007 and 2011 was used concurrently to estimate item parameters. The item difficulties of calibration were used in this study to evaluate psychometric characteristics of the selected items. Person parameters of students were estimated with plausible values estimation, which allowed comparison of population characteristics, but not at the individual level, as the uncertainties of person parameter estimates are too large (Martin & Mullis, 2012). Hence these estimates could not be used for the present study as estimates of individual student ability.

6.1.2 Item Selection and Psychometric Properties of Selected Items. For the present study, items were selected based on different criteria: 1) Items should be multiple choice and should be locally stochastically independent (i.e., no testlet-items), 2) items must be administered in the same booklet, as no planned missing data were aspired, 3) number of items should be at least ten, 4) the difficulty of the selected items should not be different from the mean difficulty of all TIMSS 2011 mathematics items, and 5) the item content should be released, so that the response options could be investigated. The MC items of blocks of booklet six provided the best fit with these criteria. In total, 18 mathematics MC items were selected which had a mean calibrated difficulty of $M = .45$ ($SD = .50$). The mean calibrated difficulty of all 303 mathematics items (including the items of TIMSS 2007 for concurrent calibration) was $M = .45$ ($SD = .57$) (Martin & Mullis, 2012).

Across 42 countries, $N = 16,380$ students responded to most of these MC items. As the students in Chile did not respond to one of the items (M032331), they were excluded from data analysis, reducing sample size to $N = 15,992$. The 18 items with the identification numbers M042016, M042024, M042067, M042150, M042260, M042041, M042077, M042235, M032295, M032679, M032331, M032047, M032398, M032352, M032507, M032424, M032738, and M032623 can be found in Foy, Arora, and Stanco (2013). Eight of the items were from the domain applying, two from the domain reasoning and eight from the domain knowing. Out of the 18 items, three included number content, eight algebra content, five geometry content, and two data and chance content.

Before scoring methods were applied, the raw data was scored with the theory- or expert-based TIMSS scoring key and the items were analyzed for basic psychometric properties - that is, internal consistency of the 18 items was estimated and their one-dimensionality was tested. To allow comparison with simulation studies, the items should at least show satisfactory internal consistency and should load on one factor. For the full data set ($N = 15\ 992$), Cronbach's alpha was $\alpha = .814$ (95% CI .810-.818). The hypothesized one-dimensional structure was supported by CFA results, $\chi^2_{WLSMV} = 3060.775$, $df = 135$, $p = .000$, $CFI = .966$, $RMSEA = .037$ with a 90% CI [.036; .038]. The distribution of sum scores of correctly answered items had a mean of $M = 9.02$ ($SD = 4.28$) and was positively-skewed, indicating a higher density lower sum scores. As the results of preliminary psychometric analysis show, the assumption of one-dimensionality of data can be maintained and the reliability estimates are above the satisfactory level.

For the full data set, missing responses were treated as if the students had given an incorrect response. However, as the scoring methods were used for raw, non-scored data, missing values needed to be either eliminated from the data set or modeled within the scoring method. In total, 2.23% of the data points were missing, either because the response to an item was omitted by a student or the item was not reached due to time limitation.[10] For the application of the scoring methods, two approaches were used to deal with the missing data:

1. Respondents with missing values due to time limitation were excluded from the data set, resulting in a sample size of $N = 15\ 732$ (data set MVI - missing values integrated). Missing values due to omission of the item were modeled as an additional response option. When missing values due to time limitation were deleted listwise, Cronbach's alpha was $\alpha = .814$, 95% CI [.810, .818]. In addition, the results of the CFA indicated one-dimensionality, $\chi^2_{WLSMV} = 2998.574$, *df* $= 135$, $p = .000$, $CFI = .966$, $RMSEA = .037$ with a 90% CI [.036; .038]. The distribution of sum scores was again positively-skewed, with a mean of $M = 9.07$ ($SD = 4.28$).
2. Respondents with missing values were excluded from data analysis, reducing sample size to $N = 13\ 452$ (data set MVD - missing values deleted). Cronbach's alpha of the 18 items was $\alpha = .823$, 95% CI [.819; .827] in the respective sample. The one-dimensionality of the items was supported by CFA results, $\chi^2_{WLSMV} = 2555.068$, *df* $= 135$, $p = .000$, $CFI = .970$, $RMSEA = .037$ with a 90% CI [.035; .038]. For this data set, the mean of correctly answered items was M = 9.25 (SD = 4.37) and the distribution of sum scores was positively-skewed.

In all data sets, 50% of the respondents were female. The mean age was 14.3 years (14.5 years for the MVI data set). The samples differed with respect to their mean sum score, because students with missing values in general had lower sum scores.

The distribution of item difficulties (item mean) is presented in Figure 17 for the MVI data set and in Figure 18 for the MVD data set. Most items had medium difficulty; few items provided medium high and medium low difficulty. However, no extreme difficulties were observed. Because the respondents with lower sum scores were deleted for MVD, the items were less difficult in this data set.

[10] For some missing values, cells were just empty and no information was provided about why the response was missing. Theses missings were treated equivalent to the time limitation missing values.

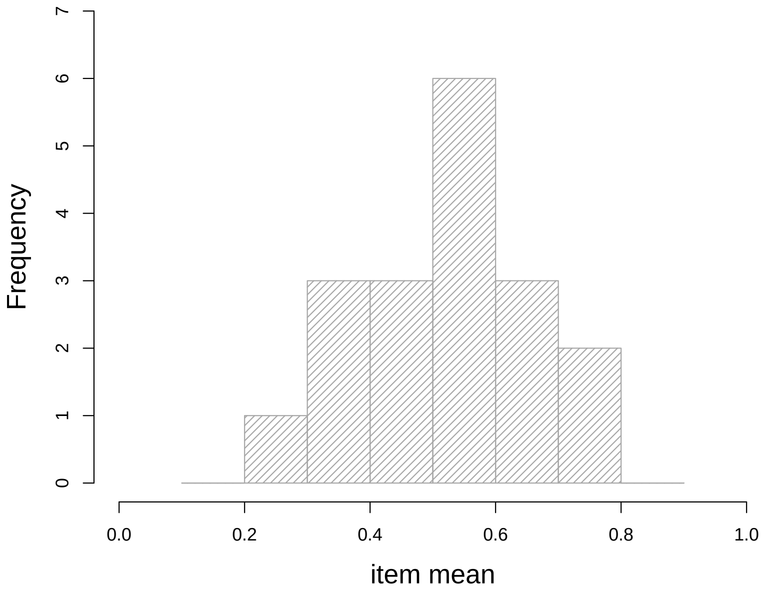


Figure 17. Histogram of item difficulties (means) for data set MVI with $N = 15\ 732$.

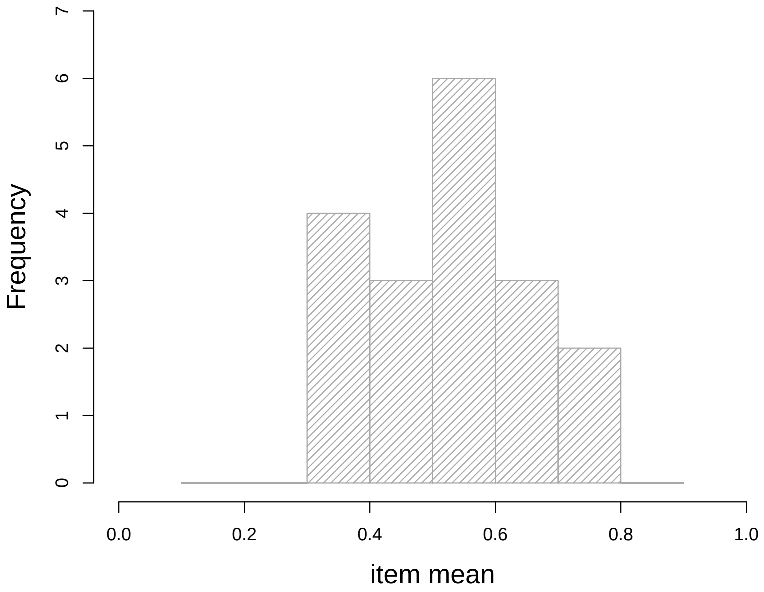


Figure 18. Histogram of item difficulties (means) for data set MVD with $N = 13\ 452$.

6.1.3 Application of Scoring Methods: Data Analysis. The scoring methods mode CBM, proportion CBM, Consensus Analysis, HOMALS, and NRM were applied to the raw TIMSS 2011 MVI and MVD data sets. Data analyses were conducted with the same R packages and scripts that were used for the simulation studies (see subsection 5.1.3). In order to compare different fixation methods for NRM, the method was applied in two versions. Firstly, the fixation of the two response options was done randomly (V2) and secondly, fixation was based on knowledge of the correct scoring key (V3). For NRM V2 random ordering was investigated further, as 100 different versions of random orders were used for each data set. For NRM V3, the correct response option was fixed to have the highest slope parameter, while one of the remaining options was randomly selected to have the lowest slope parameter.

To evaluate and compare the different scoring methods, the Pearson correlation between two ability estimates was calculated: (1) The sum of correctly answered items for each student, where the correctness of a response was evaluated on the basis of the TIMSS 2011 scoring key, and (2a) the sum of correctly answered items, where the correctness was defined with the respective scoring method mode CBM, proportion CBM or Consensus Analysis, or (2b) person parameter estimates as part of the method for HOMALS and NRM.

In order to investigate the influence of sample size and ability[11], the full data sets were moreover subdivided into different smaller data sets. For the two data sets MVD and MVI, the analysis proceeded in three steps:

1. The full data sets MVI and MVD were scored with each scoring method and the scoring methods were evaluated by the correlation between person parameter estimates as described above.

2. The full data sets MVI and MVD were divided into three subsets, by splitting the full data set into three data sets with differing mean ability levels as follows: Respondents with low ability answered four or less items correctly (mean minus one standard deviation), respondents with high ability answered more than 13 items

[11] As no items with extreme difficulties were observed and the number of items was limited to 18, difficulty and number of variables were not further investigated.

correctly (mean plus one standard deviation) and respondents with medium ability had sum scores between five and 13. For each subset, scoring methods were applied and their quality was judged using the correlation of person parameter estimates.

3. Because the sample sizes of the subsets in 2. were different, the third step included random sampling of different sample sizes of the different ability sub sets. Here, for each sample size of $N = 20$, $N = 50$, $N = 100$, $N = 200$ and $N = 1000$, five random samples were drawn from the low, medium and high ability sub set. For each data set, the scoring methods were applied and the quality of the methods was evaluated using the correlation of sum scores. For the NRM methods, only two sample sizes ($N = 200$ and $N = 1000$) were included. For Consensus Analysis, one sample size ($N = 1000$) was omitted in analogy to the simulation studies (see section 5.1.2.

6.2 Results

The results of the first analysis step are presented in Tables 33 and 34. The correlations between sum scores based on the original scoring key and the person parameter estimates of the scoring methods were highly positive (close to one) for CBM scoring methods, indicating that these scoring methods fully succeeded in identifying the correct scoring key and in providing valid person parameter estimates. As the Consensus Analysis calculation procedure was too time consuming for the full sample size, a random sample of $N = 200$ was used instead[12]. However, the correlation of ability estimates was still perfect. In contrast to the outstanding performance of consensus-based scoring methods, HOMALS had a very high negative correlation for the full data set, indicating that individuals who responded correctly to most items received low HOMALS object scores, and vice versa.

For the full data set, the NRM person parameter estimates correlated highly with sum scores based on the original scoring key when the fixation was based on theoretical assumptions (V3). When fixation was done randomly (V2), however, the results were quite

[12] The algorithm includes a factor analysis of a N × N matrix.

Table 33

Correlations of Sum Scores for the Different Scoring Methods for Data Set MVI ($N = 15,732$)

	N	Mode CBM	Prop. CBM	CA	HOMALS	NRM V3
ρ	15732	1.00	.96	1.00	$-.99$	.98
ρ_{low}	2349	$-.12$	$-.11$	$-.07$	$-.03$	.09
ρ_{medium}	10489	.76	.86	.80	$-.55$	.82
ρ_{high}	2894	1.00	.99	1.00	.08	.38

Note. Mode CBM = Mode Consensus Based Measurement; Prop. CBM = Proportion Consensus Based Measurement; CA = Consensus Analysis; HOMALS = Homogeneity with Alternating Least Squares; NRM V3 = Nominal Response Model with fixation based on correct response.

different. To get an idea of the influence of random fixation, 100 different versions of random ordering of response options were used. Across the different random orders of response categories, the mean correlation of the sum scores was $M = .21$ ($SD = .96$) for the data set MVI and $M = .33$ ($SD = .92$) for data set MVD. However, the correlations were either highly positive (61 and 67 of 100 for data set MVI and MVD, respectively) or highly negative. Hence, the mean of absolute values were $M = .98$ ($SD = .00$) for the data set MVI and $M = .98$ ($SD = .01$) for data set MVD.

In the second step of the analysis, however, dividing the sample into three sub samples of different ability offered a more differentiated view on the quality on scoring methods as presented in the lower three rows of Table 33 and Table 34. Again, for Consensus Analysis a random sample of $N = 200$ from the sub samples was used for the analysis. In the subset with low ability, all scoring methods failed to identify the correct scoring key, as the correlation between sum scores was low. However, with medium ability, the correlations were substantially higher and positive, with the exception of HOMALS for

Table 34

Correlations of Sum Scores for the Different Scoring Methods for Data Set MVD ($N = 13,452$)

	N	Mode CBM	Prop. CBM	CA	HOMALS	NRM V3
ρ	13452	1.00	.96	1.00	−.99	.98
ρ_{low}	1946	−.13	−.19	.23	−.02	.04
ρ_{medium}	8781	.76	.86	.75	−.81	.82
ρ_{high}	2725	1.00	.99	1.00	.35	.48

Note. Mode CBM = Mode Consensus Based Measurement; Prop. CBM = Proportion Consensus Based Measurement; CA = Consensus Analysis; HOMALS = Homogeneity with Alternating Least Squares; NRM V3 = Nominal Response Model with fixation based on correct response.

which the correlation was highly negative. For the highest ability subset, correlations were (close to) one for the consensus-based scoring methods. For HOMALS, correlation was of medium positive size in the high ability subset. Correlations were highest for NRM V3 in the medium ability sample, whereas they dropped in the highest ability subset.

For NRM V2, again, the correlations of sum scores were depended on random fixation, but also varied with ability. Again, 100 different versions of random fixation for the NRM were used to calculate the mean and standard deviation of the correlations of person parameter estimates and sum scores. For the low ability sub sample, the correlations were close to zero. For medium ability, correlations were high with a mean of absolute values of $M = .78$ ($SD = .03$) for data set MVI and $M = .80$ ($SD = .02$) for data set MVD. However, the correlations were either positive (53 for data set MVI and 57 for data set MVD) or negative. For high ability, correlations were of medium absolute size with a mean of absolute values of $M = .31$ ($SD = .10$) for data set MVI and $M = .39$ ($SD = .12$) for data set MVD. Again positive (49 and 52) and negative correlations were observed.

Because the sub samples with different abilities were confounded with sample size, which might also be an influencing variable on the performance of scoring methods, in the next step five random samples of different size were drawn from the different ability sub samples. For each factor level combination of sample size (five levels) and ability (three levels), the mean of the correlation across the five samples was calculated and used for graphical presentation as follows. Results of different samples are presented in Figure 19 and Figure 20 for scoring methods mode CBM, proportion CBM, Consensus Analysis as well as HOMALS.

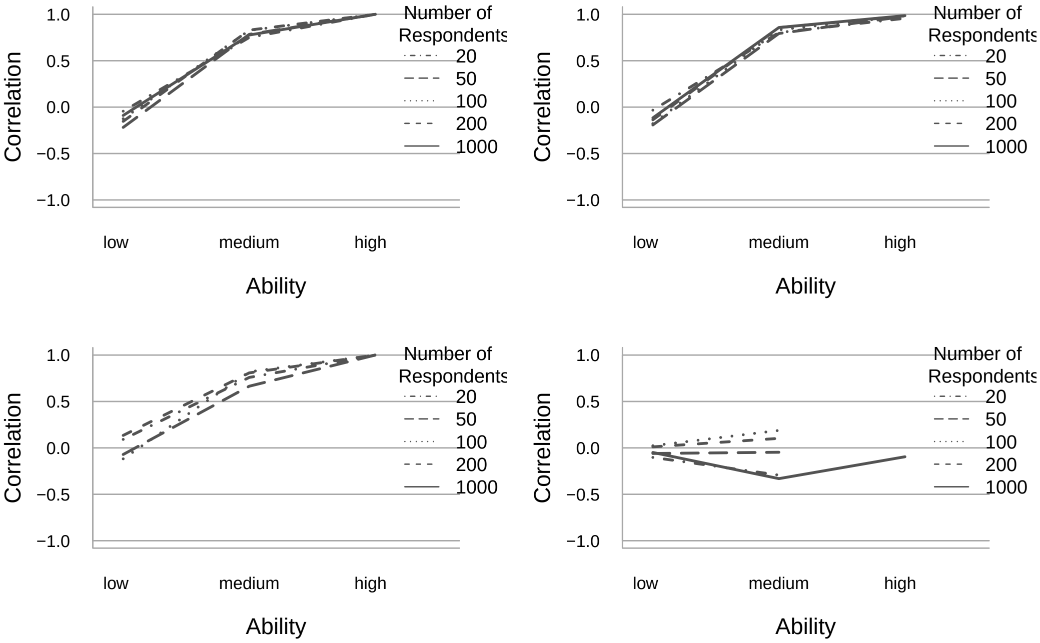


Figure 19. Interaction effects of ability and sample size on the correlation for data set MVI. Top left panel: Mode CBM; top right panel: Proportion CBM; bottom left panel: Consensus Analysis; bottom right panel: HOMALS.

For consensus-based scoring methods, the correlation did not systematically vary between different sample sizes, however, it again varied with different ability levels. Table 38 and Table 39 in Appendix G present the mean and standard deviation of the correlations for each ability level. This was highly positive for the samples with high person

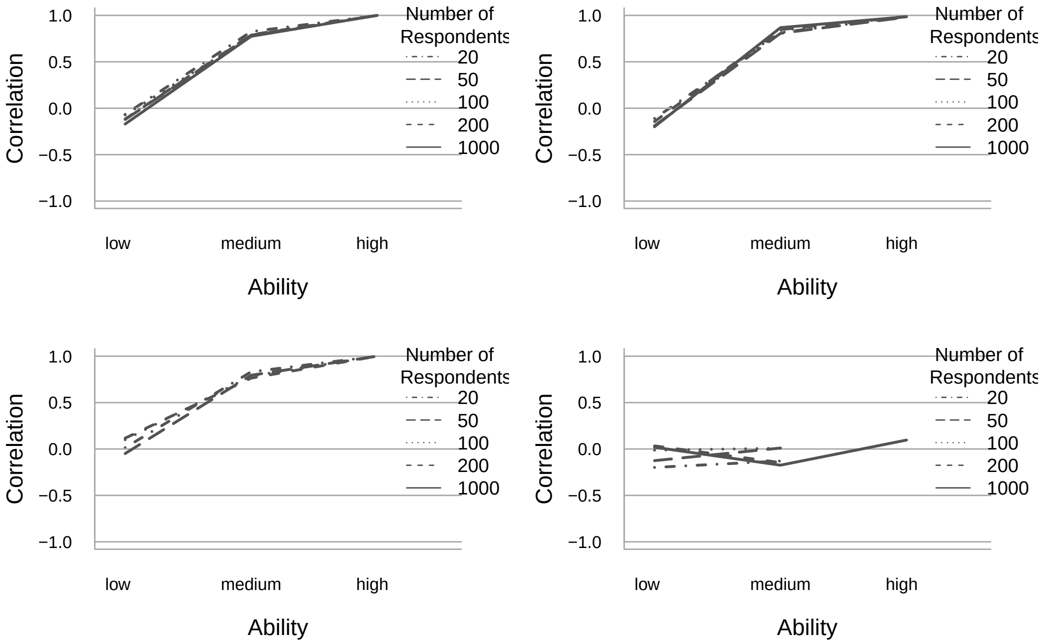


Figure 20. Interaction effects of ability and sample size on the correlation for data set MVD. Top left panel: Mode CBM; top right panel: Proportion CBM; bottom left panel: Consensus Analysis; bottom right panel: HOMALS.

parameters, that is, across all 25 drawn samples (collapsed over different sample sizes) mean correlation was (almost) $r = 1.00$ for mode CBM, proportion CBM and Consensus Analysis. For medium ability, the correlations for consensus-based scoring methods were still highly positive, whereas for low ability the correlations rapidly dropped. The different consensus-based methods performed very similarly in different data sets, however, small differences were observed. For instance, proportion CBM worked slightly better in the medium ability sample compared to mode CBM and Consensus Analysis. This is also observed for the full medium ability data set.

For different sample sizes, no systematic variation was observed for consensus-based scoring methods (see Figure 19 and Figure 20). In Table 40 and 41 in Appendix G, the mean and standard deviations of the correlations are presented for consensus-based scoring methods for each level of the factor sample size. The differences between the correlations in different factor levels are small, whereas the standard deviation in each factor level is quite large.

For HOMALS, as presented in Figure 19 and Figure 20, correlations were low and the variation was independent of the different ability levels or sample sizes. Note, that the method could not be applied for high ability sub samples with $N = 20$ and $N = 50$, as in every data set at least one variable had a zero variance. For the same reason, for $N = 100$ and $N = 200$ only one of each five data sets was used for data set MVD (1/5 for $N = 100$ and 4/5 for $N = 200$ for data set MVI). Due to these zero variances, Figure 19 and Figure 20 only include $N = 1000$ for high ability, as these data sets showed no zero variances of variables. The correlations did not vary systematically for ability, as the mean correlations (collapsed over different sample sizes) were $M = -.06$ ($SD = .23$) for low ability, $M = -.09$ ($SD = .62$) for medium ability and $M = .06$ ($SD = .22$) for high ability[13]. The variation was highest for medium ability samples, where correlations were of high absolute value, with varying positive and negative sign. In addition, no systematic variation was observed for different sample sizes, as the mean correlation (collapsed over different ability levels) varied from $-.20$ to .11 for the different sample sizes.

For the NRM V3 scoring method, results indicated that the performance of this method depended on different ability levels, but was more or less independent of the two different sample sizes. For low ability, mean correlation - collapsed over different sample sizes - was $M = .07$ ($SD = .08$) for MVI and $M = .06$ ($SD = .05$) for MVD. The correlation increased for medium ability sample with a mean of $M = .75$ ($SD = .06$) and $M = .79$ ($SD = .04$). For high ability, mean correlation was lower, with a mean correlation of $M = .22$ ($SD = .15$) and $M = .27$ ($SD = .13$). In addition, results indicated that the variation of two different sample sizes had no substantial influence on the correlation. The

[13] Results are only presented for MVD data set in the text. Results for MVI were: $M = -.04$ ($SD = .14$), $M = -.08$ ($SD = .46$) and $M = -.06$ ($SD = .10$) for low, medium and high ability. Mean and standard deviation for high ability were calculated only for ten (data set MVI) and seven (data set MVD) data sets due to zero variances in the other 18 sets.

mean correlations - collapsed over different ability levels - were $M = .33$ ($SD = .32$) as well as $M = .41$ ($SD = .36$) for $N = 200$[14] and $M = .37$ ($SD = .32$) as well as $M = .38$ ($SD = .34$) for $N = 1000$. However, the correlations were never as high as they were in the full samples of $N = 15,732$ and $N = 13,452$, indicating that sample size might influence results if varied upward.

Regarding NRM V2, 100 random fixations of response options were used for each of the 2 × 3 × 5 different data sets. The mean values and standard deviations of the correlation are presented in Figure 21 for each factor level. The mean values of correlations were not influenced by sample size or ability, however, the absolute values were influenced by ability. Hence, the standard deviation was highest for medium ability. Here, again, absolute values of the correlation were either high positive or high negative. For high ability, the correlations were of medium size, either with a negative or positive sign, whereas it was close to zero with low variation for low ability.

To summarize, the performance of consensus-based as well as NRM V3 scoring methods was very good in the full data set, however, HOMALS and NRM V2 as a scoring method did not perform well. The results did not allow determination of the impact of influencing variables on the performance of the HOMALS algorithm. The quality of NRM V2 depended on the random (good) choice of fixation, resulting in either very bad or very good results. The absolute values of the correlation between NRM person parameter estimates and sum scores depended, however, on the ability of the sample: They were highest for medium ability. In addition, consensus-based scoring methods and NRM V3 were observed to depend on the ability of the sample. Consensus-based scoring methods performed best with samples of high ability. With medium ability, results still indicate a good performance of these methods, whereas with low ability consensus-based methods did

[14] For $N = 200$ and high ability only one data set out of five could be used for NRM estimation in the data set MVD, as zero variances of variables were observed in four of five data sets. For MVI, four of five data sets with $N = 200$ and high ability were used.

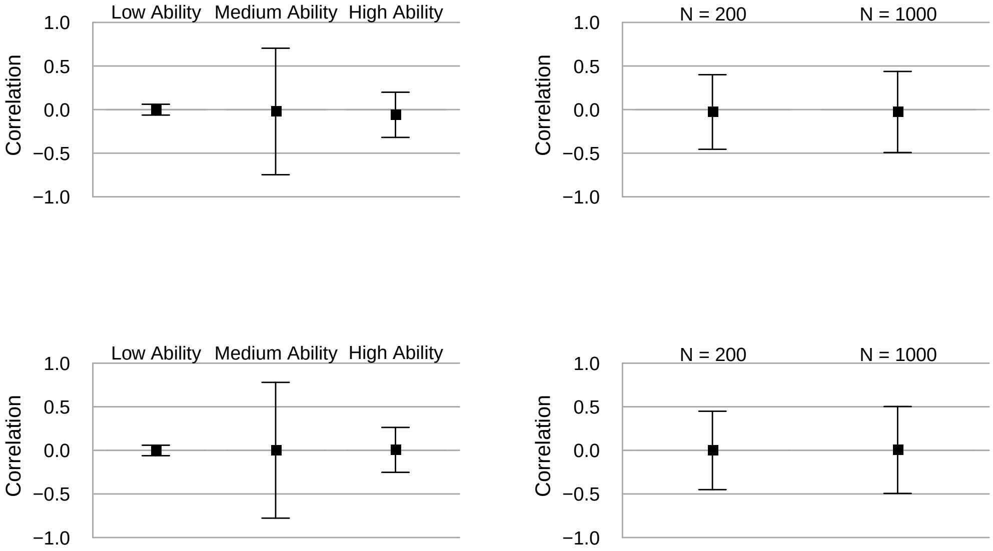


Figure 21. Mean and standard deviation of factor levels of ability and sample size for 100 random fixation versions for NRM V2. Top graphics: MVI; Bottom graphics: MVD.

not perform well. For NRM V3, performance was worst for low ability samples and best for medium ability samples. Moreover, the variation of sample size did not reveal any effect for an of the scoring methods.

7 Discussion

The results of the simulation and real-world data studies indicate that empirical scoring methods should only be applied under specific conditions and cannot be recommended for general use in ability measurement. Based on these results, recommendations differ for the methods investigated here: While some methods showed satisfying results under specific data conditions, others were not observed to function well under any of the realized conditions. This chapter will discuss and evaluate the simulation and real-world data studies separately, before integrating results, discussing limitations, and presenting conclusions.

7.1 Evaluating the Results of Simulation Studies

As the simulation studies aimed to answer two major questions, conclusions may differ with respect to these different aims. One of the aims was to investigate whether consensus-based scoring procedures could be used to identify the unknown correct scoring key in ability tests. To summarize very briefly, the results of the simulations indicate that consensus-based methods cannot be recommended for scoring in general; they are not suitable for every possible situation. However, they may be used in specific situations. Results depended acutely on the manipulated data characteristics. In particular, consensus-based methods were not able to identify the correct response option in situations in which the mean ability of the sample was low, or in which item difficulties were high. Table 35 summarizes the situation in which none of the consensus-based scoring methods showed satisfying results; that is, none of the conditions provided high ($\geq .84$) correlations between the sum scores based on scored data and true abilities. In these situations, the consensus among individuals in the worst case did not indicate the most correct, but, in contrast, the least correct response option as the seemingly *correct* response option. Thus, neither CBM nor Consensus Analysis can be recommended in these situations. The results therefore confirm concerns regarding the use of the methods, however, only in particular conditions.

Table 35

Situation with Low Performance of Consensus-based Scoring Methods for Two- and Five-categorical Data

Method	Ability	Difficulty combination
Consensus Analysis	Low	With any specified difficulty combination
	Medium	[e,m,d], [m,m,m], [d,d,e], [d,d,d]
	High	[d,d,d]
Mode CBM	Low	With any specified difficulty combination
	Medium	[e,m,d], [m,m,m], [d,d,e], [d,d,d]
	High	[e,m,d],[a] [d,d,e], [d,d,d]
Proportion CBM	Low	With any specified difficulty combination
	Medium	[e,m,d], [m,m,m], [d,d,e], [d,d,d]
	High	[d,d,e],[a] [d,d,d]

[a] No high correlations were only observed for two-categorical data.

With increasing mean ability of respondents as well as with easier items, the methods showed better results. With medium able respondents in combination with easy items and highly able respondents in combination with easy or medium difficult items, consensus-based scoring methods were more likely to succeed in identifying correct response options of items. Table 36 summarizes the situation with high ($>$ .90) and medium probability of success (.20 - .73) for five categorical data. Probability of success is defined on the basis of the relative frequencies with which the methods showed high correlations ($\geq$ .84) in the respective conditions, as presented in section 5.2.2.2. For two-categorical data, in general lower frequencies of successful situations were observed, as stated in Table 37.

Table 36

Situation with Medium and Good Performance of Consensus-based Scoring Methods for Five-categorical Data

Good performance (High probability, > .90)

Method	Ability	Difficulty combinations
Consensus Analysis	Medium	[e,e,e], [e,e,m]
	High	[e,e,e], [e,e,m], [e,m,d], [m,m,m], [d,d,e]
Mode CBM	Medium	[e,e,e]
	High	[e,e,e], [e,e,m], [m,m,m]
Proportion CBM	Medium	[e,e,e], [e,e,m]
	High	[e,e,e], [e,e,m], [m,m,m]

Medium performance (Medium probability, .20 – .73)

Method	Ability	Difficulty combinations
Mode CBM	Medium	[e,e,m]
	High	[e,m,d]
Proportion CBM	High	[e,m,d], [d,d,e]

The results summarized above indicate that the hypothesis that ability and difficulty are the most important influencing variables for the quality of consensus-based methods was fully supported. The results confirm the findings of Mohoric et al. (2010) and Barchard et al. (2013). Moreover, with higher ability, the differences in the quality of the consensus-based methods for different difficulty levels were less pronounced than with medium ability: that is, difficulty had less impact on the quality of the scoring methods for high ability respondents. The results of the simulation studies indicate that consensus-based methods can only be recommended for the particular ability and difficulty

Table 37

Situation with Medium Performance of Consensus-based Scoring Methods for Two-categorical Data

Medium performance (Medium probability, .27 − .67)

Method	Ability	Difficulty combinations
Consensus Analysis	Medium	[e,e,e], [e,e,m]
	High	[e,e,e], [e,e,m], [e,m,d], [m,m,m], [d,d,e]
Mode CBM	Medium	[e,e,e], [e,e,m]
	High	[e,e,e], [e,e,m], [m,m,m]
Proportion CBM	Medium	[e,e,e], [e,e,m]
	High	[e,e,e], [e,e,m], [e,m,d], [m,m,m]

combinations stated in the upper part of Table 36. These recommendations are based on five-categorical (GPCM) data and cannot be generalized to two-categorical 3-pl data. Moreover, it has been shown for the difficulty combination [e,e,m] that proportion CBM worked fine for medium ability samples - hence hypothesizing that this difficulty combination is the most common in EI research (see section 5.1.2.1) consensus scoring might have in fact provided valid results in EI research with mainly easy items. However, the probably more common and favorable difficulty combination [e,m,d] did not provide valid results in terms of correlations between true and reconstructed abilities. Therefore, tests with balanced item difficulties should not be scored using CBM methods and medium able samples. The results also indirectly confirm the observation that CBM does not allow the identification of difficult items (Zeidner et al., 2001, 2004).

The second, more general aim of the simulation studies was the evaluation of empirical scoring methods in general. This includes the consensus-based scoring methods, but it also includes the empirical scoring methods HOMALS and NRM, which were evaluated separately with the aim of investigating alternative methods. HOMALS did not allow the identification of the correct response options in any of the specified data conditions. The category weights did not reflect the true ordering of the response options, nor did component scores provide sufficient estimates of true ability. Hence, HOMALS - as used in the present studies - is not recommended as an empirical scoring method.

The NRM, however, did succeed in identifying the correct response option ordering in many specified data cases. While NRM with random fixation of the most correct and least correct response option could not identify the true ordering of correctness, NRM with plausible fixation showed good results. Here, the true ordering of response options was almost correctly estimated and the person parameter estimates were highly correlated with true abilities. Results thus supported the hypothesis that NRM scoring depends on the quality of fixation. However, only one instance of plausible fixation was used in the studies, in which the least correct and most correct response options were fixed with a probability of .60 and .70. Consequently, with plausible theoretical assumptions about the most correct and least correct response option, NRM can be recommended as an empirical scoring method. Although it was not explicitly investigated in the present study, using lower probabilities for fixation might most likely result in a worse performance of the NRM, as with random fixation (e.g., probability of .20 for the least and the most correct response option with five response options) the NRM was not observed to function well for scoring purposes; whereas using higher probabilities for fixation of the most and least correct response option would result in even better estimation of response option ordering and person parameter estimations. For the application of the method it is certainly of interest how the probabilities of fixation could be translated into the level of certainty or uncertainty the researcher has regarding his or her theoretical assumptions. However, the translation of probabilities into certainty levels - for instance, based on empirical evidence or the like - is not feasible. Nonetheless, the present results support the use of the NRM even with not certain theoretical ideas. A possible improvement of the NRM scoring method is proposed in section 7.4 in this chapter.

Alongside the more general aims, other specific data characteristics were hypothesized to influence the quality of the methods. The potentially influencing variables sample size and number of items were systematically varied in the present simulation studies. However, the hypotheses made about these variables were not fully supported by the results.

The assumption that Consensus Analysis shows advantages over CBM, especially with small samples sizes, was not supported by the results, as no dependence upon sample size was observed for CBM or Consensus Analysis. According to Weller (1987, 2007) Consensus Analysis - in contrast to regular consensus - can provide good results even for very small sample sizes. Sample size requirements for Consensus Analysis depend on the agreement of respondents, that is, the quality of the method is arguably high for small sample sizes when agreement among respondents is high. Weller (2007) states the sample size requirements for different levels of agreement among respondents (cultural competence) and different proportions of correctly classified items. For 99% correctly classified items and an average cultural competence of .50, a sample size of more than 30 respondents is needed, whereas with an average cultural competence of .90 only six respondents are required for the same proportion of correctly classified items. In the present studies, however, agreement among respondents was not systematically modeled in terms of the correlation between respondents or the variance in their abilities. The difference in the quality of the method dependent on sample size and agreement may not be revealed in this study because levels of agreement were not manipulated in the simulated data.

Moreover, CBM and Consensus Analysis did not show any advantages in data with a higher number of items. Therefore, hypotheses regarding the positive influence of a higher number of variables were not supported. Based on the results of Barchard et al. (2013), it was hypothesized that the correlation of true abilities and CBM-scoring abilities may increase with more items. However, in the present study the correlation between true abilities and CBM ability estimates was almost independent of the number of variables. Here, the number of items had a barely observable effect on the absolute values of correlation in the expected direction: They were slightly higher with more items. The difference between the present study and the study of Barchard et al. (2013) is most probably explained by the effect of measurement error. The reliability of test scores is higher for tests with more items and is known to influence the correlation between them (Spearman, 1904). In the respective study, the test scores based on the veridical key and based on the CBM scoring methods had Cronbach's alpha coefficients between $\alpha = .52$ and $\alpha = .98$ (Barchard et al., 2013). In the present studies, in contrast, the true abilities - which were not affected by measurement error - were known. They were correlated with scores resulting from scoring methods, which were affected by measurement error. Hence, the effects that measurement error has on the correlations are likely to be much higher in Barchard et al. (2013) study compared to the present studies, especially for lower number of items. For Consensus Analysis, the quality of the competence estimation based on the agreement among individuals was hypothesized to improve with more items. This assumption was not supported by the present results. However, as the simulation of the data was based on a model for which agreement among respondents was not a central parameter, results might be different with data based on other psychometric models.

Moreover, no differences in the quality of the NRM method were observed for different sample sizes, although one might expect the sample size to be a major determinant of the quality of estimation (De Ayala & Sava-Bolesta, 1999; DeMars, 2003). However, as only two different sample sizes were used in this study, the variation may have been too limited to reveal any effect; in particular, because the relevant studies recommended higher sample sizes for estimation, for instance $N = 1200$ for twelve items with five response options (De Ayala & Sava-Bolesta, 1999). Sample size was not varied with respect to the total number of parameters estimated or number of parameters per item, both of which influence item parameter estimation recovery (DeMars, 2003). On the other hand, the quality of the NRM scoring method with a plausible fixation depended on the number of items. Based on the results of simulation studies (DeMars, 2003), it was assumed that a reduction in the number of variables would slightly improve item parameter estimation and possibly, therefore, NRM quality as a scoring method. With regard to its quality, different directions of effects were observed, depending on the dependent variable. For the correlation of NRM person parameter estimates and true abilities, an increase was observed with more items, whereas the agreement of scoring keys (Kendall's τb) increased with less items. For the latter dependent variable, the small effect of number of variables on item parameter estimation is consistent with the results of DeMars (2003) and therefore with the hypotheses. The effect of number of items on the correlation between true and estimated abilities was not hypothesized, but might be explained as follows. The NRM item parameters were estimated with marginal ML estimation using the EM algorithm and the NRM person parameters were estimated with the EAP procedure (Bock & Aitkin, 1981; Chalmers, 2012). The different direction of effects might be explained by the different estimation procedures used for the estimation of item and person parameters. EAP person parameter estimation for IRT in general is less biased with more items (Embretson & Reise, 2000). Hence, the effect of number of items on the person parameter estimation might be the reason for the increasing correlation between ability estimates and true abilities for the NRM with more items.

Regarding the number of response categories, no clear evaluation is possible on the basis of the simulation studies, because different models were used for data generation of two- and five-categorical data. However, differences between two- and five-categorical data were observed, with the latter showing better results. Although these differences do not allow conclusions to be drawn about general difference for number of response categories, it seems plausible to conclude that the scoring methods worked better for five-categorical GPCM data than they did for two-categorical 3-pl data. Independent of the functioning of the scoring method and of the type of data, the results of the simulation studies provide no evidence that empirical scoring methods influence the one-dimensional structure of the data. For mode and proportion CBM scoring, this result is consistent with the findings of

MacCann et al. (2004), who found both CBM scoring methods to results in one-dimensional scores. For the other methods, testing the structure after scoring was conducted for the first time and results so far indicate stability in one-dimensionality.

7.2 Evaluating the Results of the Real-world Data Study

The two main aims of the research project were also central to the real-data study based on TIMSS 2011 data: the evaluation of empirical and consensus-based scoring methods. The results indicate that consensus-based scoring methods function perfectly for the full data set. Consensus Analysis and CBM methods succeeded in identifying the true correct scoring key and were therefore able to provide excellent person parameter estimates. For NRM with plausible fixation, a similar picture was observed. These results were rather surprising, given the concerns about the consensus-based scoring methods discussed in the literature. In contrast, neither HOMALS nor NRM with random fixation provided valid person parameter estimates, being partly consistent with the expectation. In particular, the influence of fixation was supported for the NRM. For NRM with random fixation, either very high positive or very high negative correlations between the person parameter estimates and the sum scores based on the TIMSS key were observed for random fixation. This result suggests that with random fixation (and a high sample size as well as sufficiently reasonable assumptions about the mean ability of the sample) either a very good or a very bad estimate of individual ability would occur. While the first case works perfectly well, the second one might at least provide valuable information. Ability measurement is not without theoretical assumptions about the ability; so even with very simple theories about the construct, it might be possible to determine whether the scoring key is reversed.

Regarding the two investigated data characteristics, ability and sample size, the stated hypotheses were partly confirmed for the real-world data study. Consistent with the hypotheses based on the relevant literature, the ability of the sample was observed to have great influence on scoring method quality. For the sub sample with a low ability, no empirical scoring method functioned well. However, the quality of the consensus-based scoring method increased with increasing ability and best results were observed with the highest ability sub sample for Consensus Analysis and CBM. Ability was not hypothesized to have an effect on NRM scoring, however, best results were observed for the medium ability sub samples. Even if sample size was controlled by randomly drawing samples of different sizes, the results remained stable. Sample size was not observed to affect the quality of the scoring methods. However, it can be assumed that sample size has an effect with NRM, as the correlations between person parameter estimates for $N = 200$ and $N = 1000$ were never as high as they were for the full sample.

Neither the number of items nor their difficulty were systematically varied for the TIMSS data set, as the number of items was restricted to 18. Moreover it was not possible to manipulate the number of response categories for TIMSS data. Therefore, the influence of the difficulty of items, number of items and number of response categories (or different response models) was not investigated with real-world data.

7.3 Integrating the Results

The results of both the simulation studies and the TIMSS study have shown that ability and difficulty are the major variables influencing the quality of consensus-based scoring methods. So far, only two other studies have systematically investigated CBM methods, whereas for Consensus Analysis no studies investigating the scoring quality of ability tests have been published. Mohoric et al. (2010) and Barchard et al. (2013) have shown that ability and difficulty are important influencing variables for CBM methods. The present studies not only confirm these results, they also highlight the influence of these variables under controlled conditions and in a real-data situation of high quality standards. A major advantage of the present simulation studies is that the true correct response key is known, as it is defined by the simulation model. The scoring key of the mathematical TIMSS items is based on mathematical operations for which one true correct response option can be unequivocally defined. Consequently, the results of the present studies replicate and extend previous findings and emphasize the impact of the two influencing variables ability of respondents and difficulty of items.

Neither the simulation studies nor the real-world data study indicate that Mode and Proportion CBM differ systematically in their performance. Although differences were observed, none of the CBM methods consistently showed substantially better results. For instance, Mode CBM scoring showed small advantages for the full TIMSS data set compared to Proportion CBM, whereas for the medium ability sub sample of TIMSS data, Proportion CBM was observed to function better. In the simulation studies, Proportion CBM provided slightly better results. However, the differences between the CBM methods were not substantial. Previous studies came to conflicting conclusions regarding the choice between Mode and Proportion CBM. For instance, Mohoric et al. (2010) observed higher correlation for Mode CBM based sum scores with scores based on a veridical scoring key. The present studies indicate that the quality of CBM scores depends less on the choice of CBM method than on sample and item characteristics. However, as Proportion CBM showed slightly better results, it might be preferred.

The quality of Consensus Analysis did not differ substantially from CBM scoring methods in both the simulation studies and real-world study. However, small differences were observed in the simulation studies, as Consensus Analysis performed better with the difficulty combination [d,d,e] compared to the CBM methods. Hence, if more difficult items are expected, one might prefer to use Consensus Analysis. In the TIMSS data set, however, no consistent differences between Consensus Analysis and CBM methods were observed.

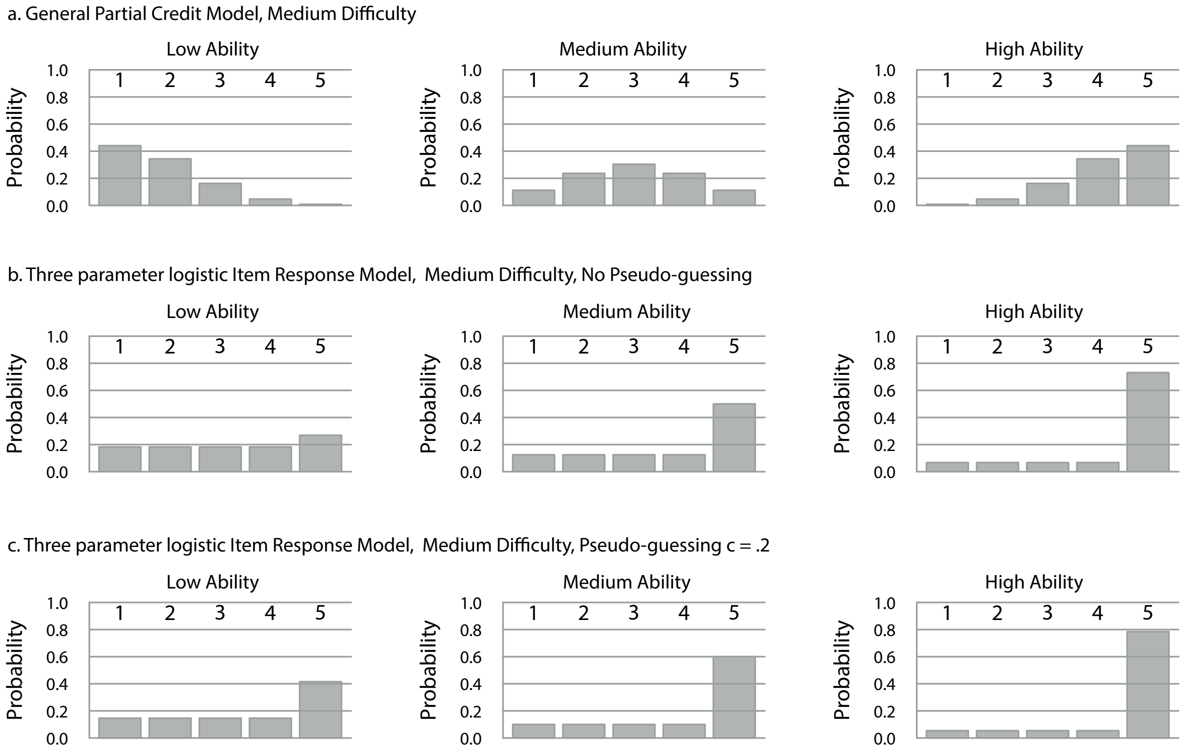
The simulation and the real-world data studies agree with regard to the importance of fixation for NRM scoring. Plausible assumptions about the most correct as well as least correct response option are important for the performance of the method. In addition, both studies are consistent in their results for HOMALS: This method did not show good results.

However, the results of the simulation and real-world data studies differed in one respect: The results of consensus-based scoring methods were substantially better in the real-world data study than in the simulation studies. Using TIMSS data, the scoring keys according to consensus-based scoring methods were identical to the original TIMSS scoring key. In the simulation studies, however, the scoring keys were only identical in the case of a very high mean ability of respondents or low item difficulty. For the TIMSS data set, neither a high mean ability of respondents nor a low item difficulty seem to be the reason for the good performance of consensus-based scoring methods. The TIMSS sample was representative for eighth-grade students. The items were mostly of medium difficulty (both according to classical test theory and IRT difficulty estimates) and no extreme difficulties were observed. The differences in the results might instead point to the influence of different psychometric models underlying the simulation and the real-world data studies. These potential differences might constrain the conclusions with regard to external validity, that is, to what degree the conclusions based on the simulation studies might be generalizable to all relevant real-world situations (Skrondal, 2000).

In the simulation studies, the five-categorical data was simulated on the basis of the GPCM to incorporate the view that response options can be ordered regarding their correctness, that is, that each response option provides a different degree of information about the underlying ability. However, the TIMSS data was modeled with a 3-pl IRT model which assumes that only one response option is correct. Here, items were constructed to have only one correct response and three incorrect distractors. The construction of the items most likely aimed at developing distractors that were definitely incorrect, for instance, by eliminating those that were positively correlated to the total score. Moreover, unlike the simulation studies, the TIMSS data model included respondents' guessing.

On the basis of the mathematical model of the 3-pl and the GPCM data, one can illustrate how differences between models affect response probabilities and therefore might affect consensus-based scoring methods. The GPC model which incorporates the idea that response options can be ordered for correctness assumes that the probability of endorsing a more correct response option increases with increasing ability. A test taker has the highest probability of endorsing the response option for which the difficulty corresponds to his or her ability, as illustrated in Figure 22a for the three different ability levels used in the present simulation studies as well as an item for which the distribution of the difficulty parameters of the response categories represents a medium difficult item (compare section 5.1.2.1). Therefor, the resulting item response distribution has highest endorsement rates for the option for which the difficulty corresponds to the mean sample ability. However, response options with a difficulty different from the mean ability are endorsed less frequently. This concept of response behavior is reasonable for many example items for the measurement of EI or SI. For instance, the response option of the items in Table 1 for measuring judgment in social situation, in Table 2 for measuring social insight, in Table 3 and in Table 4 for measuring emotion management include behavior which cannot be categorized into correct or incorrect options, but which more appropriately can be ordered with respect to correctness (or effectiveness). For these items, the GPCM might adequately model response behavior.

For items modeled with a 3-pl model, the idea of response behavior is quite different. The 3-pl model assumes that the probability of endorsing the correct response option increases with increasing ability, whereupon only one response option is correct and all other response options are incorrect. With low ability, the probability of endorsing an incorrect response option is higher than that of endorsing the correct response option; however, the probability of endorsing an incorrect response option is assumed to be equally distributed among the four distractors (at least under the assumption that each distractor is equally attractive, see Figure 22b). The higher the ability, the more likely is the endorsement of the correct response option (which is here named response option five). With an additional pseudo guessing parameter of $c = .20$, the model-based probability for endorsement of the correct response option is even higher (see Figure 22c). The response behavior might be reasonable for SI and EI tests, for instance see the item measuring observation of human behavior in Table 1 and the item measuring emotional understanding in Table 4. More importantly, the 3-pl model might be most appropriate to model the response behavior for items measuring classical intelligence, for instance figural matrices, or mathematical competence (see Chapter 6). The 3-pl is probably the most appropriate model for these items, because the item construction in classical intelligence tests usually aimed to eliminate distractors which positively correlate with test scores (Lienert & Raatz, 1998). Partly correct responses might however be possible for some type of tasks. For example, for figural matrices one can construct item response options which are partly correct when only a part of the construction rules are applied. For these possible distractor a partial credit might be suitable; but it is more likely that the distractor is eliminated in a construction phase - or not modeled as partly correct.


a. General Partial Credit Model, Medium Difficulty

b. Three parameter logistic Item Response Model, Medium Difficulty, No Pseudo-guessing

c. Three parameter logistic Item Response Model, Medium Difficulty, Pseudo-guessing c = .2

Figure 22. Response probabilities for five-categorical data with GPCM and 3pl model.

The choice of a response behavior model influences the hypotheses regarding the performance of consensus-based scoring methods: In the case of multi-categorical 3-pl data with equal attractiveness of distractors, consensus-based scoring methods are assumed to show better results than for GPCM data. This assumption could possibly explain differences in the results of the real-world data and the simulation studies. However, as argued above, both the 3-pl model and the GPCM are important for the measurement of abilities, and both can be assumed to underlie ability data. Therefore, the conclusions of the simulation studies and the real-world data study are both important. However, the differences in the results motivate the posing of further research questions. To systematically investigate the empirical scoring methods for multi-categorical data, simulation studies with multi-categorical 3-pl data (with equally attractive distractors) might be conducted. One could also model the attractiveness of distractors, for instance using a nested logit model (Suh & Bolt, 2011) which allows to model the probability of a correct response with a 2-pl or 3-pl IRT model whereas the probability of an incorrect response - and hence the probability of selecting one of the distractors - is modeled by a version of the NRM (Bock, 1972). Moreover, these simulation studies should also include a variation of the pseudo-guessing parameter which was held constant in the present studies. On the basis of the illustrated effect of pseudo-guessing (see Figure 22), the investigation of this parameter is important as guessing might influence the quality of CBM methods (assuming that distractors are equally attractive).

The external validity of the results of the present simulation studies, however, could be investigated with data for which the GPCM idea seems more plausible, that is, where partly correct responses can be identified. SI and EI tests assumedly include partly correct responses; however, as the assumptions about correctness are not based on sound theories, data from these tests is not suitable for external validation. Test of classical intelligence, where the correctness of response options is most often less arguable, however, do often not include partly correct responses, even if it would be possible to design partly correct responses. For this reason the question which data might be appropriate for the purpose of external validation of the GPCM simulation study results remains unanswered.

To summarize, the differences between the real-world data study and the simulation studies indicate that response behavior for different types of items and appropriate psychometric models for these behavior may show effects in the quality of empirical scoring methods. Scoring methods worked better with real data than they did in simulation studies. Further research is needed to investigate the influence of the underlying model and different types of items. Moreover, the question what psychometric model is more appropriate for the measurement of SI and EI should be discussed further. Thus, does the measurement of EI and SI include items for which only one response option is correct and the other response options are clearly incorrect? Or does the measurement include items for which the response options can be ordered with respect to their correctness? However, despite the differences between the simulation and real-world data studies, the results agree with regard to the influence of the independent variables ability (and difficulty) on consensus-based scoring methods, as well as the influence of the fixation for NRM based scoring.

7.4 Limitations and Future Studies

Limited time and computer resources prevented investigation of every conceivably interesting influencing variable in the simulation studies. Hence, the generalization of results of these studies only applies to data that can be modeled with a similar psychometric model. Data characteristics that were not considered include multi-dimensionality, pseudo-guessing, different response scales, different difficulty combinations, and other distributions of ability. Further investigation of these data characteristics - in addition to the further research questions stated in the last section - would contribute to a more detailed understanding of empirical scoring methods. For instance, item response data based on two abilities might lead to different results regarding the quality of scoring methods, as the methods might be hypothesized to be influenced by multi-dimensionality because one-dimensionality is a common model assumption, for example for the NRM scoring method as it has been used here.

The investigation of models where higher ability is explicitly modeled as higher agreement between respondents would contribute to the understanding of consensus-based scoring methods. Agreement among respondents is most frequently conceptualized as a positive correlation between respondents or low variance (Davis-Stober et al., 2014). The idea of higher agreement between more able respondents is not necessarily aligned with the item response models which have been used in the present studies to simulate data. In simulated data according to the GPCM, agreement among respondents does not depend on ability levels: The model does not state that agreement for high ability individuals is higher than for low ability individuals. The assumption that more able respondents show higher agreement is however implicitly used in the 3-pl model for multi-categorical data. If all distractors are assumed to be equally attractive and truly incorrect, so that minimal knowledge results in pure guessing, then the more able respondents will agree on one response option whereas the less able respondents will disagree. However, with some distractors more attractive than others, the assumption of equally distributed disagreement might be untenable. Moreover, variance in abilities might influence agreement, in particular for GPCM data. The more similar the abilities of respondents, the more they will agree. However, if they differ highly in ability, they will tend to show lower agreement. Therefore the variance of abilities in the sample might influence the performance of the method. Especially for Consensus Analysis, lower variance and correspondingly higher agreement might show effects.

Additional limitations concern the Consensus Analysis estimation method for five-categorical data. The algorithm apparently identified response options as most correct that were not chosen by any respondent, resulting in zero variance (i.e., all respondents received a zero score for this item) of one or more variables after scoring[15]. The supposed algorithm failure only systematically affected CFAs, as data sets with zero variance after

[15] For these data sets, no zero variances were observed before scoring (see paragraph 5.2.1.3)

scoring could not be used for factor analyses. The failure of the algorithm was evidently observed more frequently in situations in which the difficulty levels of items were unequal. However, the zero variance did not influence the quality of the scoring methods in terms of the dependent variables correlation between true and estimated abilities and Euclidean distance between scoring keys, as Consensus Analysis did not show poorer results in these conditions compared to equal difficulty conditions and compared to CBM methods; in fact, Consensus Analysis provided more high correlations in mixed difficulty combinations than CBM methods. Presumably, the two dependent variables could not reveal these situations and the failure was not systematic, that is, it did not occur in every data set and not every variable in the affected data sets was observed to have zero variance. For instance, the correlation between true and estimated abilities might not be affected if the incorrect response was chosen by all respondents for one variable. The reasons for the algorithm failure are probably violations of assumptions. Different item difficulties represent a violation of the assumption of equal item difficulties (see subsection 3.1.1). Moreover, the assumption of common truth (see subsection 3.1.1) may also be violated, as the competencies of all respondents may not by accounted for by only one factor in all conditions of simulated data and this precondition was not tested before the application of the algorithm. Note that although the simulated data was one-dimensional, the dimensionality of agreement among respondents was not tested. The Consensus Analysis factor-analysis based algorithm should only be used with great care, checking model assumptions before using the algorithm. Other Consensus Analysis algorithm might be more robust in these situations and might be preferred.

In particular, different estimators could be used for Consensus Analysis, which might result in improvements. The factor-analysis based estimation procedures used in the present studies has been criticized because it only allows estimation of the most restricted model (Aßfalg & Erdfelder, 2012a; Karabatsos & Batchelder, 2003; Oravecz et al., 2014). As no published systematic simulation study is available to compare different estimation methods, the possible benefits of estimations procedures ML, MCMC and HBM estimation is an open research question. However, these methods would allow the estimation of additional parameters, such as difficulty of items or guessing of respondents. Hence, other estimation procedures might improve the performance of Consensus Analysis as a scoring method. However, only some of the alternative estimation methods can handle five-categorical data.

Regarding the confirmation of the one-dimensional structure, the use of the estimator ML for Proportion CBM and HOMALS was a problematic choice which limit conclusions. These scores were not necessarily multivariate-normally distributed and this precondition was not tested. Although, it may have been possible to transform data into integer values, for example based on ranks, we decided to remain with original values given by the scoring method, as those values will most likely be used in practice for further analysis. Because of the high risk of non-normal distributions, especially for some conditions, other estimators, such as ML with robust standard errors and Satorra-Bentler corrections (MLM, Satorra & Bentler, 1994), might have been a better estimator choice. The present results based on the ML estimator need to be interpreted carefully and future studies should take account of the problems observed in the present simulation studies.

The NRM scoring method might be improved for future applications by using a exploratory step-wise procedure, in which the slope parameters of the response options are estimated repeatedly, adapting the parameters on the basis of previous results. As Chalmers (2012) states, greater slope parameters values may indicate a poor choice in the fixation of the lowest and highest categories. After model estimation, these values could be identified and the fixation might be changed on the basis of those empirical results in combination with theoretical assumptions. However, this exploratory procedure requires cross validation with a different sample. As the estimation of the NRM model itself requires a large sample (De Ayala & Sava-Bolesta, 1999), cross validation raises the requirement of even higher sample sizes. The advantages and disadvantages of this procedure have not been discussed for empirical scoring and should be investigated further.

7.5 Conclusions and Recommendations

The problem of scoring is one of the biggest challenges for the measurement of social-cognitive abilities such as EI and SI. Consensus-based scoring methods have been frequently used for the measurement of EI and SI and are subject to fundamental criticism of the related measurement approaches. The present studies investigated the quality of consensus-based and empirical scoring methods. They are the first systematic comparative studies of consensus-based scoring methods using the controlled, highly internal valid conditions of simulation studies as well as high quality real-world data from ability measurement. Because the scoring keys were known in the present simulation studies, the evaluation of the empirical scoring methods was based on the true correctness of response options. For the first time, therefore, the consensus-based scoring methods were investigated using veridical criteria. In addition, the present studies are the first to systematically investigate alternative empirical scoring methods and to aim at enlarging the possible pool of methods. They therefore contribute to the improvement and development of measurement approaches for new ability constructs.

The present thesis addressed two major points: The related rationale of the consensus-based scoring methods has not been sufficiently discussed in the past, nor have these methods been systematically investigated. Although researchers pointed out that scoring of EI and SI tests does not meet the standards of scoring in classical intelligence tests, the literature review has revealed that scoring has also not been sufficiently addressed for the measurement of classical intelligence. The rationales for deciding which response option is the correct one have been widely neglected in intelligence research. However, as one of the rationales for evaluation of response options, agreement among several respondents or mean response has been proposed as an indicator of the correctness, not only for the measurement of new ability constructs but also for the measurement of classical ability constructs (Guttman & Levy, 1991; Jensen, 1998; Nevo, 1993; E. L. Thorndike et al., 1926). However, from a theoretical point of view, there is no clear link between knowledge (correctness) and general consensus as a response behavior. Nonetheless, the consensus of respondents with high ability has consistently - even in philosophical literature - been argued to indicate correctness, with high ability being a key precondition for consensus to indicate correctness.

In the literature on Consensus Analysis, it has been argued that individuals with higher ability show higher agreement, whereas individuals with lower ability show less agreement (e.g., as they guess) (Batchelder & Romney, 1988). However, this assumption has not been discussed more thoroughly in relevant literature on the theoretical foundation of consensus. It may be assumed that this idea is dependent upon nongeneric characteristics in ability measurement, in particular the type of question and the underlying item response model. If only one response option is correct and the distractors are both equally attractive and false (i.e., 3-pl data), individuals with higher ability should show higher agreement. If response options can be ordered with respect to their correctness (i.e., GPCM data), as it remains an exception for many SI and EI items, respondents with higher ability might not necessarily shower higher agreement compared to individuals with lower abilities. Also here is valid that when guessing occurs for individuals with very low ability, individuals with lower ability could show less agreement compared to respondents with higher ability. The results of the present studies have raised the question of the effect of the item type and the underlying psychometric model on the quality of empirical scoring methods.

Likewise, the empirical investigation of consensus-based scoring methods and other suggested alternatives shows that empirical scoring can lead to wrong scoring keys for the measurement of abilities for both type of data (i.e., in the simulation studies and in the real-world data study). Ability of respondents as one of the key preconditions for consensus to indicate truth is strongly supported by the results. Hence, consensus may provide information about the correct response if the respondents have high abilities. Information about the construct is also needed for one of the investigated alternative scoring methods, the NRM. Accordingly, the results of the simulation studies indicate that empirical scoring methods should not be used without any theoretical knowledge about the ability to be measured. Knowledge about the construct is necessary because the good performance of scoring methods either requires assumptions about the ability of the tested individuals or plausible assumptions about the most correct and least correct response option. Knowledge of ability as a precondition for using empirical scoring methods is especially critical for those areas in which these methods have been or might be used, because adequately elaborated theories are lacking (Maul, 2012b).

The findings hinge on the longstanding problem of psychological measurement and research: the "general lack of detailed theory" (Borsboom et al., 2004, p. 1064) in psychology. The lack of concepts and rationale for scoring rules, as well as the finding that empirical scoring methods might be misleading without the background of theoretical knowledge, are examples of this problem. The lack of theory results in limitations of validity of the measurement approaches, as it remains unclear how and why test behavior allows inferences about underlying ability. The lack of theory also results in difficulties defining experts in these areas. As Borsboom et al. (2004) states:

> The reason for this is that researchers expect to get an answer to the question of what the test measures, without having a hypothesis on how the test works. If one attempts to sidestep the most important part of test behavior, which is what happens between item administration and item response, then one will find no clarity in tables of correlation coefficients. No amount of empirical data can fill a theoretical gap. (Borsboom et al., 2004, p. 1068)

However, this conclusion is anything but new, as it was addressed in early discussions about EI measurement (Roberts et al., 2001; Zeidner et al., 2001). A thorough conceptualization of EI has in fact not been published until today and, clearly, more detailed theories of EI are required. For SI, theories have been elaborated in more detail (Conzelmann et al., 2013; Weis & Süß, 2005; Weis, 2008), integrating SI into structure models of intelligence. For the measurement of social memory and social perception, Weis (2008) used typical intelligence tests rationales for defining test performance, for instance using reaction times or recall of information. For social understanding, however, scoring was provided by targets, still resulting in challenges due to this particular empirical scoring method. Recently, attempts were also made to integrate EI into the structure and facet models of intelligence (MacCann, Joseph, Newman, & Roberts, 2014; Mayer, Caruso, & Salovey, 2016). MacCann et al. (2014) provided results supporting the idea that EI can be integrated into the theory of primary mental abilities, with EI representing an ability distinct from, but correlated with crystallized and fluid intelligence. Recently, Mayer et al. (2016) took up the idea and presented a more refined four branch model of EI. They also started to integrate their ideas about problem solving in the area of emotions with facet models of intelligence, including different operators (e.g. knowing, recognizing) or contents (e.g., emotional expression, situational aspects). However, Mayer et al. (2016) did not address scoring issues from the theoretical point. These studies can at best be seen as a first step in a more thorough theory development aiming to provide more theoretical input for either identifying experts or for drawing hypotheses about the correctness of response options. Structure models of intelligence as well as intelligence measurement, however, can in part provide valuable information for scoring purposes, as it has been used by Seidel (2007) and Weis (2008). For instance, based on a theoretical definition of emotional understanding as reasoning in the domain of emotions, a test design was developed which incorporates the idea of other reasoning tests (Hellwig, 2016; Schulze & Jobmann, 2016). Here, information about individual behavior is given which unequivocally defines correct responses, and the test takers have to understand the provided information and induce correct predictions about the behavior. This and similar developments are important for the progress in research on EI and SI. Beyond that, theory development for new intelligences requires idiosyncratic theories about how areas on the latent social-cognitive abilities are related to test behavior.

Moreover, the development of measurement approaches with empirical scoring methods should be maintained under the specific conditions mentioned above. The real-world data study provided some evidence that consensus-based scoring method might also work with non-expert samples and an item sample which includes difficult items (however, not with very low ability). Therefore, it is hypothesized that consensus-based scoring methods might better work with 3-pl multi-categorical items. If items measuring EI and SI are suitable for 3-pl models, which has already been shown for a brief version of the STEU (Allen, Weissman, Hellwig, MacCann, & Roberts, 2014), consensus-based scoring procedures might provide some valuable information with medium ability samples. If, however, response behavior for EI and SI items is better modeled by GPCM, only consensus of high ability samples might provide information about the correctness of response options. In addition, the NRM has shown to be of high value when plausible hypotheses about the most and least correct response options exist. Also it might be possible to adapt the hypotheses using an exploratory procedures based on the results of NRM item parameter estimation (Chalmers, 2012). When using empirical scoring methods - that is, NRM, CBM or Consensus Analysis - the following recommendations are derived based on the results of the present studies:

- For the application of CBM scoring methods, high ability samples are recommended in combination with medium difficult and easy items. That is, expert scoring should be preferred to group consensus scoring based on a representative sample, because in the latter case only easy items can be used. For Consensus Analysis, however, more difficult items appear manageable.
- NRM should only be used with some knowledge of most correct and least correct response option. In addition, the small effects of number of variables, difficulty and ability (as well as their interaction) give reasons to define some limiting factors for

the application of the scoring method over and above fixation. Hence, using NRM models with uncertain, but reasonable (e.g. theory-based) fixation, one should avoid situations in which low ability and mostly difficult items are used.

The present studies have contributed to the challenge of scoring for the measurement of new ability constructs. The results highlight the importance of more detailed assumptions about the underlying ability. Given these assumption, some of the empirical scoring methods can be validly applied. Future research might investigate the influence of item type and psychometric model on the quality of scoring as well as possible improvements and combinations of the methods.

8 References

Allen, V., Weissman, A., Hellwig, S., MacCann, C., & Roberts, R. D. (2014). Development of the situational test of emotional understanding Ű brief (STEU-B) using item response theory. *Personality and Individual Differences*, *65*, 3-7. doi: 10.1016/j.paid.2014.01.051

Alvarado, N. (1996). Congruence of meaning between facial expressions of emotion and selected emotion terms. *Motivation and Emotion*, *20*, 33–61. doi: 10.1007/BF02251006

Alvarado, N., & Jameson, K. (1996). New findings on the contempt expression. *Cognition and Emotion*, *10*, 379–408. doi: 10.1080/026999396380196

Alvarado, N., & Jameson, K. A. (2002). Varieties of anger: The relation between emotion terms and components of anger expressions. *Motivation and Emotion*, *26*, 153–182. doi: 10.1023/A:1019815402873

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*, 151–181. doi: 10.1007/s11336-013-9382-9

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. doi: 10.1007/BF02293814

Arikan, S. (2015). Construct validity of TIMSS 2011 mathematics cognitive domains for turkish students. *International Online Journal of Educational Sciences*, *7*, 29–44. doi: 10.15345/iojes.2015.01.003

Asch, S. E. (1955). Opionions and social pressure. *Scientific American*, *193*, 31–35. doi: 10.1038/scientificamerican1155-31

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against unanimous majority. *Psychological Monographs*, *70*, 1–70. doi: 10.1037/h0093718

Aßfalg, A., & Erdfelder, E. (2012a). CAML maximum likelihood consensus analysis. *Behavior Research*, *44*, 189–201. doi: 10.3758/s13428-011-0138-0

Aßfalg, A., & Erdfelder, E. (2012b). *Consensus analysis: A comparison of methods.*

(Unpublished manuscript, University of Mannheim, Germany)

Austin, E. J. (2010). Measurement of ability emotional intelligence: Results for two new tests. *British Journal of Psychology*, *101*, 563–578. doi: 10.1348/000712609X474370

Barchard, K. A., Hensley, S., & Anderson, E. (2013). When proportion consensus scoring works. *Personality and Individual Differences*, *55*, 14–18. doi: 10.1016/j.paid.2013.01.017

Barchard, K. A., & Russell, J. A. (2006). Bias in consensus scoring, with examples from ability emotional intelligence tests. *Psicothema*, *18*, 49–54.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*, 71–92. doi: 10.1007/BF02294195

Baumgarten, M., Süß, H.-M., & Weis, S. (2015). The cue is the key: The relevance of cues and contextual information in the social understanding tasks of the Magdeburg Test of Social Intelligence. *European Journal of Psychological Assessment*, *31*, 38–44. doi: 10.1027/1015-5759/a000204

Bechtoldt, M. N. (2008). Emotional intelligence, professional qualifications, and psychologists' need for gender research. In N. C. Karafyllis & G. Ulshöfer (Eds.), *Sexualized brains: Scientific modeling of emotional intelligence from a cultural perspective* (pp. 117–130). Cambridge, MA: The MIT Press.

Becker, N., Preckel, F., Karbach, J., Raffel, N., & Spinath, F. M. (2014). Die Matrizenkonstruktionsaufgabe: Validierung eines distraktorfreien Aufgabenformats zur Vorgabe figuraler Matrizen [the matrices construction task: Validation of a distractor-free task format for figural matrices]. *Diagnostica*, *61*. doi: 10.1026/0012-1924/a000111

Beckermann, A. (1972). Die realistischen Voraussetzungen der Konsenstheorie von J. Habermas [The realist preconditions of the consensus theory of J. Habermas]. *Zeitschrift für Allgemeine Wissenschaftstheorie*, *3*, 63–80. doi: 10.1007/BF01800820

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006).

Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14*, 223–235. doi: 10.1111/j.1468-2389.2006.00345.x

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test score* (pp. 397–479). Reading, MA: The MIT Press.

Blickle, G., Momm, T., Liu, Y., Witzki, A., & Steinmayr, R. (2011). Construct validation of the Test of Emotional Intelligence (TEMINT): A two-study investigation. *European Journal of Psychological Assessment*, *27*, 282-289. doi: 10.1027/1015-5759/a000075

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. doi: 10.1007/BF02291411

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. doi: 10.1007/BF02293801

Borsboom, D. (2009). *Measuring the mind: Conceptual issues in contemporary psychometrics.* doi: 10.1017/CBO9780511490026

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Brackett, M. A., & Mayer, J. D. (2003). Convergent, discriminant, and incremental validity of competing measures of emotional intelligence. *Personality and Social Psychology Bulletin*, *29*, 1147–1158. doi: 10.1177/0146167203254596

Brody, N. (2004). What cognitive intelligence is and what emotional intelligence is not. *Psychological Inquiry*, *15*(3), 234-238.

Broom, M. E. (1928). A note on the validity of a test of social intelligence. *Journal of Applied Psychology*, *12*(4), 426–428.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* doi: 10.1017/CBO9780511571312

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*, 1–22. doi: 10.1037/h0046743

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29. doi: 10.18637/jss.v048.i06

Chapin, F. S. (1942). Preliminary standardization of a social insight scale. *American Sociological Review*, *7*, 214–225. doi: 10.2307/2085176

Chapin, F. S. (1960). *The chapin social insight test.* Redwood City, CA: Mind Garden.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*, 83–117. doi: 10.1111/j.1744-6570.2009.01163.x

Ciarrochi, J. V., Chan, A. Y., & Caputi, P. (2000). A critical evaluation of the emotional intelligence construct. *Personality and Individual Differences*, *28*, 539–561. doi: 10.1016/S0191-8869(99)00119-1

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559–583. doi: 10.1016/0169-2070(89)90012-5

Colberg, M., Nester, M. A., & Trattner, M. H. (1985). Convergence of the inductive and deductive models in the measurement of reasoning abilities. *Journal of Applied Psychology*, *70*, 681–694. doi: 10.1037/0021-9010.70.4.681

Comrey, A. L. (1962). The minimum residual method of factor analysis. *Psychological Reports*, *11*, 15–18. doi: 10.2466/pr0.1962.11.1.15

Conzelmann, K., Weis, S., & Süß, H.-M. (2013). New findings about social intelligence: Development and application of the Magdeburg Test of Social Intelligence (MTSI). *Journal of Individual Differences*, *34*, 119–137. doi: 10.1027/1614-0001/a000106

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York, NY:

Harper Collins Publishers.

Davey, T., Nering, M. L., & Thompson, T. (1997). Realistic simulation of item response data. *ACT Research Report Series*, *97*(4), 11–37.

Davies, M., Stankov, L., & Roberts, R. D. (1998). Emotional intelligence: In search of an elusive construct. *Journal of Personality and Social Psychology*, *75*, 989–1015. doi: 10.1037/0022-3514.75.4.989

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*, 79–101. doi: 10.1037/dec0000004

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement*, *16*, 327–343. doi: 10.1177/014662169201600403

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement*, *23*, 3–19. doi: 10.1177/01466219922031130

De Leeuw, J. (1984). The gifi system of nonlinear multivariate analysis. In E. Diday (Ed.), *Data analysis and informatics, III* (pp. 415–424). Amsterdam, Netherlands: Elsevier.

De Leeuw, J., & Mair, P. (2009). Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software*, *31*, 1–20. doi: 10.18637/jss.v031.i04

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology*, *63*, 453–482. doi: 10.1146/annurev-psych-120710-100353

DeMars, C. E. (2003). Sample size and recovery of nominal response model item parameters. *Applied Psychological Measurement*, *27*, 275–288. doi: 10.1177/0146621603027004003

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influence upon individual judgment. *Journal of Abnormal and Social Psychology*, *51*, 629–636. doi: 10.1037/h0046408

Devitt, M. (2006). Scientific realism. In P. Greenough & M. P. Lynch (Eds.), *Truth and* *realism* (pp. 100–124). doi: 10.1093/acprof:oso/9780199288878.003.0006

Dewey, J. (1909). *Moral principals in education.* Boston, MA: Houghtin Mifflin Company. Retrieved from http://www.gutenberg.org/files/25172/25172-h/25172-h.htm

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* New York, NY: Psychology Press.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Journal of Psychology*, *217*, 108–124. doi: 10.1027/0044-3409.217.3.108

Eysenck, H. J. (1939). The validity of judgments as a function of the number of judges. *Journal of Experimental Psychology*, *25*, 650–654. doi: 10.1037/h0058754

Fan, H., Jackson, T., Yang, X., Tang, W., & Zhang, J. (2010). The factor structure of the Mayer-Salovey-Caruso Emotional Intelligence Test V 2.0 (MSCEIT): A meta-analytic structural equation modeling approach. *Personality and Individual Differences*, *48*, 781–785. doi: 10.1016/j.paid.2010.02.004

Fan, X. (2012). Designing simulation studies. In H. Cooper (Ed.), *APA handbook of research methods in psychology: Vol. 2. research designs* (pp. 427–444). Washington, DC: American Psychological Association.

Farrelly, D., & Austin, E. J. (2007). Ability EI as an intelligence? Associations of the MSCEIT with performance on emotion processing and social tasks and with cognitive ability. *Cognition and Emotion*, *21*, 1043–1063. doi: 10.1080/02699930601069404

Ferrara, A. (1987). A critique of Habermas' consensus theory of truth. *Philosophy & Social Criticism*, *13*, 39–67. doi: 10.1177/019145378701300104

Field, H. (1972). Tarski's theory of truth. *The Journal of Philosophy*, *69*, 347–375. doi: 10.2307/2024879

Ford, M. E., & Tisak, M. S. (1983). A further search for social intelligence. *Journal of Educational Psychology*, *75*, 196–206. doi: 10.1037/0022-0663.75.2.196

Foy, P., Arora, A., & Stanco, G. M. (Eds.). (2013). *TIMSS 2011 user guide for the* *international database.* TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Foy, P., Brossman, B., & Galia, J. (2012). Scaling the TIMSS and PIRLS 2011 Achievement data. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in timss and pirls 2011.* Chestnut Hill: MA: TIMSS & PIRLS International Study Center, Boston College.

Frederikson, N., Carlson, S., & Ward, W. C. (1984). The place of social intelligence in a taxonomy of cognitive abilities. *Intelligence*, *8*, 315–337. doi: 10.1016/0160-2896(84)90015-1

Gaissmaier, W., & Marewski, J. N. (2011). Forecasting elections with mere recognition from small, lousy samples: A comparison of collective recognition, wisdom of crowds, and representative polls. *Judgment and Decision Making*, *6*(1), 73–88.

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* doi: 10.1002/pam.4050030422

Gifi, A. (1989). *Algorithm description for ANACOR, HOMALS, PRINCALS, and OVERALS* (Tech. Rep. No. 89-01). University of Leiden, Netherlands. Retrieved from `http://www.datatheory.nl/pdfs/89/89_01.pdf`

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester, England: John Wiley & Sons Ltd.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgements. *Psychological Bulletin*, *121*, 149–167. doi: 10.1037/0033-2909.121.1.149

Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ.* New York, NY: Bantam Books.

Gough, H. G. (1965). A validational study of the Chapin Social Insight Test. *Psychological Reports*, *17*, 355–368. doi: 10.2466/pr0.1965.17.2.355

Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin*, *53*, 267–293. doi: 10.1037/h0040755

Guilford, J. P. (1967). *The nature of human intelligence.* New York, NY: McGraw-Hill.

Guilford, J. P. (1981). Higher-order structure-of-intellect abilities. *Multivariate Behavioral Research*, *16*, 411–435. doi: 10.1207/s15327906mbr1604_1

Guo, H., Zu, J., Kyllonen, P., & Schmitt, N. (2016). Evaluation of different scoring rules for a noncognitive test development. *ETS Research Report Series*, *RR-16-03*. doi: 10.1002/ets2.12089

Guttman, L., & Levy, S. (1991). Two structural laws for intelligence tests. *Intelligence*, *15*, 79–103. doi: 10.1016/0160-2896(91)90023-7

Habermas, J. (1984). Wahrheitstheorien (1972) [Truth theories (1972)]. In *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns [Pre-studies and supplements to the theory of communicative action]* (pp. 127–183). Frankfurt am Main: Suhrkamp.

Hallquist, M., & Wiley, J. (2013). *MplusAutomation: Automating Mplus model estimation and interpretation (R package version 0.6-1-9) [Computer software].* Retrieved from `http://CRAN.R-project.org/package=MplusAutomation`

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101–125. doi: 10.1177/014662169602000201

Healy, P. (1987). Is Habermas's consensus theory a theory of truth? *Irish Philosophical Journal*, *4*, 145–152. doi: 10.5840/irishphil198741/27

Hellwig, S. (2016). *Die Erfassung von Emotional Understanding mit dem Empathic Agent Paradigma [Assessing emotional understanding with the Empathic Agent Paradigm]* (Doctoral dissertation, University of Wuppertal, Germany). Retrieved from `http://elpub.bib.uni-wuppertal.de/`

Herzog, S. M., & Hertwig, R. (2011). The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition. *Judgment and Decision Making*, *6*(1), 58–72.

Hesse, M. (1978). Habermas' consensus theory of truth. *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *2*, 373–396. doi: 10.1086/psaprocbienmeetp.1978.2.192479

Hill, S., & Ready-Campbell, N. (2011). Expert stock picker: The wisdom of (experts in) crowds. *International Journal of Electronic Commerce*, *15*, 73–102. doi: 10.2753/JEC1086-4415150304

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Convential criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. doi: 10.1080/10705519909540118

Hueffer, K., Fonseca, M. A., Leiserowitz, A., & Taylor, K. M. (2013). The wisdom of crowds: Predicting a weather and climate-related event. *Judgement and Decision Making*, *8*(2), 91–105.

Hunt, T. (1928). The measurement of social intelligence. *Journal of Applied Psychology*, *12*, 317–334. doi: 10.1037/h0075832

Intelligence, & its measurement: A Symposium. (1921). *Journal of Educational Psychology*, *12*, 124–127. doi: 10.1037/h0076078

Jäger, A. O. (1967). *Dimensionen der Intelligenz [Dimensions of intelligence]*. Göttingen, Germany: Hogrefe.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger Publishers.

Johnson-Laird, P. N. (1994a). Mental models and probabilistic thinking. *Cognition*, *50*, 189–209. doi: 10.1016/0010-0277(94)90028-0

Johnson-Laird, P. N. (1994b). A model theory of induction. *International Studies of the Philosophy of Science*, *8*, 5–29. doi: 10.1080/02698599408573474

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in Cognitive Sciences*, *5*, 434–442. doi: 10.1016/s1364-6613(00)01751-4

Kalender, I. (2012). catcher: A software program to detect answer copying in multiple-choice tests based on nominal response model. *Applied Psychological Measurement*, *36*, 625–626. doi: 10.1177/0146621612456452

Karabatsos, G., & Batchelder, W. H. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika*, *68*, 373–389. doi: 10.1007/BF02294733

Kaufman, A. S., & Kaufman, J. C. (2001). Emotional intelligence as an aspect of general intelligence: What would David Wechsler say? *Emotion*, *1*, 258–264. doi: 10.1037/1528-3542.1.3.258

Kaufmann, F. (1940). Truth and logic. *Philosophy and Phenomenological Research*, *1*, 59–69. doi: 10.2307/2103196

Keating, D. P. (1978). A search for social intelligence. *Journal of Educational Psychology*, *70*, 218–223. doi: 10.1037/0022-0663.70.2.218

Keele, S. M., & Bell, R. C. (2009). Consensus scoring, correct responses and reliability of the MSCEIT V2. *Personality and Individual Differences*, *47*, 740–747. doi: 10.1016/j.paid.2009.06.013

Kirkham, R. L. (1999). *Theories of truth.* Cambridge, MA: The MIT Press.

Krause, S., James, R., Faria, J. J., Ruxton, G. D., & Krause, J. (2011). Swarm intelligence in humans: diversity can trump ability. *Animal Behaviour*, *81*, 941–948. doi: 10.1016/j.anbehav.2010.12.018

Landy, F. J. (2006). The long, frustrating, and fruitless search for social intelligence: A cautionary tale. In K. R. Murphy (Ed.), *A critique of emotional intelligence: What are the problems and how can they be fixed?* (pp. 81–123). doi: 10.4324/9781315820927

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers of social psychology: Social psychology and decision making.* Philadelphia, PA: Psychology Press.

Lee, M. D., Zhang, S., & Shi, J. (2011). The wisdom of the crowd playing the price is right. *Memory & Cognition*, *39*, 914–923. doi: 10.3758/s13421-010-0059-7

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a likert-based testing procedure. *Intelligence*, *21*, 247–266. doi: 10.1016/0160-2896(95)90016-0

Legree, P. J., Martin, D. E., & Psotka, J. (2000). Measuring cognitive aptitude using unobtrusive knowledge tests: A new survey technology. *Intelligence*, *28*, 291–308. doi: 10.1016/s0160-2896(99)00039-2

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: An international handbook* (pp. 155–179). Cambridge, MA: Hogrefe & Huber Publishers.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse [Test construction and analysis]* (6th ed.). Weinheim, Germany: Beltz.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2006). *Intelligenz-Struktur-Test 2000 R* (2nd ed.). Göttingen, Germany: Hogrefe.

Lima Passos, V., Berger, M. P. F., & Tan, F. E. (2007). Test design optimization in CAT early stage with the nominal response model. *Applied Psychological Measurement*, *31*, 213–232. doi: 10.1177/0146621606291571

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 9020–9025. doi: 10.1073/pnas.1008636108

MacCann, C. (2006). *New approaches to measuring emotional intelligence: Exploring methodological issues with two new assessment tools* (Doctoral dissertation, University of Sydney, Australia). Retrieved from `https://ses.library.usyd.edu.au/handle/2123/934`

MacCann, C., Joseph, D. L., Newman, D. A., & Roberts, R. D. (2014). Emotional intelligence is a second-stratum factor of intelligence: Evidence from hierarchical and bifactor models. *Emotion*, *14*. doi: 10.1037/a0034755

MacCann, C., Matthews, G., Zeidner, M., & Roberts, R. D. (2003). Psychological assessment of emotional intelligence: A review of self-report and performance-based testing. *The International Journal of Organizational Analysis*, *11*, 247–274. doi: 10.1108/eb028975

MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, *8*, 540–551. doi: 10.1037/a0012746

MacCann, C., Roberts, R. D., Matthews, G., & Zeidner, M. (2004). Consensus scoring and empirical option weighting of performance-based emotional intelligence (EI) tests. *Personality and Individual Differences*, *36*, 645–662. doi: 10.1016/S0191-8869(03)00123-5

MacCann, C., Schulze, R., Matthews, G., Zeidner, M., & Roberts, R. (2008). Emotional intelligence as pop science, misled science, and sound science: A review and critical synthesis of perspectives from the field of psychology. In N. C. Karafyllis & G. Ulshöfer (Eds.), *Sexualized brains: Scientific modeling of emotional intelligence from a cultural perspective* (pp. 131–148). Cambridge, MA: The MIT Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*, 276–299. doi: 10.1037/a0036677

Marlowe, H. A. (1986). Social intelligence: Evidence for multidimensionality and construct independence. *Journal of Educational Psychology*, *78*, 52–58. doi: 10.1037/0022-0663.78.1.52

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS) - Special issue on uniform random number generation*, *8*, 3-30. doi: 10.1145/272991.272995

Matthews, G., Roberts, R. D., & Zeidner, M. (2004). Seven myths about emotional intelligence. *Psychological Inquiry*, *15*, 179–196. doi: 10.1207/s15327965pli1503_01

Matthews, G., Zeidner, M., & Roberts, R. D. (2005). Emotional intelligence: An elusive ability? In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 79–99). doi: 10.4135/9781452233529

Matthews, G., Zeidner, M., & Roberts, R. D. (2012). Emotional intelligence: A promise unfulfilled? *Japanese Psychological Research*, *54*, 105–127. doi: 10.1111/j.1468-5884.2011.00502.x

Maul, A. (2012a). Examining the structure of emotional intelligence at the item level: New perspectives, new conclusions. *Cognition and Emotion*, *26*, 503–520. doi: 10.1080/02699931.2011.588690

Maul, A. (2012b). The validity of the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) as a measure of emotional intelligence. *Emotion Review*, *4*, 394–402. doi: 10.1177/1754073912445811

Maul, A., Wilson, M., & Irribarra, D. T. (2013). On the conceptual foundations of psychological measurement. *Journal of Physics: Conference Series*, *459*, 1–6. doi: 10.1088/1742-6596/459/1/012008

Mayer, J. D., Caruso, D. R., & Salovey, P. (2000). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, *27*, 267–298. doi: 10.1016/s0160-2896(99)00016-1

Mayer, J. D., Caruso, D. R., & Salovey, P. (2016). The ability model of emotional intelligence: Principles and updates. *Emotion Review*, *8*. doi: 10.1177/1754073916639667

Mayer, J. D., DiPaolo, M., & Salovey, P. (1990). Perceiving affective content in ambiguous visual stimuli: A component of emotional intelligence. *Journal of Personality Assessment*, *54*, 772–781. doi: 10.1080/00223891.1990.9674037

Mayer, J. D., & Geher, G. (1996). Emotional intelligence and the identification of emotion. *Intelligence*, *22*, 89–113. doi: 10.1016/S0160-2896(96)90011-2

Mayer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology*, *59*, 507–536. doi: 10.1146/annurev.psych.59.103006.093646

Mayer, J. D., & Salovey, P. (1993). The intelligence of emotional intelligence. *Intelligence*, *17*, 433–442. doi: 10.1016/0160-2896(93)90010-3

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. J. Sluyter (Eds.), *Emotional development and emotional intelligence: Educational implications* (pp. 3–31). New York, NY: Harper Collins.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2002). *Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT): User's manual.* Toronto, Canada: Multi-Health Systems.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2012). The validity of the MSCEIT: Additional analyses and evidence. *Emotion Review*, *4*, 403–408. doi: 10.1177/1754073912445815

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion*, *1*, 232–242. doi: 10.1037/1528-3542.1.3.232

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion*, *3*(1), 97–105.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, *9*, 103–113. doi: 10.1111/1468-2389.00167

McLeod, A. (2011). *Kendall: Kendall rank correlation and mann-kendall trend test (R* *package version 2.2) [Computer software]*. Retrieved from `http://CRAN.R-project.org/package=Kendall`

Merkle, E. C., & Steyvers, M. (2011). A psychological model for aggregating judgements of magnitude. In J. S. et al. (Ed.), *Social computing, behavioral modeling, and prediction* (pp. 236–343). doi: 10.1007/978-3-642-19656-0

Michailidis, G., & de Leeuw, J. (1998). The gifi system of descriptive multivariate analysis. *Statistical Science*, *13*, 307–336. doi: 10.1214/ss/1028905828

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *RR-03-16*. doi: 10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (Tech. Rep. No. 9). Principled Assessment Designs for Inquiry. Retrieved from `http://padi.sri.com/downloads/TR9_ECD.pdf`

Mittring, G., & Rost, D. H. (2008). Die verflixten Distraktoren: Über den Nutzen einer theoretischen Distraktorenanalyse bei Matrizentests (für besser Begabte und Hochbegabte) [The accursed distractors: About the benefit of a theoretical distractor analysis for matrix tests (for gifted and highly gifted)]. *Diagnostica*, *54*, 193–201. doi: 10.1026/0012-1924.54.4.193

Mohoric, T., Taksic, V., & Duran, M. (2010). In search of "the correct answer" in an ability-based emotional intelligence (EI) test. *Studia Psychologica*, *52*(3), 219–228.

Moss, F. A., Hunt, T., Omwake, K. T., & Woodward, L. G. (1955). *George Washington University Series Social Intelligence Test* (2nd ed.). Center for Psychological Service, Washington, DC.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2005). A theoretical basis for situational judgment tests. In J. A. Weekley & R. E. Ployhardt (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 57–81). doi: 10.4324/9780203774878

Mullis, I. V. S., Drucker, K. T., Preuschoff, C., Arora, A., & Stanco, G. M. (2012). Assessment framework and instrument development. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011.* Chestnut Hill: MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics.* Chestnut Hill: MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment frameworks.* Chestnut Hill: MA: TIMSS & PIRLS International Study Center, Boston College.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. doi: 10.1177/014662169201600206

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Nevo, B. (1993). In search of a correctness typology for intelligence. *New Ideas in Psychology*, *11*, 391–397. doi: 10.1016/0732-118x(93)90009-3

Nofer, M., & Hinz, O. (2014). Are crowds on the internet wiser than experts? The case of a stock prediction community. *Jorunal of Business Economics*, *84*, 303–338. doi: 10.1007/s11573-014-0720-x

Oravecz, Z., Anders, R., & Batchelder, W. H. (2015). Hierarchical bayesian modeling for test theory without an answer key. *Psychometrika*, *80*, 341–364. doi: 10.1007/s11336-013-9379-4

Oravecz, Z., Vandekerckhove, J., & Batchelder, W. H. (2014). Bayesian cultural consensus theory. *Field Methods*, *26*, 207–222. doi: 10.1177/1525822x13520280

Orchard, B., MacCann, C., Schulze, R., Matthews, G., Zeidner, M., & Roberts, R. D. (2009). New directions and alternative approaches to the measurement of emotional intelligence. In C. Stough, D. H. Saklofske, & J. D. A. Parker (Eds.), *Assessing* *emotional intelligence: theory, research, and application* (pp. 321–344). New York, NY: Springer.

O'Sullivan, M., & Guilford, J. P. (1975). Six factors of behavioral cognition: Understanding other people. *Journal of Edducational Measurement*, *12*, 255–271. doi: 10.1111/j.1745-3984.1975.tb01027.x

O'Sullivan, M., Guilford, J. P., & deMille, R. (1965). *Measurement of social intelligence* (Tech. Rep. No. 1976). The University of Southern California, Los Angeles, CA. Retrieved from `http://files.eric.ed.gov/fulltext/ED010278.pdf`

Palmer, B. R., Gignac, G., Manocha, R., & Stough, C. (2005). A psychometric evaluation of the Mayer-Salovey-Caruso Emotional Intelligence Test Version 2.0. *Intelligence*, *33*, 285–305. doi: 10.1016/j.intell.2004.11.003

Petrides, K. V., & Furnham, A. (2000). On the dimensional structure of emotional intelligence. *Personality and Individual Differences*, *29*, 313–320. doi: 10.1016/s0191-8869(99)00195-6

Petrides, K. V., & Furnham, A. (2001). Trait emotional intelligence: Psychometric investigation with reference to established trait taxonomies. *European Journal of Personality*, *15*, 425–448. doi: 10.1002/per.416

Preckel, F. (2003). *Diagnostik intellektueller Hochbegabung: Testentwicklung zur Erfassung der fluiden Intelligenz [Assessment of intellectual high giftedness: Construction of a test of fluid intelligence].* Göttingen, Germany: Hogrefe.

Preston, K., Reise, S., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement*, *71*, 523–550. doi: 10.1177/0013164410382250

Psychometric Society. (1979). Publication policy regarding monte carlo studies. *Psychometrika*, *44*(2), 133–134.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rescher, N. (1993). *Pluralism: Against the demand for consensus.* Oxford, England: Clarendon Press.

Rescher, N. (2012). Die Kriterien der Wahrheit (1973) [Criteria of truth (1973))]. In G. Skirbekk (Ed.), *Wahrheitstheorien: Eine Auswahl aus den Diskussionen über Wahrheit im 20. Jahrhundert [Truth theories: A selection of discussions about truth in the 20th century]* (Vol. 11, pp. 337–390). Frankfurt am Main, Germany: Suhrkamp.

Revelle, W. (2013). *psych: Procedures for psychological, psychometric, and personality research (R package version 1.3.2) [Computer software].* Retrieved from `http://CRAN.R-project.org/package=psych`

Riggio, R. E. (1986). Assessment of basic social skills. *Journal of Personality and Social Psychology*, *51*, 649–660. doi: 10.1037/0022-3514.51.3.649

Rijmen, F., Tuerlinckx, F., DeBoeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205. doi: 10.1037/1082-989x.8.2.185

Roberts, R. D., Markham, P. M., Matthews, G., & Zeidner, M. (2005). Assessing intelligence: Past, present, and future. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 333–360). doi: 10.4135/9781452233529

Roberts, R. D., Schulze, R., & MacCann, C. (2008). The measurement of emotional intelligence: A decade of progress? In G. Boyle, G. Matthews, & D. Saklofske (Eds.), *The Sage handbook of personality theory and assessmen. Vol. 2: Personality measurement and testing* (pp. 461–482). doi: 10.4135/9781849200479

Roberts, R. D., Schulze, R., Reid, J., O'Brien, K., MacCann, C., & Maul, A. (2006). Exploring the validity of the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) with established emotions measures. *Emotion*, *6*, 663–669. doi:

10.1037/1528-3542.6.4.663

Roberts, R. D., Zeidner, M., & Matthews, G. (2001). Does emotional intelligence meet traditional standards for an intelligence? Some new data and conclusions. *Emotion*, *1*, 196–231. doi: 10.1037/1528-3542.1.3.196

Romney, A. K. (1999). Culture consensus as a statistical model. *Current Anthropology*, *40*, 103–115. doi: 10.1086/200062

Romney, A. K., Batchelder, W. H., & Weller, S. C. (1987). Recent applications of cultural consensus theory. *American Behavioral Scientist*, *31*, 163–177. doi: 10.1177/000276487031002003

Romney, A. K., Boyd, J. P., Moore, C. C., Batchelder, W. H., & Brazill, T. J. (1996). Culture as shared cognitive representations. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 4699–4705. doi: 10.1073/pnas.93.10.4699

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*, 313–338. doi: 10.1525/aa.1986.88.2.02a00020

Rossen, E., Kranzler, J. H., & Algina, J. (2008). Confirmatory factor analysis of the Mayer-Salovey-Caruso Emotional Intelligence Test V 2.0 (MSCEIT). *Personality and Individual Differences*, *44*, 1258–1269. doi: 10.1016/j.paid.2007.11.020

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, Cognition, and Personality*, *9*, 185–211. doi: 10.2190/dugg-p24e-52wk-6cdg

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications to developmental research* (pp. 399–419). Thousand Oaks, CA: SAGE Publications.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley & R. E. Ployhardt (Eds.), *Situational judgment tests: Theory,* *measurement, and application* (pp. 135–155). doi: 10.4324/9780203774878

Schulze, R., & Jobmann, A.-L. (2016). *Der Zusammenhang zwischen Emotional Understanding und Arbeitsgedächtniskapazität [The relation between emotional understanding and working-memory-capacity].* Paper presented at the 50th congress of the Deutsche Gesellschaft für Psychologie, Leipzig, Germany.

Schulze, R., & Roberts, R. (Eds.). (2005). *Emotional intelligence: An international handbook.* Cambridge, MA: Hogrefe & Huber Publishers.

Schulze, R., Wilhelm, O., & Kyllonen, P. C. (2007). Approaches to the assessment of emotional intelligence. In G. Matthews, M. Zeidner, & R. D. Roberts (Eds.), *Emotional intelligence: Knowns and unknowns* (p. 199-229). doi: 10.1093/acprof:oso/9780195181890.003.0008

Seidel, K. (2007). *Social intelligence and auditory intelligence: Useful constructs?* (Doctoral dissertation, Otto-von-Guericke-Universität Magdeburg, Germany). Retrieved from `http://edoc2.bibliothek.uni-halle.de/urn/urn:nbn:de:gbv:ma9:1-2674`

Shye, S. (1988). Inductive and deductive reasoning: A structural reanalysis of ability tests. *Journal of Applied Psychology*, *73*, 308–311. doi: 10.1037/0021-9010.73.2.308

Simmons, J. P., Nelson, L. D., Galak, J., & Frederick, S. (2010). Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research*, *38*, 1–15. doi: 10.1086/658070

Skrondal, A. (2000). Design and analysis of monte carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*, 137–167. doi: 10.1207/S15327906MBR3502_1

Solomon, M. (2006). Groupthink versus the wisdom of crowds: The social epistemology of deliberation and dissent. *The Southern Journal of Philosophy*, *44*, 28–42. doi: 10.1111/j.2041-6962.2006.tb00028.x

Spearman, C. (1904). The proof and measurement of association between two things. *The*

American Journal of Psychology, *15*. doi: 10.2307/1422689

Stephen, A., & Lee, L. (2010). Are crowds always wiser? *Advances in Consumer Research*, *37*, 94–97.

Sternberg, R. J. (1978). The nature of mental abilities. *American Psychologist*, *34*, 214–230. doi: 10.1037/0003-066x.34.3.214

Strang, R. (1930). Measures of social intelligence. *American Journal of Sociology*, *36*, 263–269. doi: 10.1086/215342

Strang, R. (1932). An analysis of errors made in a test of social intelligence. *The Journal of Educational Sociology*, *5*, 291–299. doi: 10.2307/2961660

Stricker, L. J., & Rock, D. A. (1990). Interpersonal competence, social intelligence, and general ability. *Personality and Individual Differences*, *11*, 833–839. doi: 10.1016/0191-8869(90)90193-u

Suh, Y., & Bolt, D. M. (2011). A nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement*, *48*, 188-205. doi: 10.1111/j.1745-3984.2011.00139.x

Surowiecki, J. (2004). *The wisdom of crowds.* New York, NY: Anchor Books.

Süß, H.-M., & Beauducel, A. (2005). Faceted models of intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 313–332). doi: 10.4135/9781452233529

Tarski, A. (1935). Der Wahrheitsbegriff in den formalisierten Sprachen [The truth term in formalized languages]. *Studia Philosophica*, *1*, 261–405.

Tenenhaus, M., & Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical data. *Psychometrika*, *50*, 91–119. doi: 10.1007/BF02294151

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory* *models* (pp. 43–75). doi: 10.4324/9780203861264

Thorndike, E. L. (1920). Intelligence and its use. *Harper's Magazine*, *140*, 227–235.

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence.* New York, NY: Teachers College, Columbia University. Retrieved from `https://archive.org/details/measurementofint00thoruoft`

Thorndike, R. L. (1936). Factor analysis of social and abstract intelligence. *Journal of Educational Psychology*, *27*, 231–233. doi: 10.1037/h0059840

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago, IL: The University of Chicago Press.

Tuerlinckx, F., & Wang, W.-C. (2004). Models for polytomous data. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and monlinear approach* (pp. 75–109). doi: 10.1007/978-1-4757-3990-9

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, *95*, 261–289. doi: 10.1007/s10994-013-5401-4

Van Rooy, D. I. L., Viswesvaran, C., & Pluta, P. (2005). An evaluation of construct validity: What is this thing called emotional intelligence? *Human Performance*, *18*, 445–462. doi: 10.1207/s15327043hup1804_9

Von Aster, M. G., Neubauer, A., & Horn, R. (2006). *Wechsler-Intelligenztest für Erwachsene (WIE-III). Deutschsprachige Bearbeitung und Adaptation des WAIS-III von David Wechsler [Wechsler intelligence test for adults (WIE-III). German adaption of the WAIS-III of David Wechsler]* (2nd ed.). Frankfurt am Main, Germany: Pearson Assessment.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*, 645–647. doi: 10.1111/j.1467-9280.2008.02136.x

Warwick, J., Nettelbeck, T., & Ward, L. (2010). AEIM: A new measure and method of scoring abilities-based emotional intelligence. *Personality and Individual Differences*, *48*, 66–71. doi: 10.1016/j.paid.2009.08.018

Waubert de Puiseau, B., Aßfalg, A., Erdfelder, E., & Bernstein, D. M. (2012). Extracting the truth from conflicting eyewitness reports: A formal modeling approach. *Journal of Experimental Psychology: Applied*, *18*, 390–403. doi: 10.1037/a0029801

Webster, C. M., Iannucci, A. L., & Romney, A. K. (2002). Consensus analysis for the measurement and validation of personality traits. *Field Methods*, *14*, 46–64. doi: 10.1177/1525822x0201400104

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). doi: 10.1037/11329-000

Weekley, J. A., & Ployhart, R. E. (2006). An introduction to situational judgment testing. In J. A. Weekley & R. E. Ployhardt (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 1–10). doi: 10.4324/9780203774878

Weis, S. (2008). *Theory and measurement of social intelligence as a cognitive performance construct* (Doctoral dissertation, Otto-von-Guericke-Universität Magdeburg, Germany). Retrieved from http://diglib.uni-magdeburg.de/Dissertationen/2008/susweis.pdf

Weis, S., & Süß, H.-M. (2005). Social intelligence: A review and critical discussion of measurement concepts. In R. Schulze & R. Roberts (Eds.), *Emotional intelligence: An international handbook* (pp. 203–230). Cambridge, MA: Hogrefe & Huber Publishers.

Weis, S., & Süß, H.-M. (2007). Reviving the search for social intelligence: A multitrait-multimethod study of its structure and construct validity. *Personality and Individual Differences*, *42*, 3–14. doi: 10.1016/j.paid.2006.04.027

Weller, S. C. (1987). Shared knowledge, intracultural variation, and knowledge aggregation. *American Behavioral Scientist*, *31*, 178–193. doi: 10.1177/000276487031002004

Weller, S. C. (2007). Cultural consensus theory: Applications and frequently asked questions. *Field Methods*, *19*, 339–368. doi: 10.1177/1525822X07303502

White, A. P., & Zammarelli, J. E. (1981). Convergence principle: Information in the answer sets of some multiple-choice intelligence tests. *Applied Psychological Measurement*, *5*, 21–27. doi: 10.1177/014662168100500103

Wilhelm, O. (2000). *Psychologie des schlussfolgernden Denkens: Differentialpsychologische Prüfung von Strukturüberlegungen [Psychology of reasoning: Testing structural theories].* Hamburg, Germany: Dr Kovac.

Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 373–392). doi: 10.4135/9781452233529

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science*, *36*, 452–470. doi: 10.1111/j.1551-6709.2011.01223.x

Zeidner, M., Matthews, G., & Roberts, R. D. (2001). Slow down, you move too fast: Emotional intelligence remains an "elusive" intelligence. *Emotion*, *1*, 265–275. doi: 10.1037/1528-3542.1.3.265

Zeidner, M., Roberts, R. D., & Matthews, G. (2004). The emotional intelligence bandwagon: Too fast to live, too young to die? *Psychological Inquiry*, *15*, 239–248. doi: 10.1207/s15327965pli1503_04

A Individual Analysis of Mode CBM (Graphics)

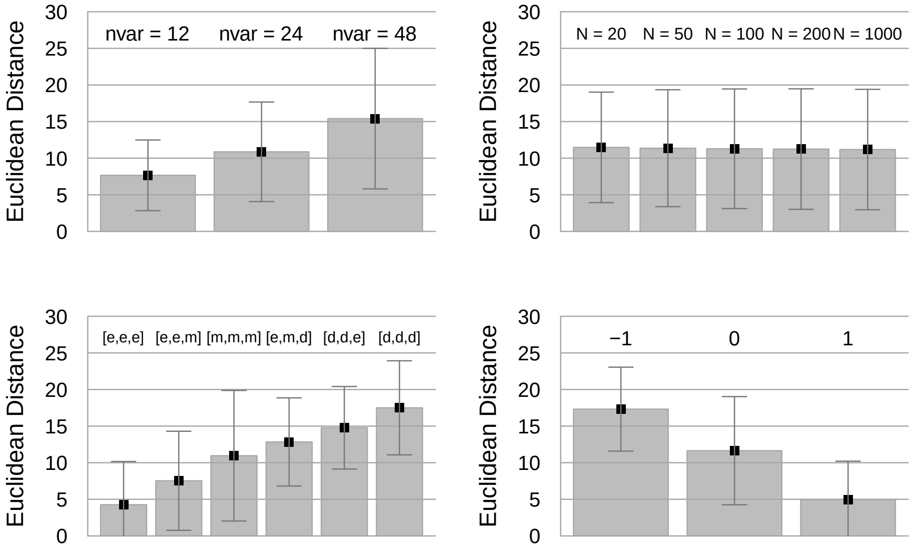


Figure 23. Five-categorical data: Main effects of independent variables on Euclidean distance (Mode CBM). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

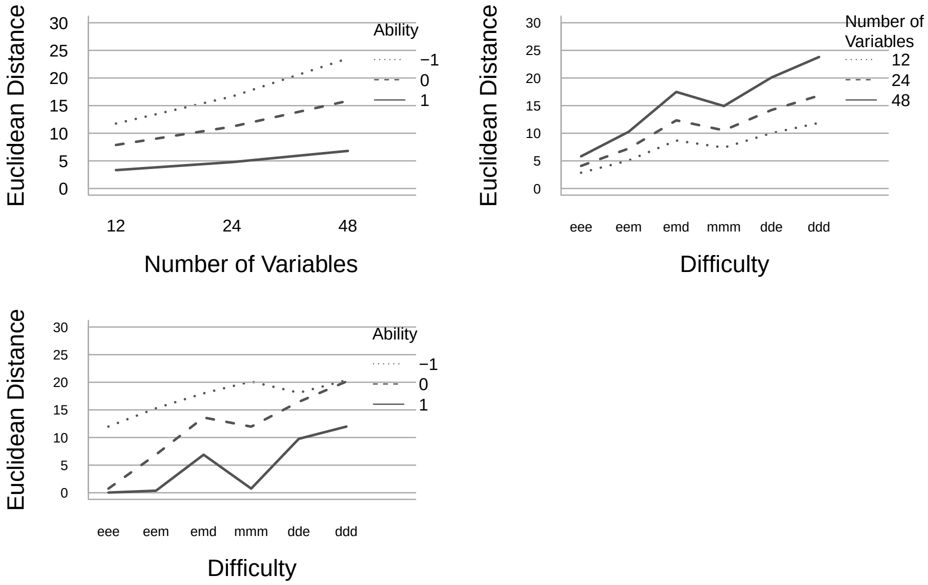


Figure 24. Five-categorical data: Interaction effect of independent variables on Euclidean distance (Mode CBM). Top left panel: Interaction of number of variables and ability; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

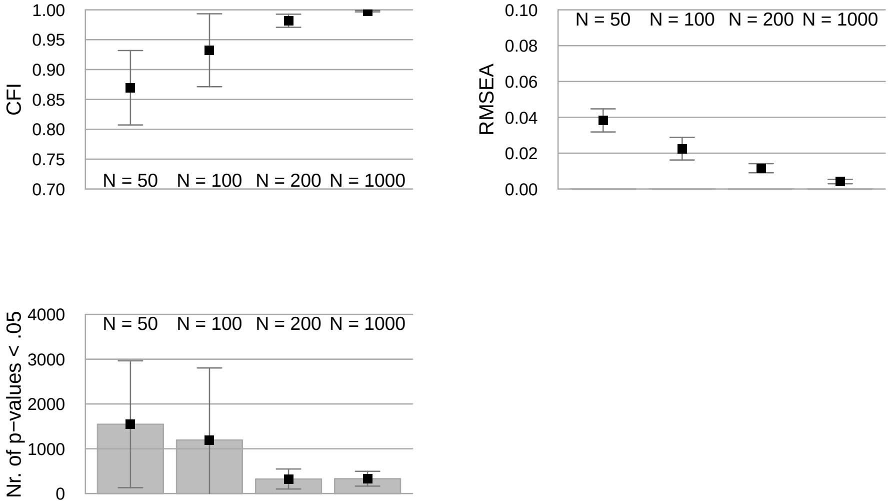


Figure 25. Two-categorical data: Main effect of number of respondents on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Mode CBM).

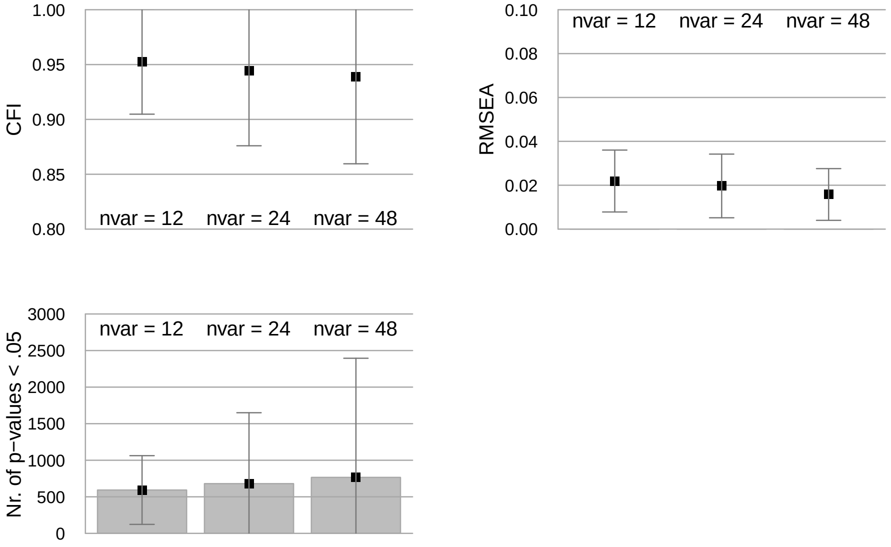


Figure 26. Two-categorical data: Main effect of number of variables on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Mode CBM).

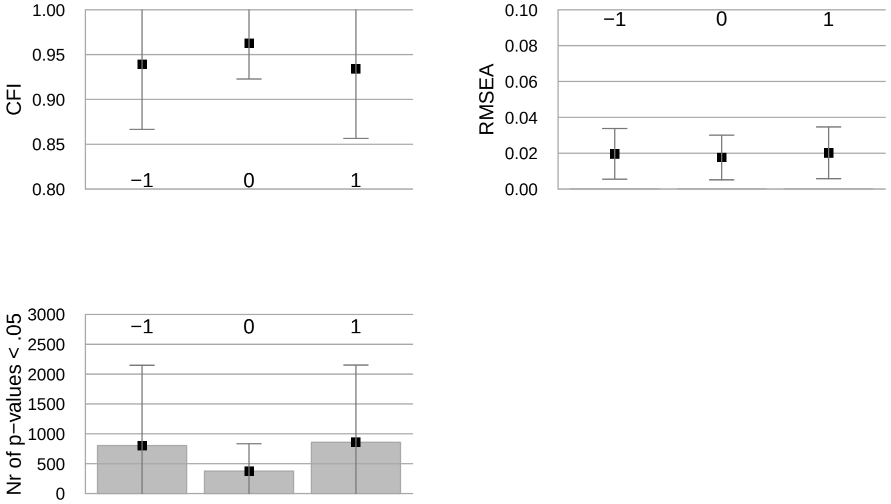


Figure 27. Two-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Mode CBM).

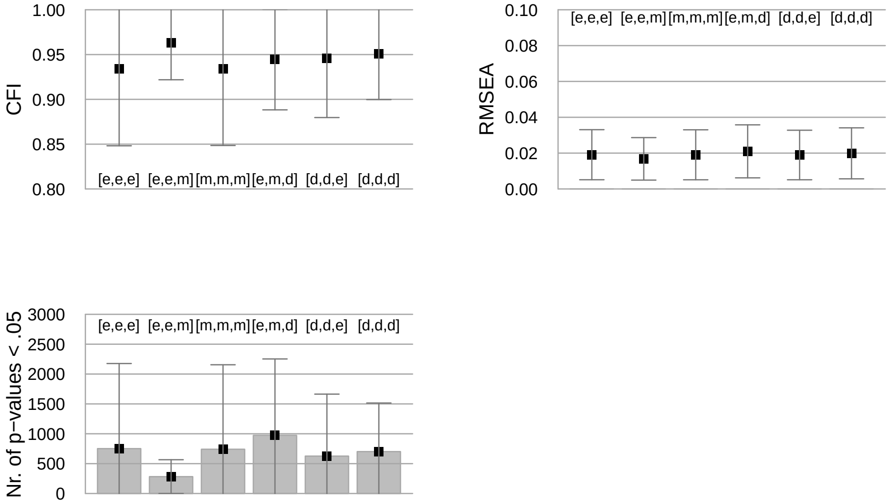


Figure 28. Two-categorical data: Main effect of difficulty on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Mode CBM).

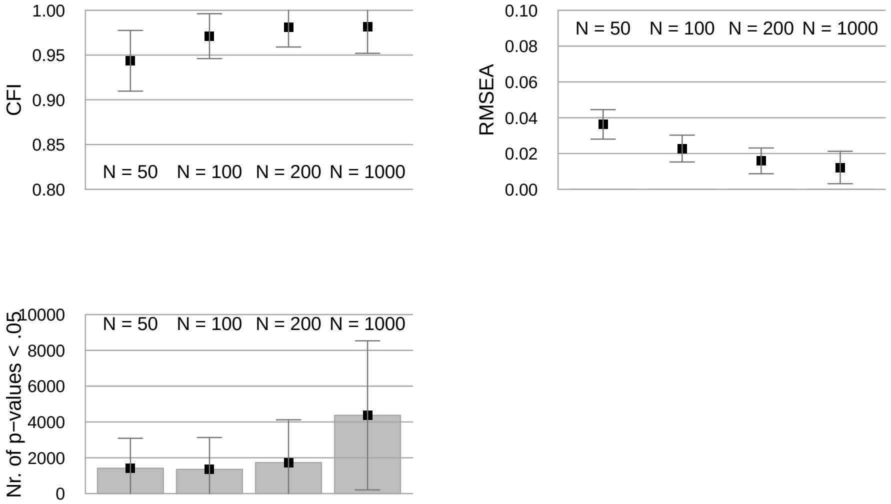


Figure 29. Five-categorical data: Main effect of number of respondents on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Mode CBM).

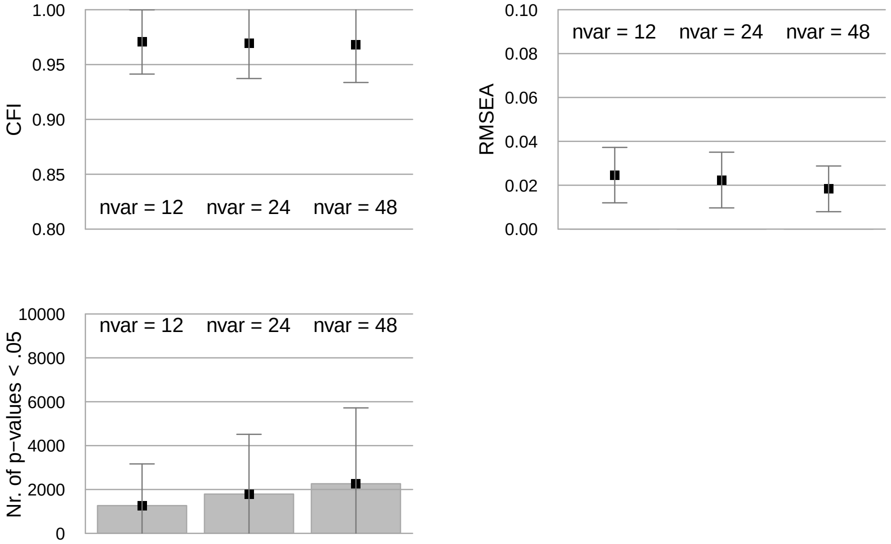


Figure 30. Five-categorical data: Main effect of number of variables on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Mode CBM).

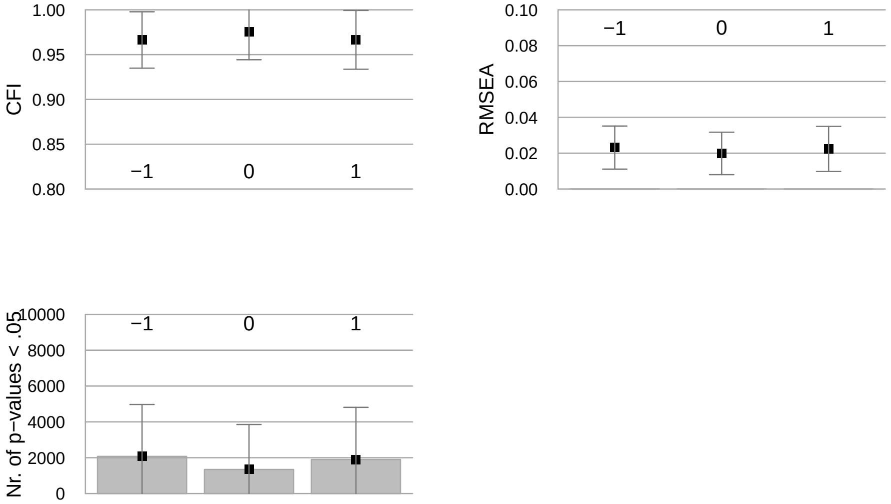


Figure 31. Five-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Mode CBM).

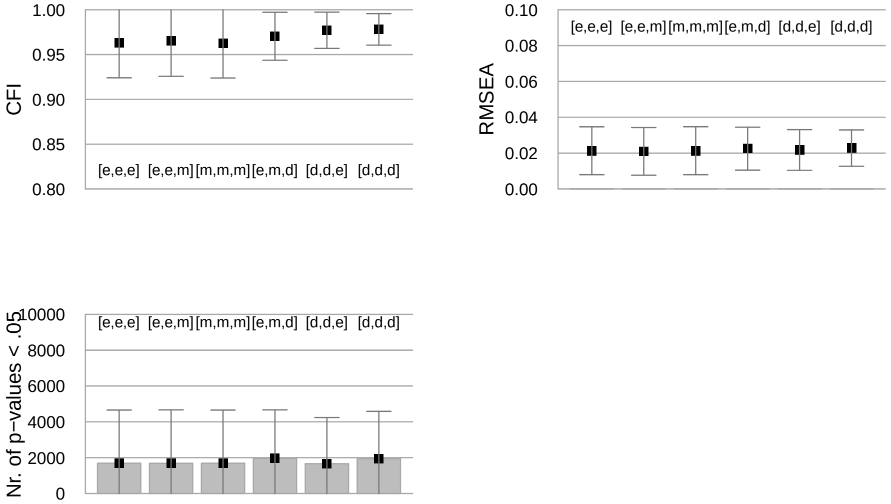


Figure 32. Five-categorical data: Main effect of difficulty on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Mode CBM).

B Individual Analysis of Proportion CBM (Graphics)

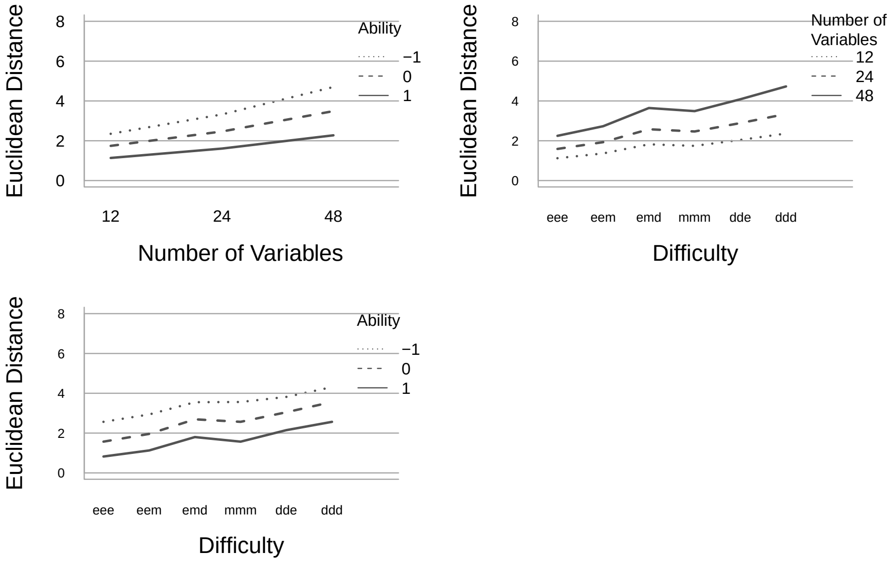


Figure 33. Two-categorical data: Interaction effect of independent variables on Euclidean distance (Proportion CBM). Top left panel: Interaction of number of variables and ability; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

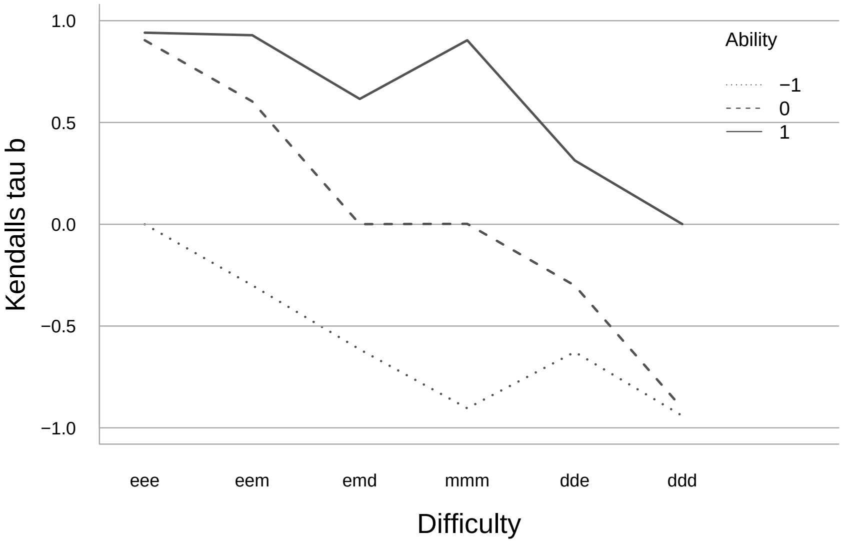


Figure 34. Five-categorical data: Interaction effect of ability and difficulty on Kendall's τb (Proportion CBM).

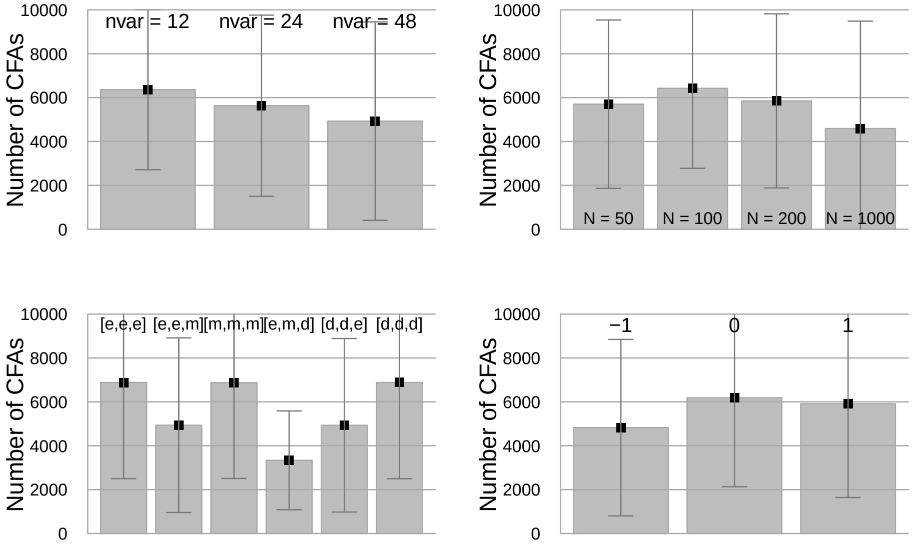


Figure 35. Two-categorical data: Main effects of independent variables on number of successful CFAs (Proportion CBM). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

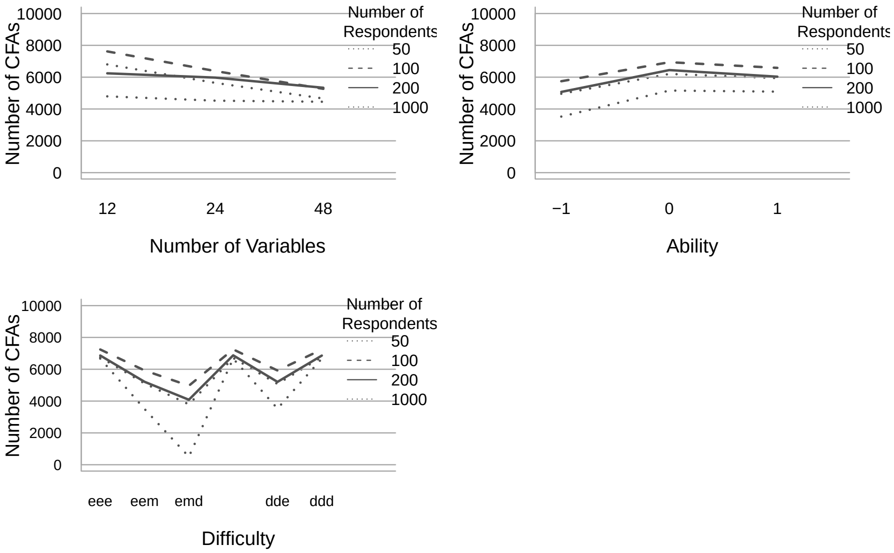


Figure 36. Two-categorical data: Interaction effect of independent variables on number of successful CFAs (Proportion CBM). Top left panel: Interaction of number of variables and number of respondents; Top right panel: Interaction of number of respondents and ability; Bottom left panel: Interaction of number of respondents and difficulty.

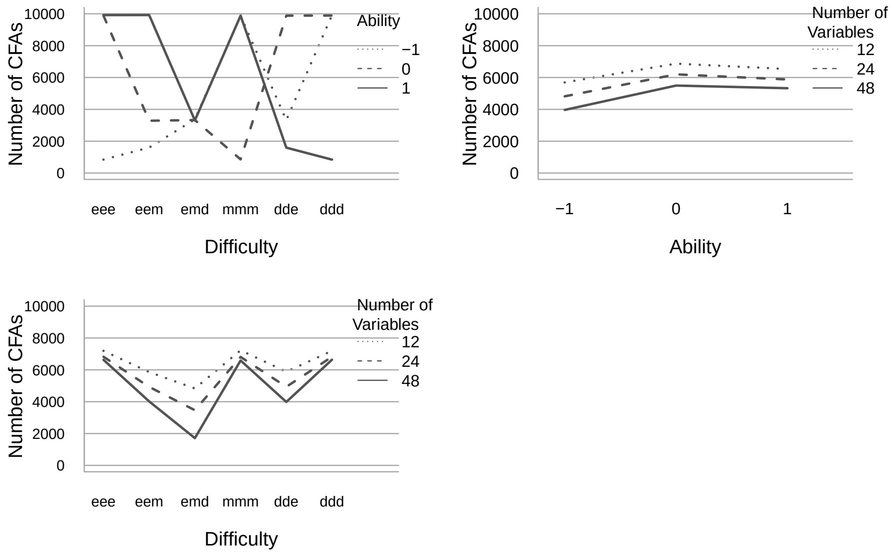


Figure 37. Two-categorical data: Interaction effect of independent variables on number of successful CFAs (Proportion CBM). Top left panel: Interaction of ability and difficulty; Top right panel: Interaction of number of variables and ability; Bottom left panel: Interaction of number of variables and difficulty.

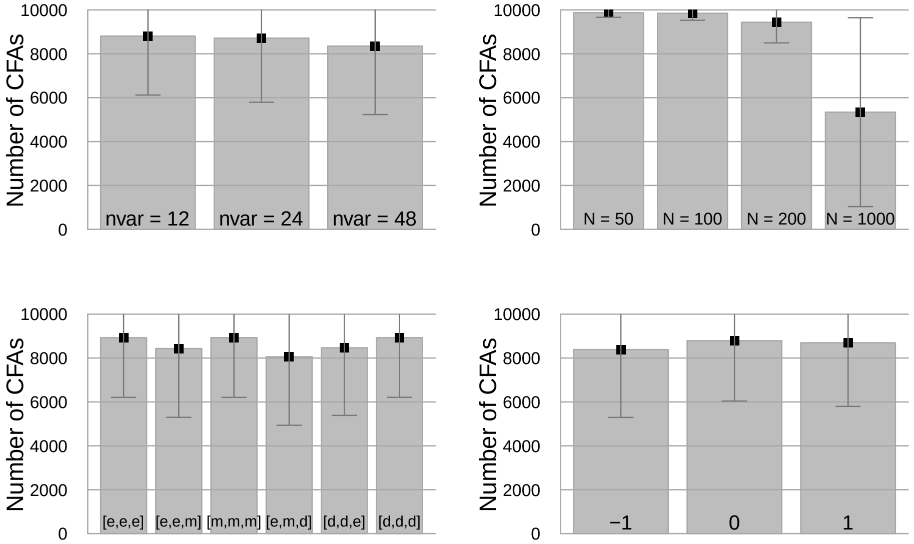


Figure 38. Five-categorical data: Main effects of independent variables on number of successful CFAs (Proportion CBM). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

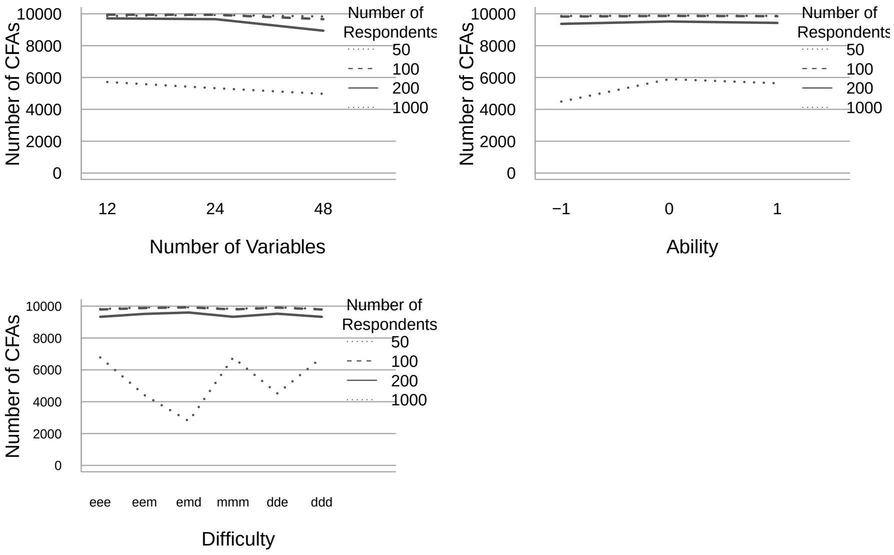


Figure 39. Five-categorical data: Interaction effect of independent variables on number of successful CFAs (Proportion CBM). Top left panel: Interaction of number of variables and number of respondents; Top right panel: Interaction of number of respondents and ability; Bottom left panel: Interaction of number of respondents and difficulty.

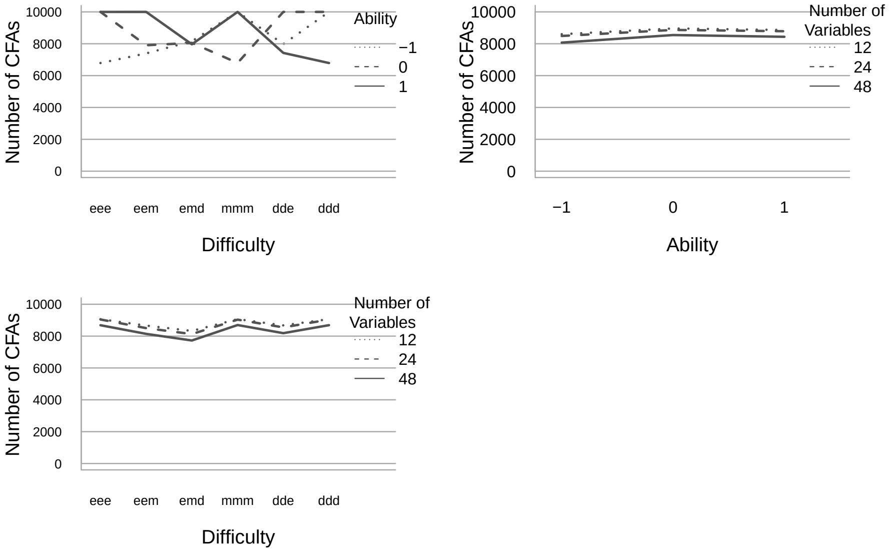


Figure 40. Five-categorical data: Interaction effect of independent variables on number of successful CFAs (Proportion CBM). Top left panel: Interaction of ability and difficulty; Top right panel: Interaction of number of variables and ability; Bottom left panel: Interaction of number of variables and difficulty.

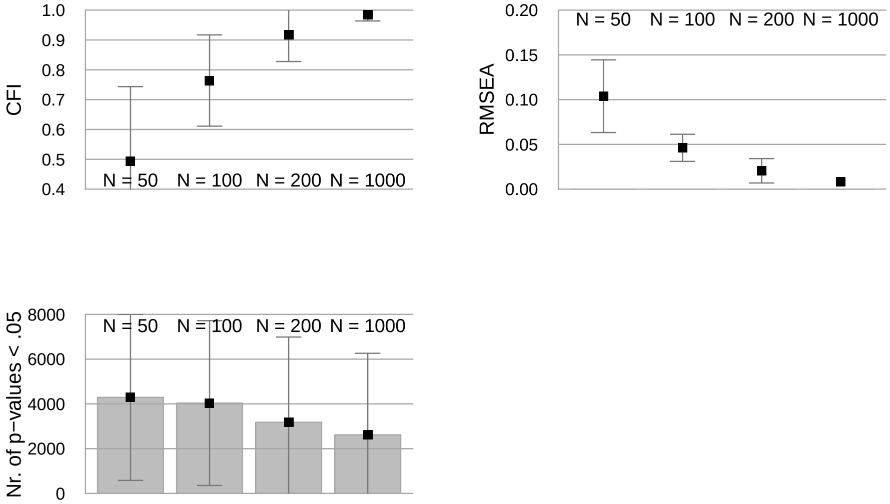


Figure 41. Two-categorical data: Main effect of number of respondents on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Proportion CBM).

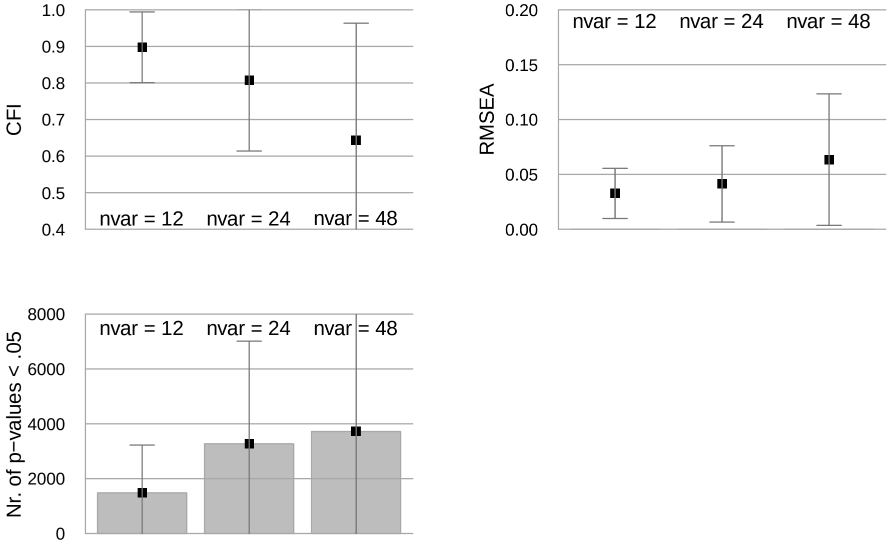


Figure 42. Two-categorical data: Main effect of number of variables on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Proportion CBM).

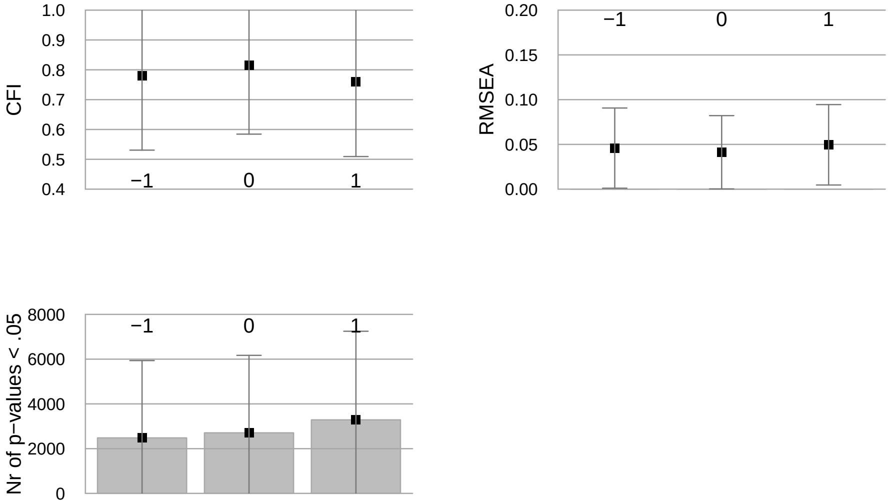


Figure 43. Two-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Proportion CBM).

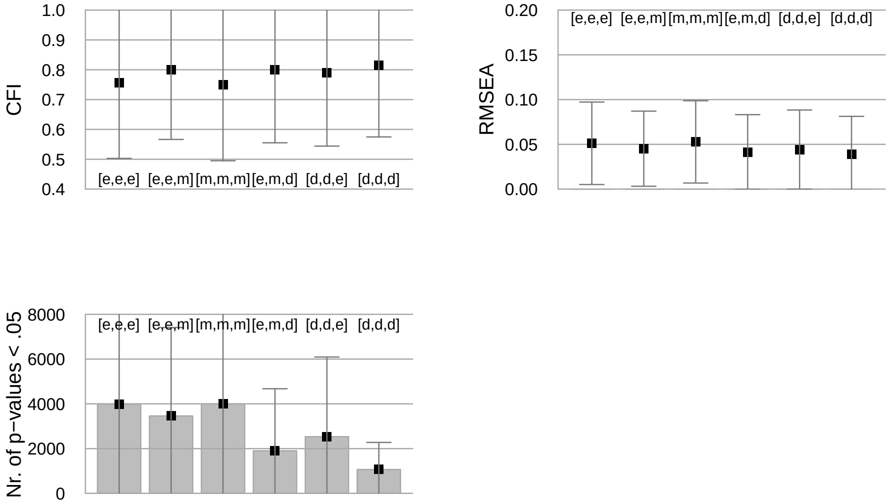


Figure 44. Two-categorical data: Main effect of difficulty on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Proportion CBM).

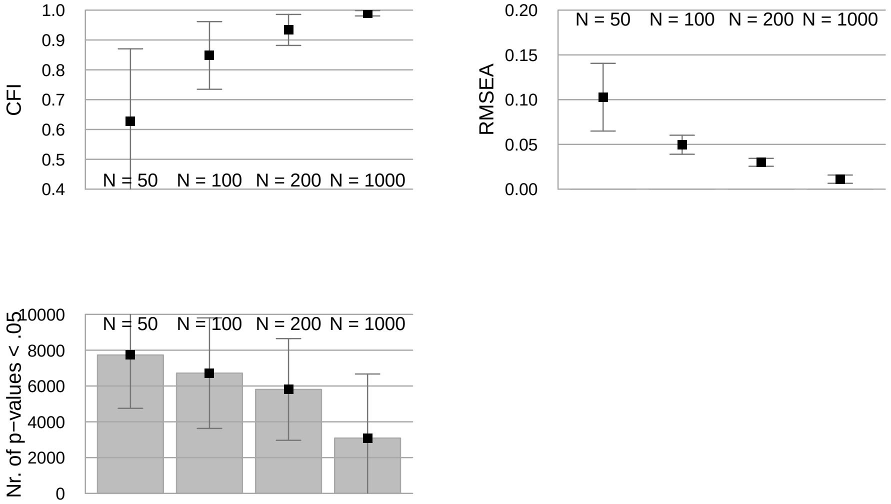


Figure 45. Five-categorical data: Main effect of number of respondents on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Proportion CBM).

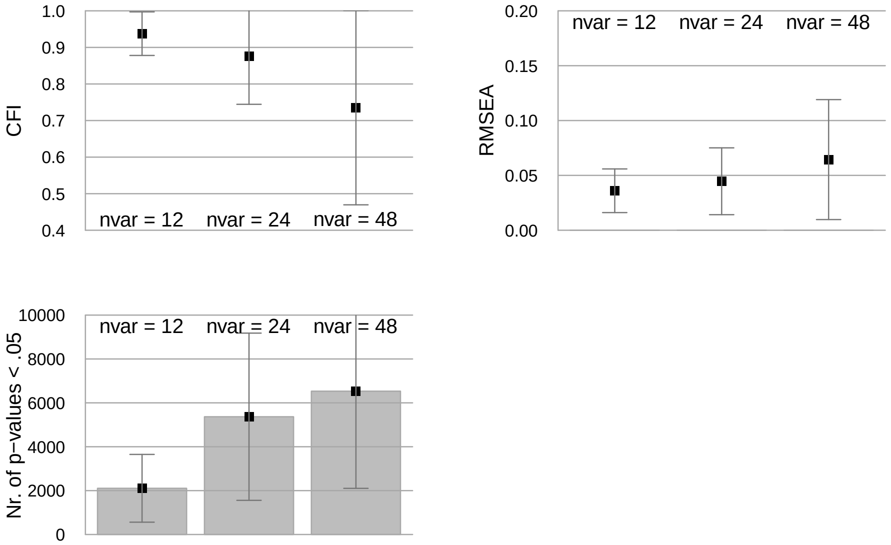


Figure 46. Five-categorical data: Main effect of number of variables on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Proportion CBM).

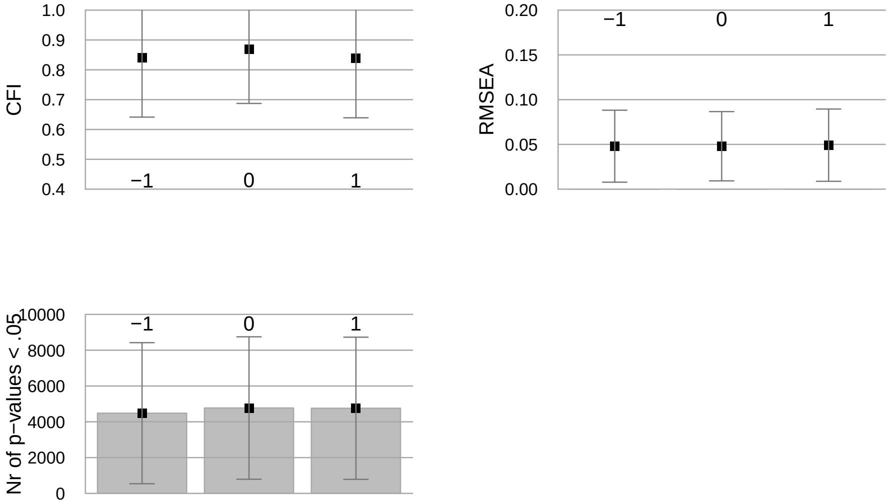


Figure 47. Five-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Proportion CBM).

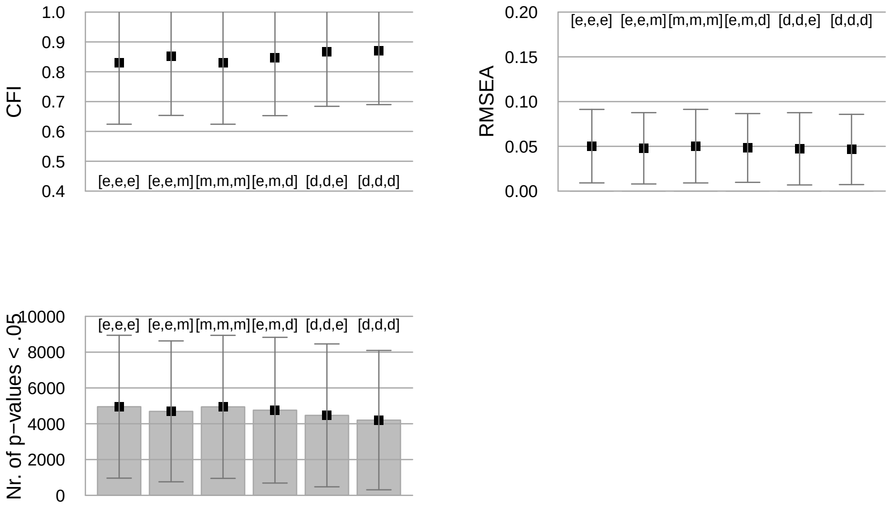


Figure 48. Five-categorical data: Main effect of difficulty on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Proportion CBM).

C Individual Analysis of Consensus Analysis (Graphics)

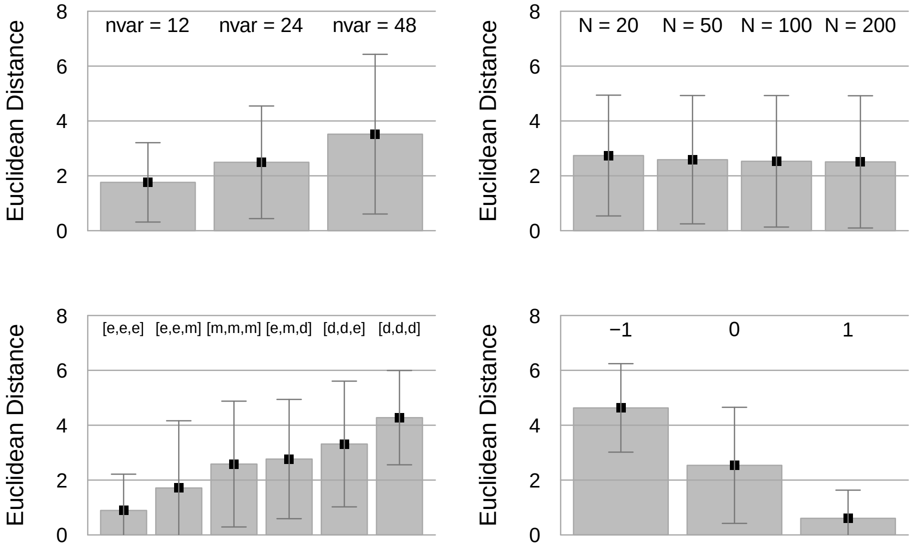


Figure 49. Two-categorical data: Main effects of independent variables on Euclidean distance (Consensus Analysis). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

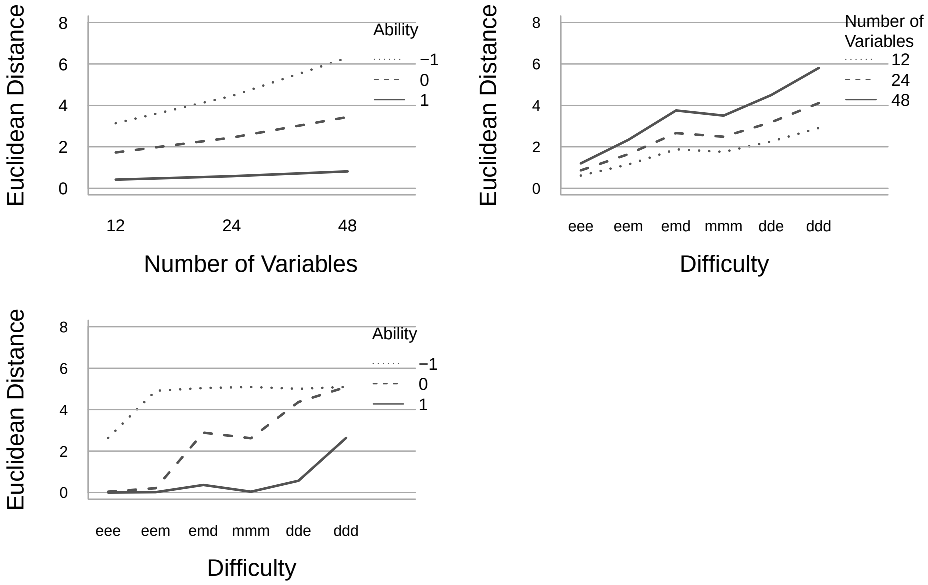


Figure 50. Two-categorical data: Interaction effects of independent variables on Euclidean distance (Consensus Analysis). Top left panel: Interaction of number of variables and ability; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

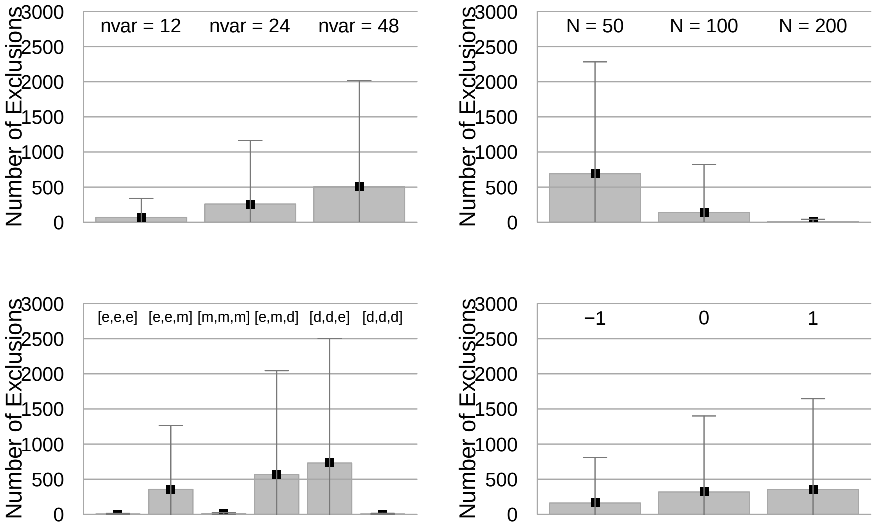


Figure 51. Five-categorical data: Main effect of independent variables on number of exclusions due to zero variances after scoring (Consensus Analysis). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

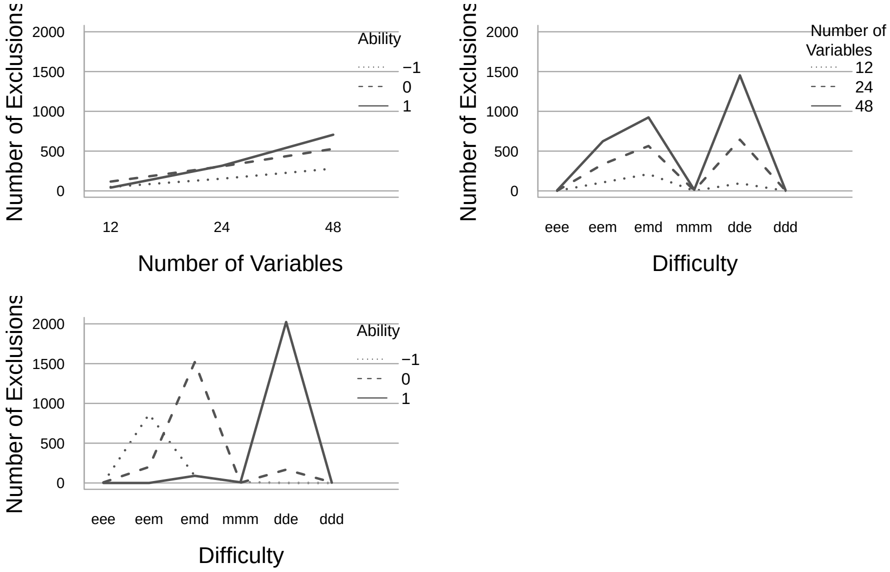


Figure 52. Five-categorical data: Interaction effect of independent variables on number of exclusions due to zero variances after scoring (Consensus Analysis). Top left panel: Interaction of number of variables and ability; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

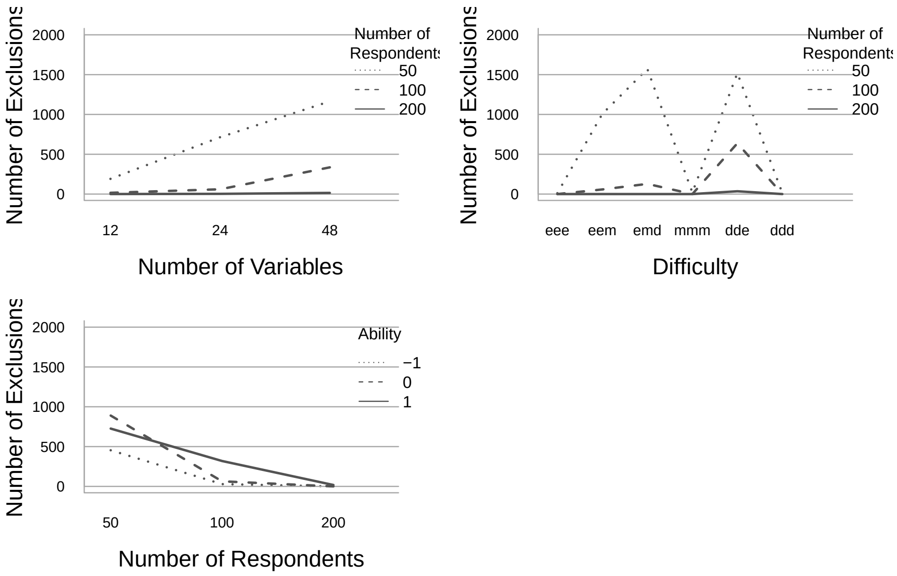


Figure 53. Five-categorical data: Interaction effect of independent variables on number of exclusions due to zero variances after scoring (Consensus Analysis). Top left panel: Interaction of number of variables and number of respondents; Top right panel: Interaction of number of respondents and difficulty; Bottom left panel: Interaction of ability and number of respondents.

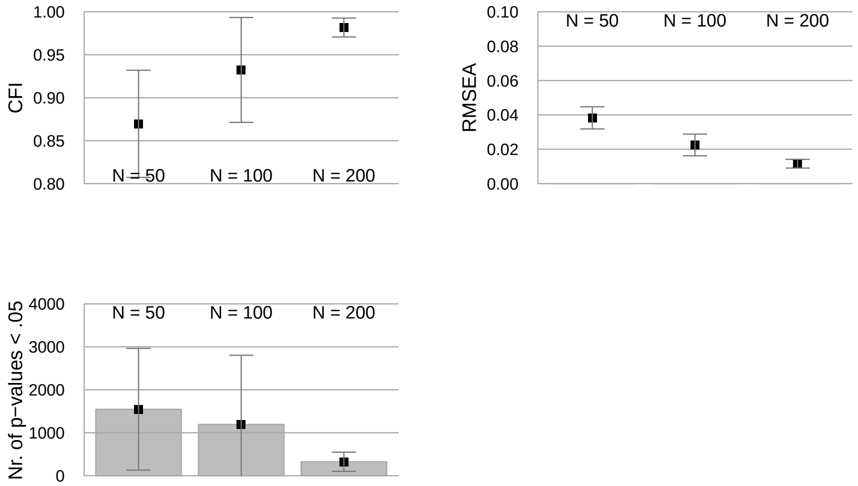


Figure 54. Two-categorical data: Main effect of number of respondents on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Consensus Analysis).

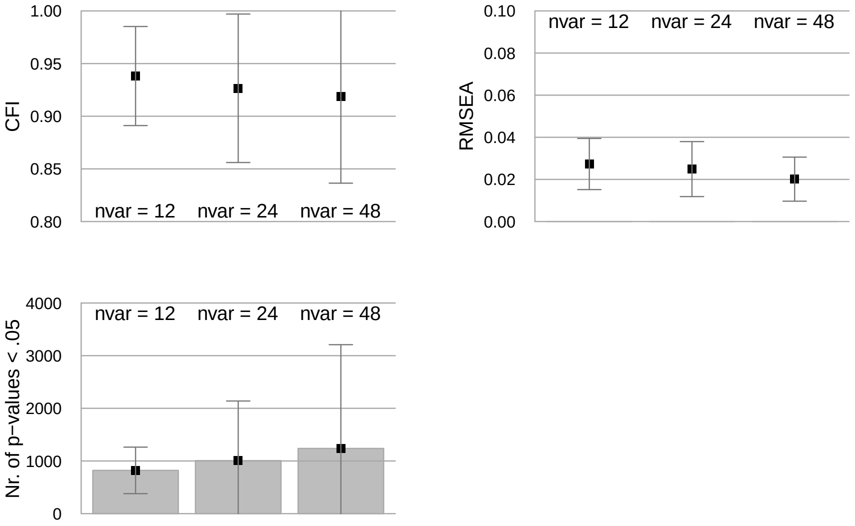


Figure 55. Two-categorical data: Main effect of number of variables on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Consensus Analysis).

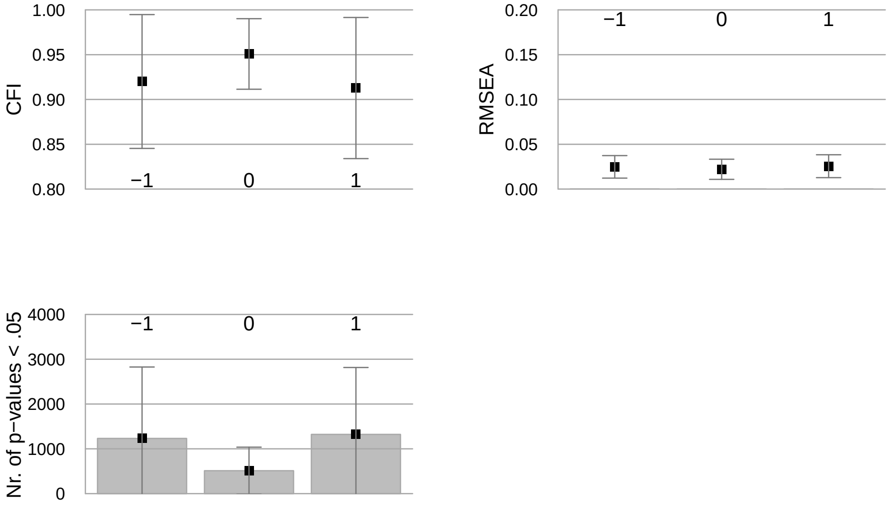


Figure 56. Two-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Consensus Analysis).

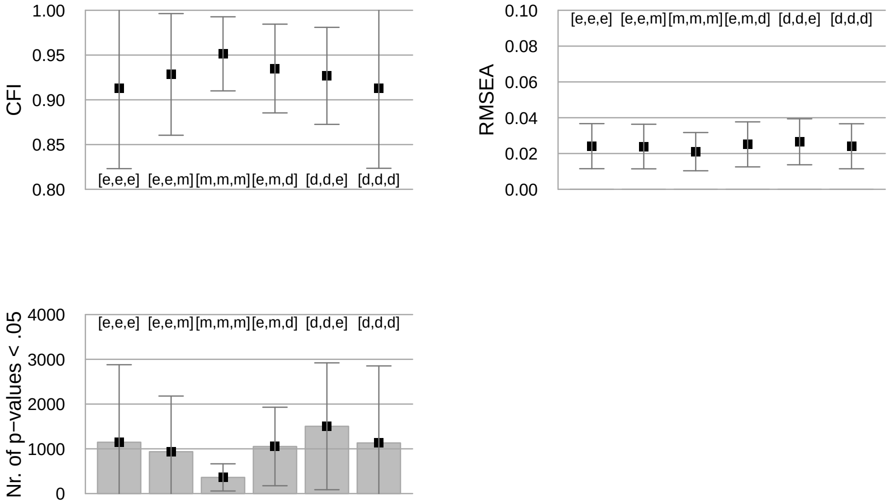


Figure 57. Two-categorical data: Main effect of difficulty on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Consensus Analysis).

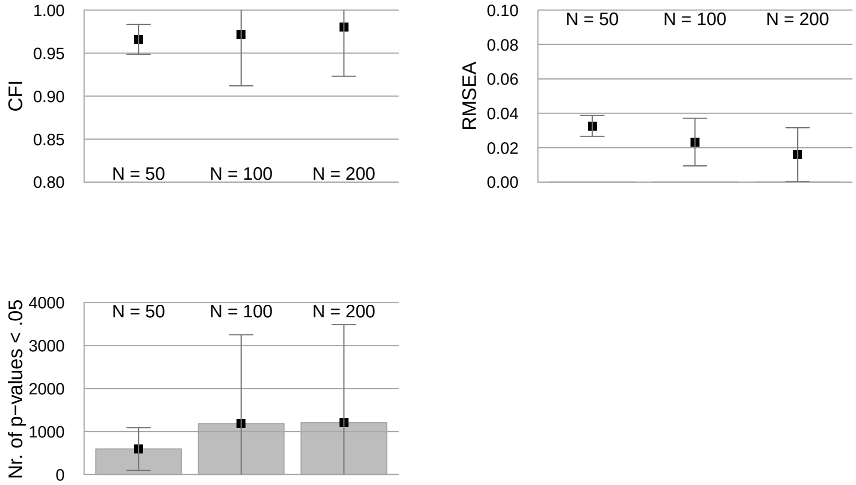


Figure 58. Five-categorical data: Main effect of number of respondents on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Consensus Analysis).

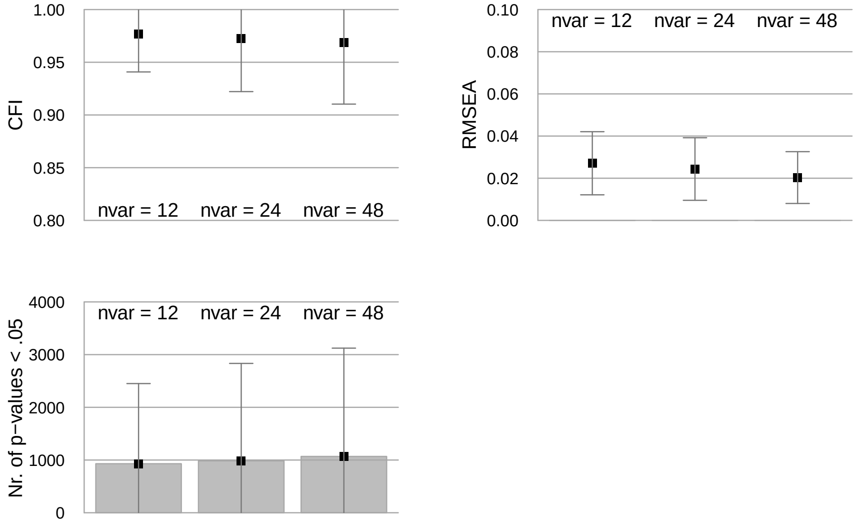


Figure 59. Five-categorical data: Main effect of number of variables on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Consensus Analysis).

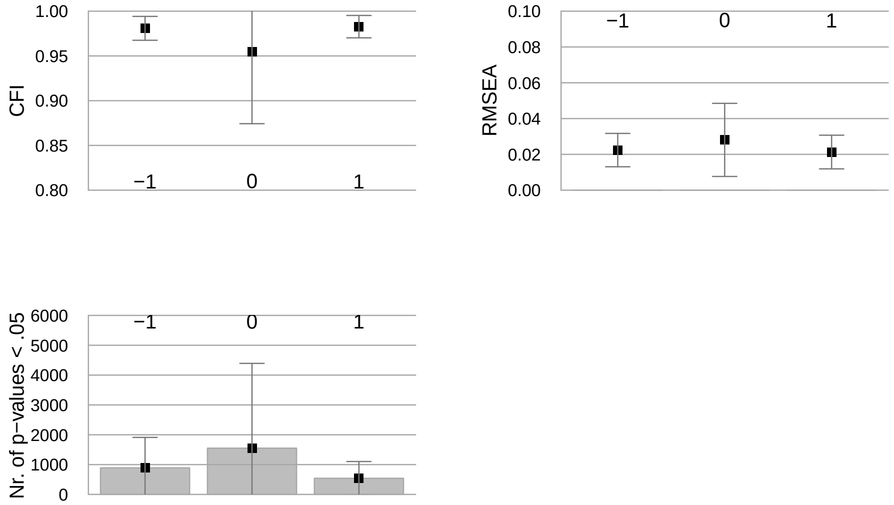


Figure 60. Five-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Consensus Analysis).

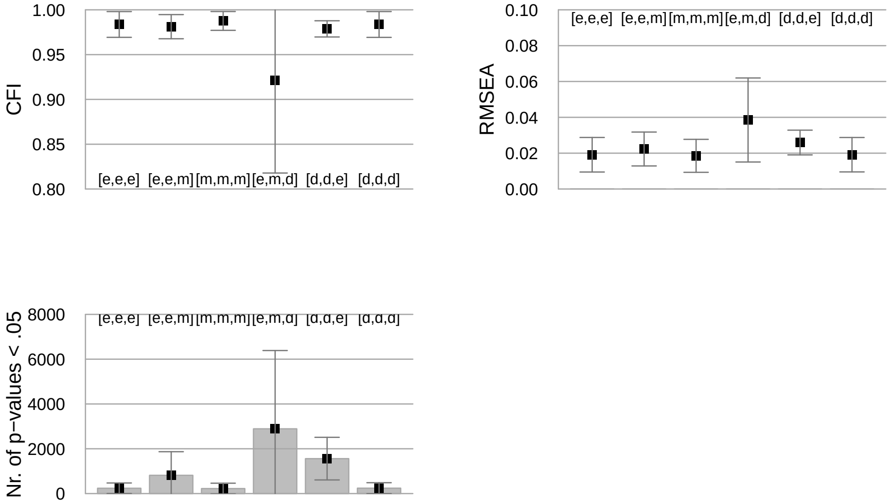


Figure 61. Five-categorical data: Main effect of difficulty on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (Consensus Analysis).

D Individual Analysis of HOMALS (Graphics)

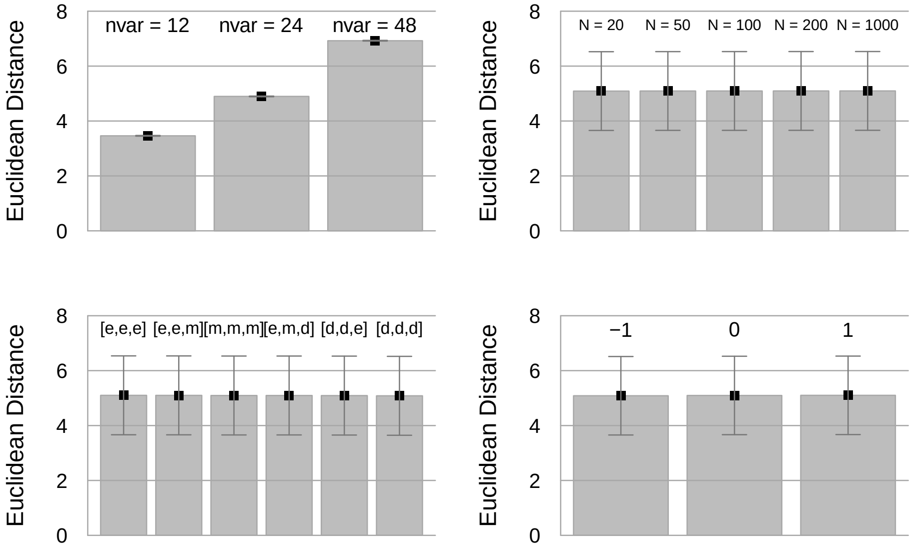


Figure 62. Two-categorical data: Main effect of independent variables on Euclidean distance (HOMALS). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

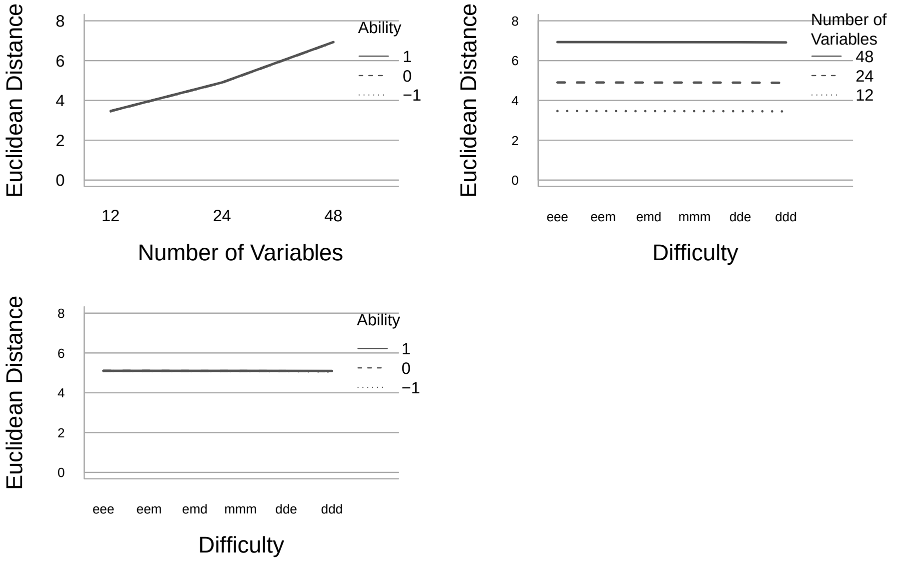


Figure 63. Two-categorical data: Interaction effects of independent variables on Euclidean distance (HOMALS). Top left panel: Interaction of number of variables and ability; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

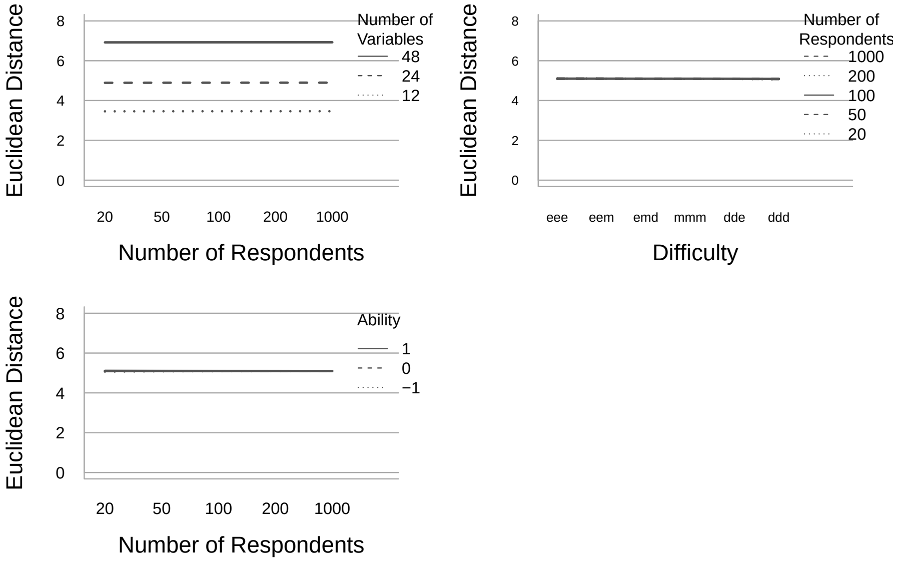


Figure 64. Two-categorical data: Interaction effects of independent variables on Euclidean distance (HOMALS). Top left panel: Interaction of number of respondents and number of variables; Top right panel: Interaction of number of respondents and difficulty; Bottom left panel: Interaction of ability and number of respondents.

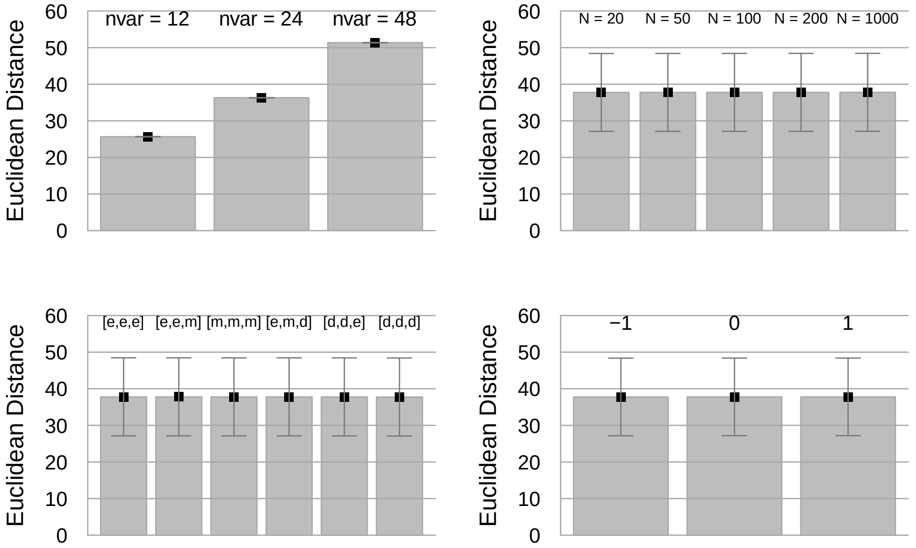


Figure 65. Five-categorical data: Main effect of independent variables on Euclidean distance (HOMALS). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

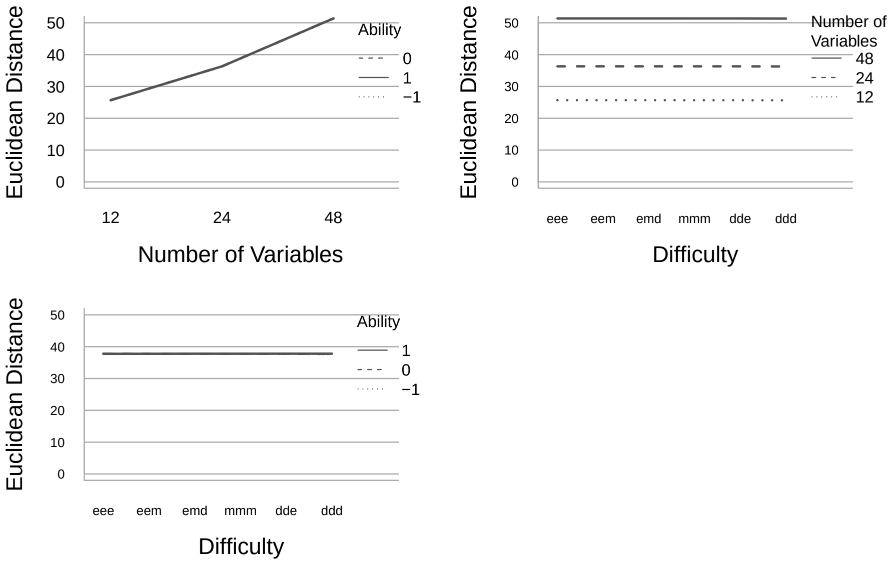


Figure 66. Five-categorical data: Interaction effects of independent variables on Euclidean distance (HOMALS). Top left panel: Interaction of number of variables and ability; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

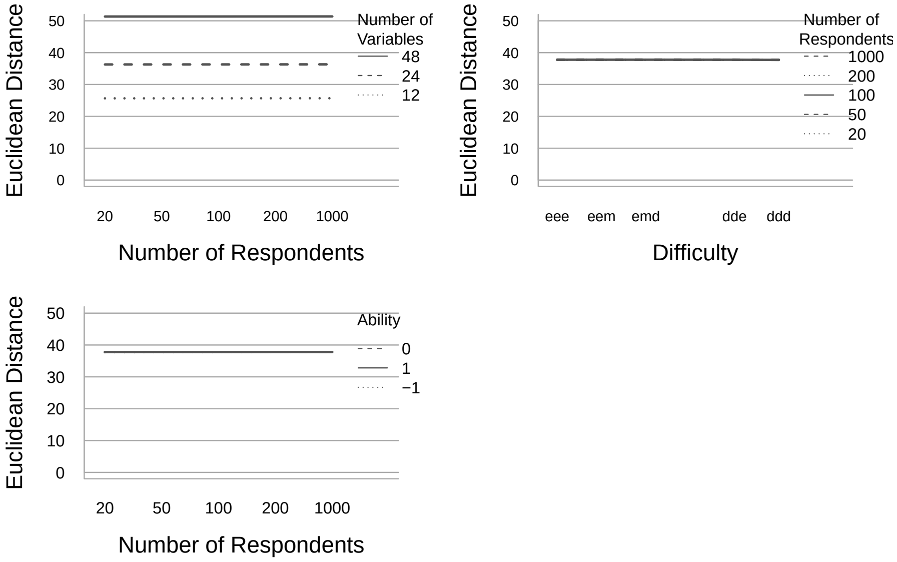


Figure 67. Five-categorical data: Interaction effects of independent variables on Euclidean distance (HOMALS). Top left panel: Interaction of number of respondents and number of variables; Top right panel: Interaction of number of respondents and difficulty; Bottom left panel: Interaction of ability and number of respondents.

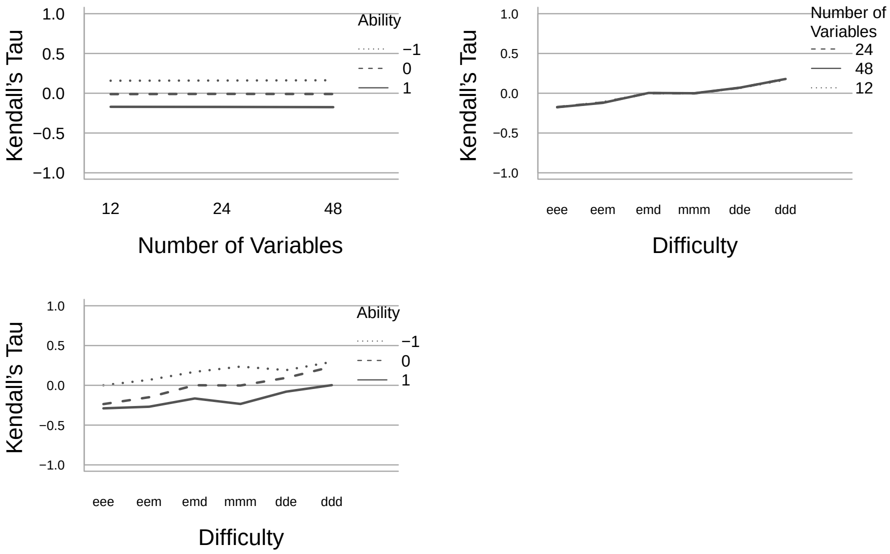


Figure 68. Five-categorical data: Interaction effects of independent variables on Kendall's τb (HOMALS). Top left panel: Interaction of number of variables and ability; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

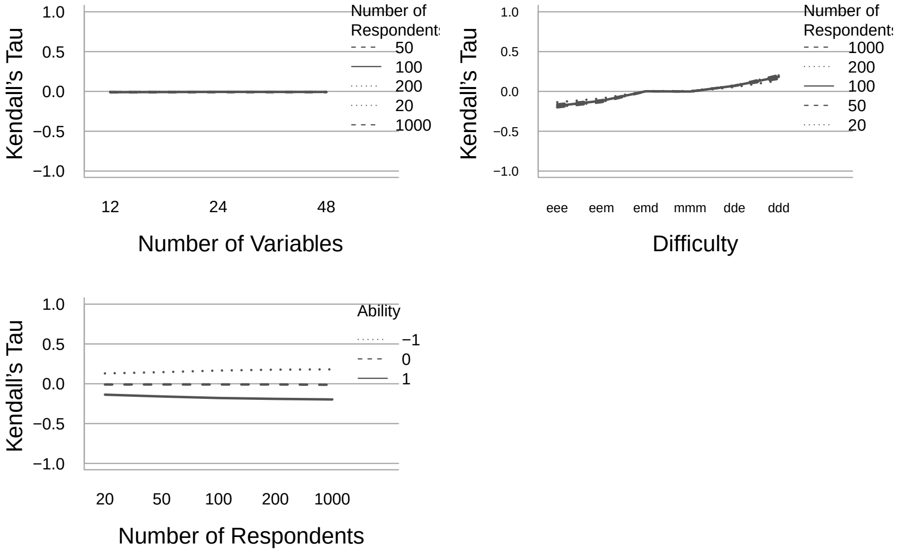


Figure 69. Five-categorical data: Interaction effects of independent variables on Kendall's τb (HOMALS). Top left panel: Interaction of number of respondents and number of variables; Top right panel: Interaction of number of respondents and difficulty; Bottom left panel: Interaction of ability and number of respondents.

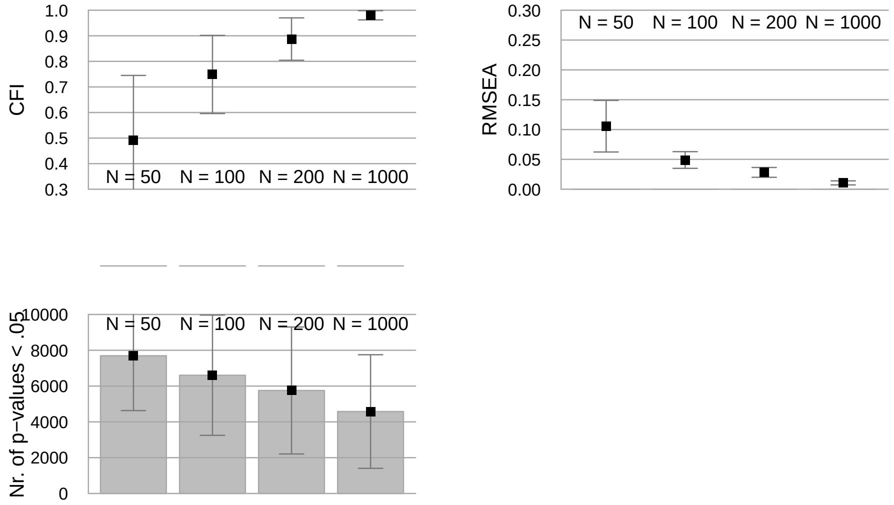


Figure 70. Two-categorical data: Main effect of number of respondents on CFI (top left panel), RMSEA (top right panel) and number of p-values $<$ .05 (bottom left panel) (HOMALS).

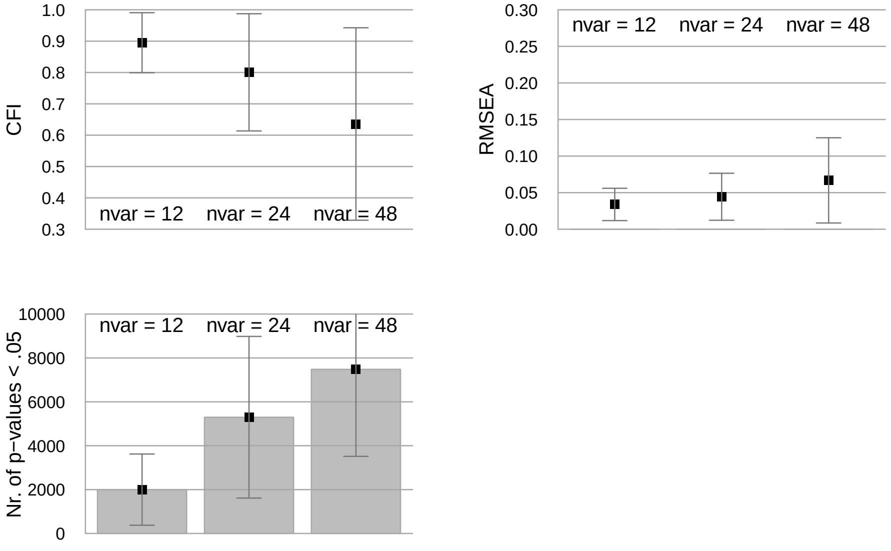


Figure 71. Two-categorical data: Main effect of number of variables on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (HOMALS).

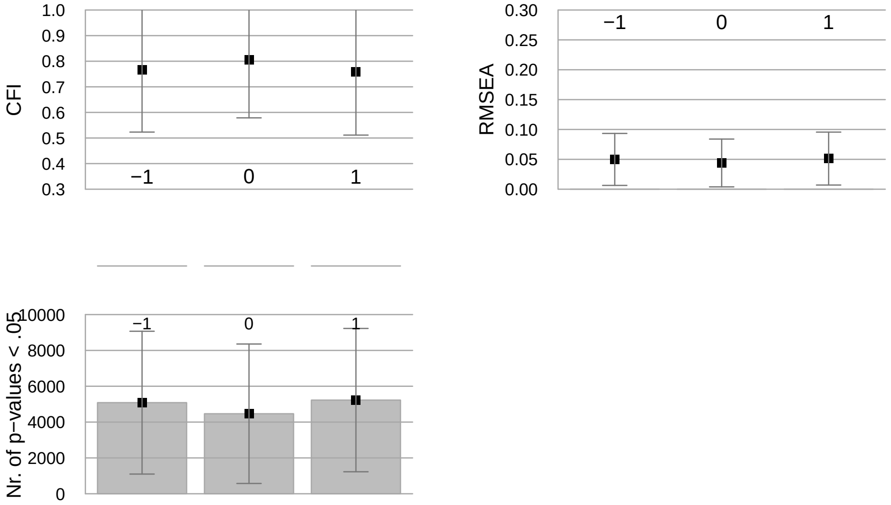


Figure 72. Two-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (HOMALS).

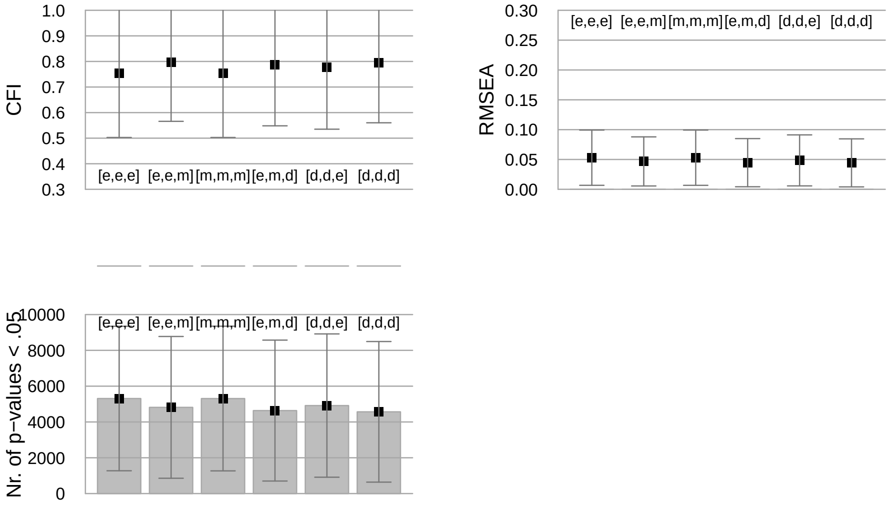


Figure 73. Two-categorical data: Main effect of difficulty on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (HOMALS).

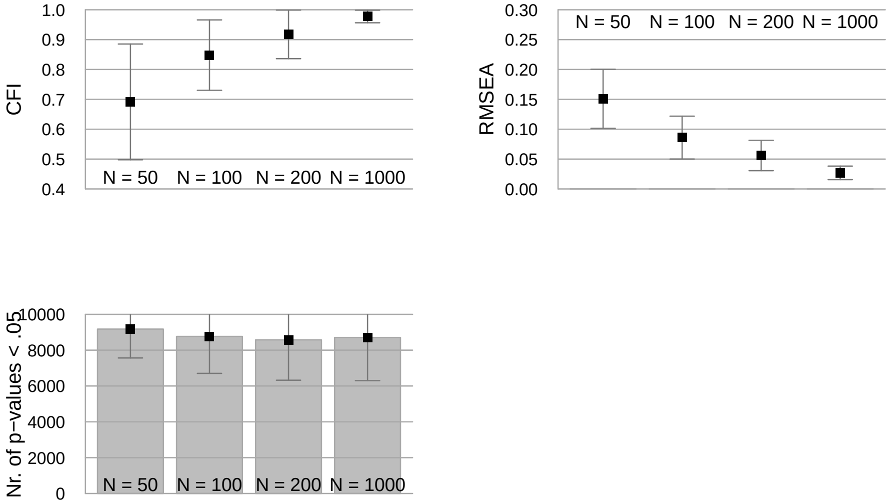


Figure 74. Five-categorical data: Main effect of number of respondents on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (HOMALS).

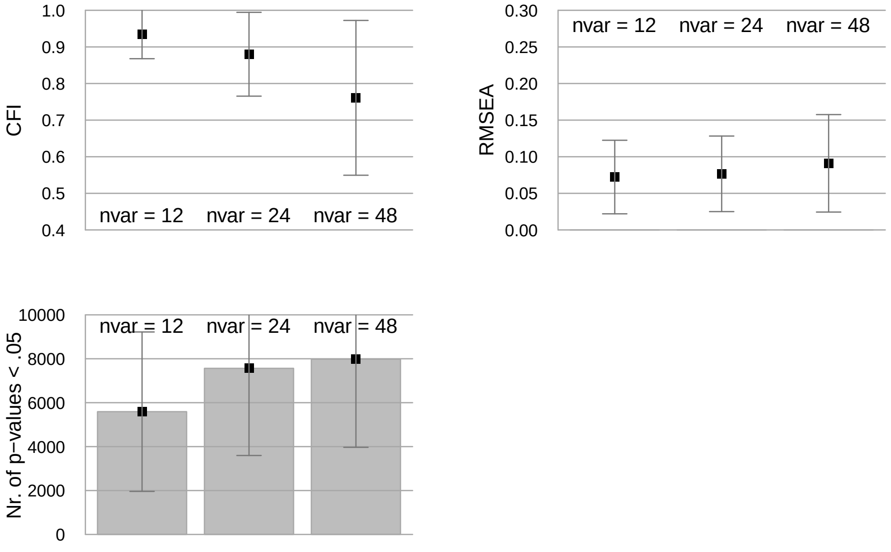


Figure 75. Five-categorical data: Main effect of number of variables on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (HOMALS).

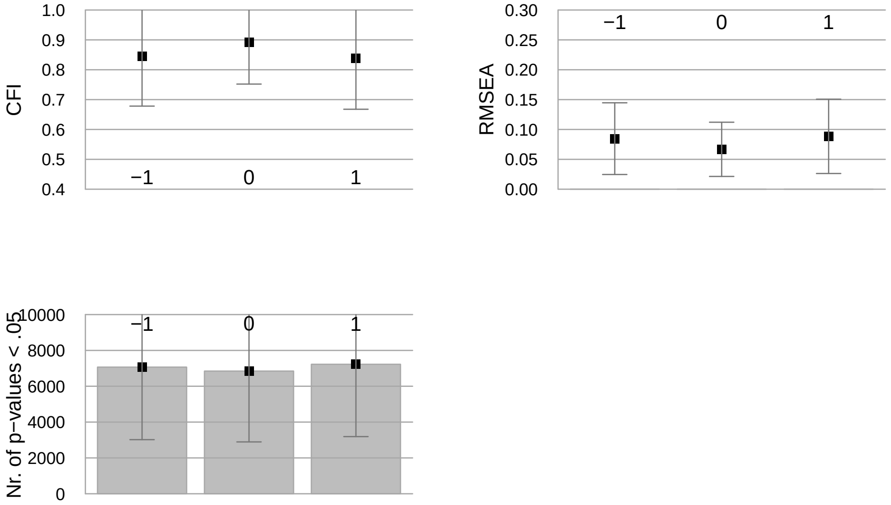


Figure 76. Five-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (HOMALS).

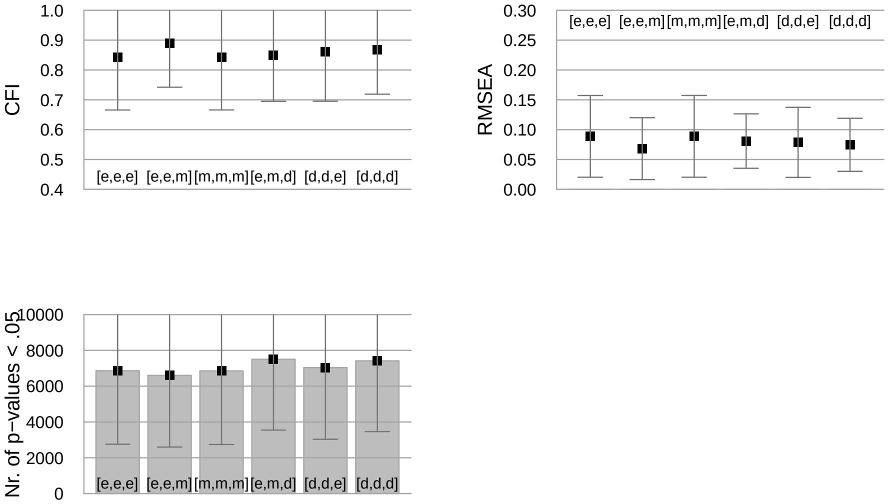


Figure 77. Five-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (HOMALS).

E Individual Analysis of the Nominal Response Model (Graphics)

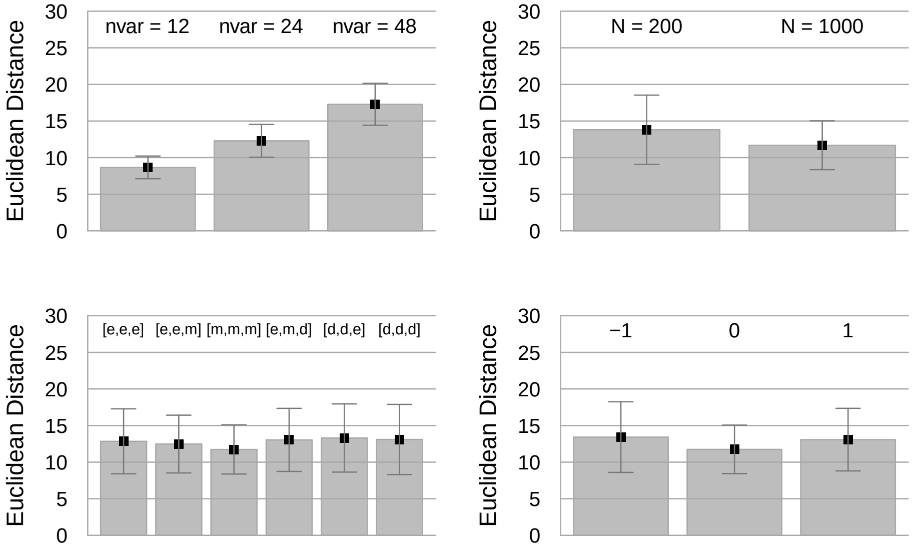


Figure 78. Five-categorical data: Main effect of independent variables on Euclidean distance (NRMV1). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

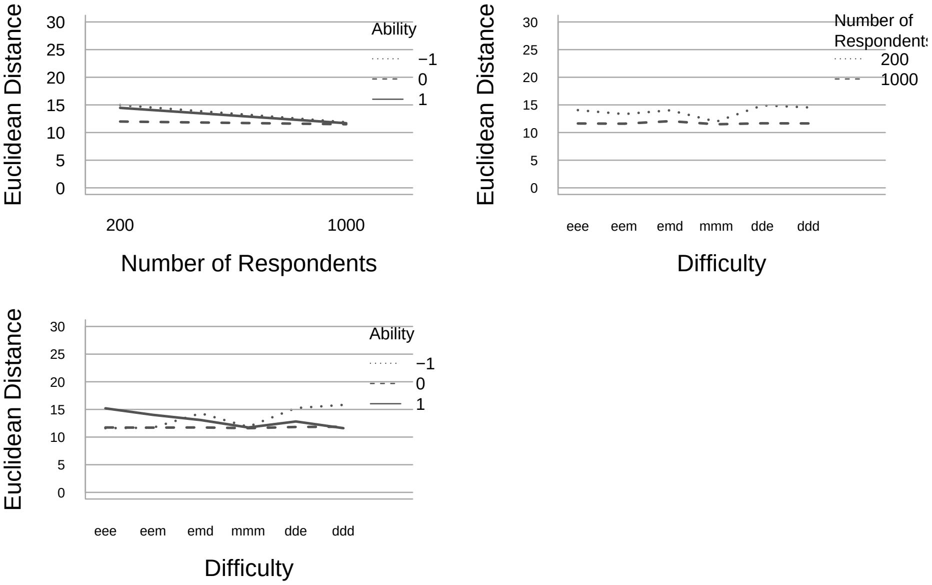


Figure 79. Five-categorical data: Interaction effects of independent variables on Euclidean distance (NRMV1). Top left panel: Interaction of number of respondents and ability; Top right panel: Interaction of number of respondents and difficulty; Bottom left panel: Interaction of ability and difficulty.

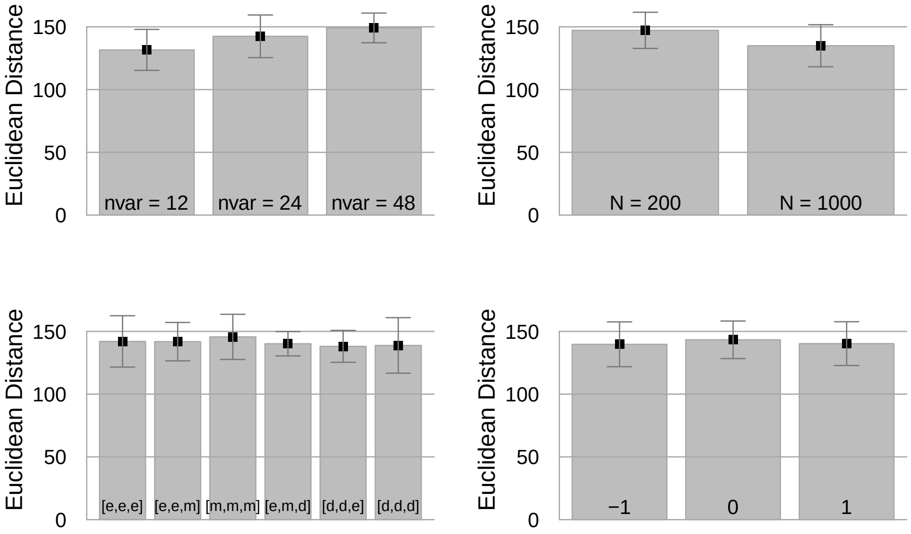


Figure 80. Five-categorical data: Main effect of independent variables on Euclidean distance (NRMV2). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

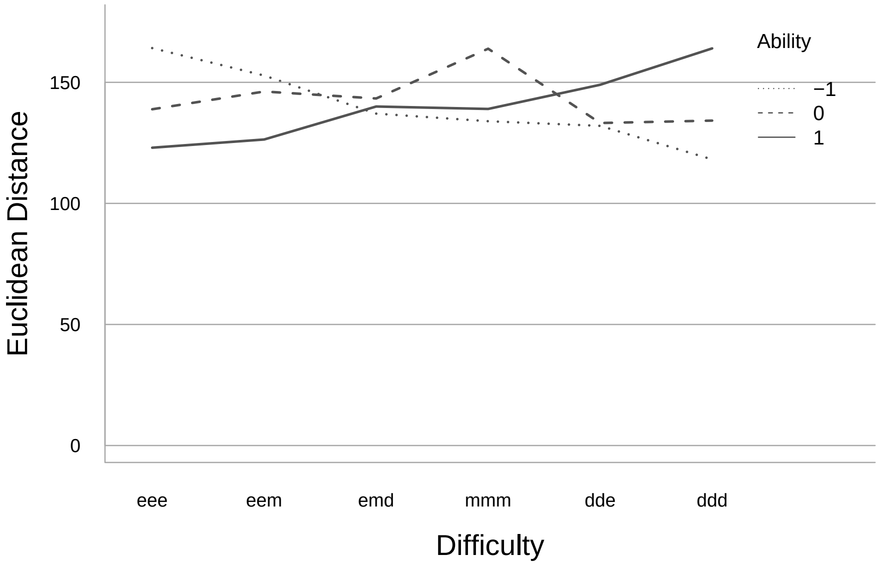


Figure 81. Five-categorical data: Interaction effects of ability and difficulty on Euclidean distance (NRMV2).

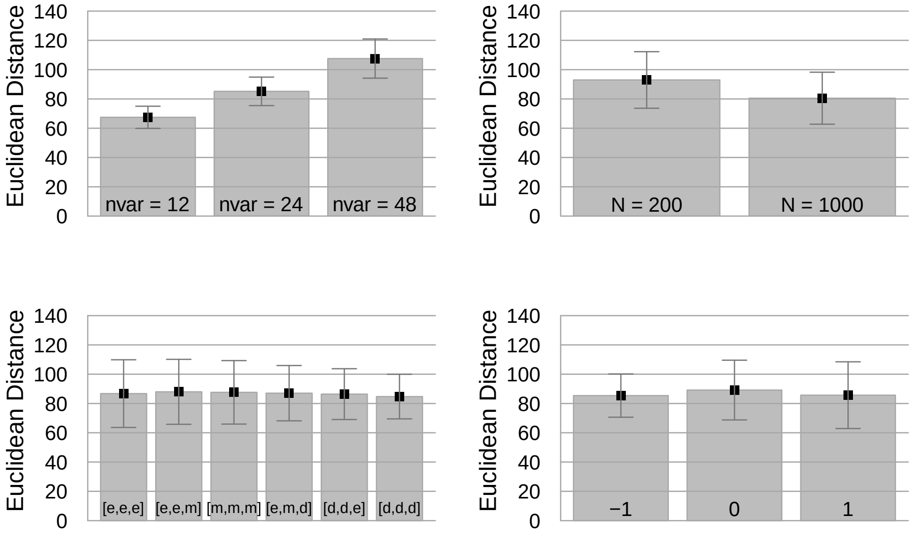


Figure 82. Five-categorical data: Main effect of independent variables on Euclidean distance (NRMV3). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

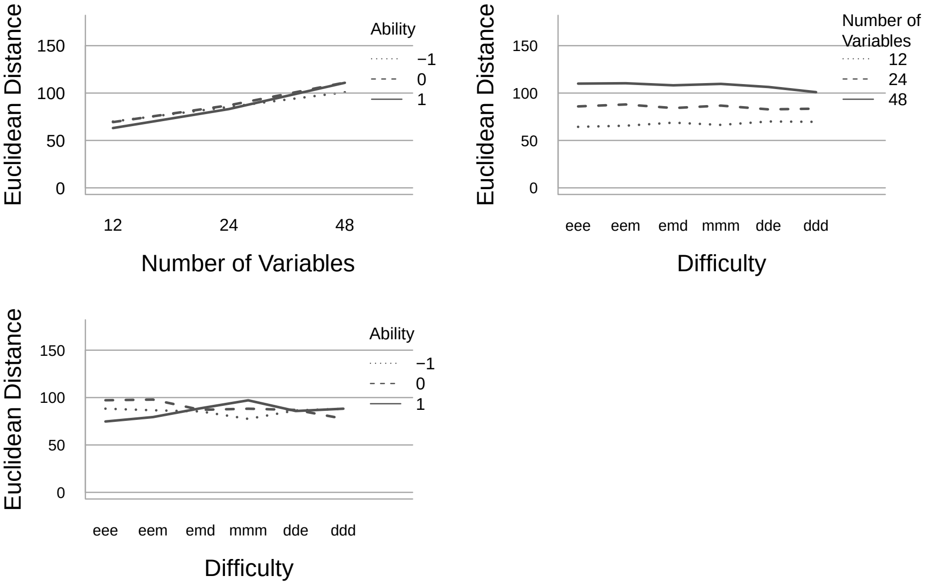


Figure 83. Five-categorical data: Interaction effects of independent variables on Euclidean distance (NRMV3). Top left panel: Interaction of ability and number of variables; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

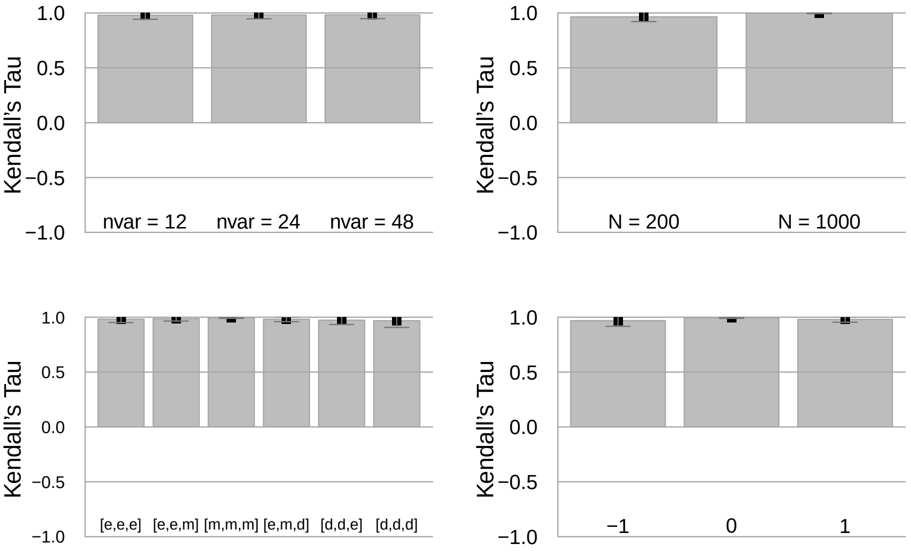


Figure 84. Five-categorical data: Main effect of independent variables on Kendall's τb (NRMV1). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

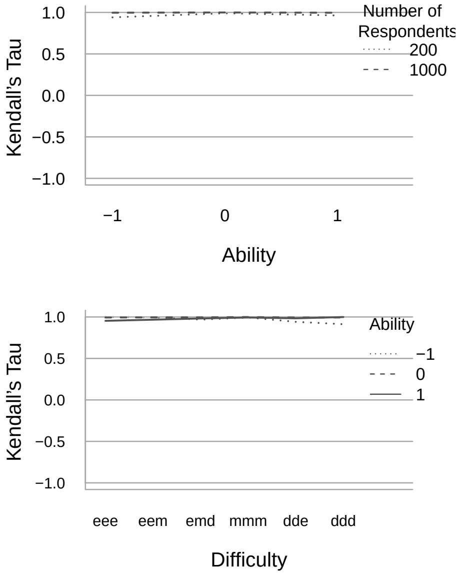


Figure 85. Five-categorical data: Interaction effects of independent variables on Kendall's τb (NRMV1). Top left panel: Interaction of number of respondents and ability; Top right panel: Interaction of number of respondents and difficulty; Bottom left panel: Interaction of ability and difficulty.

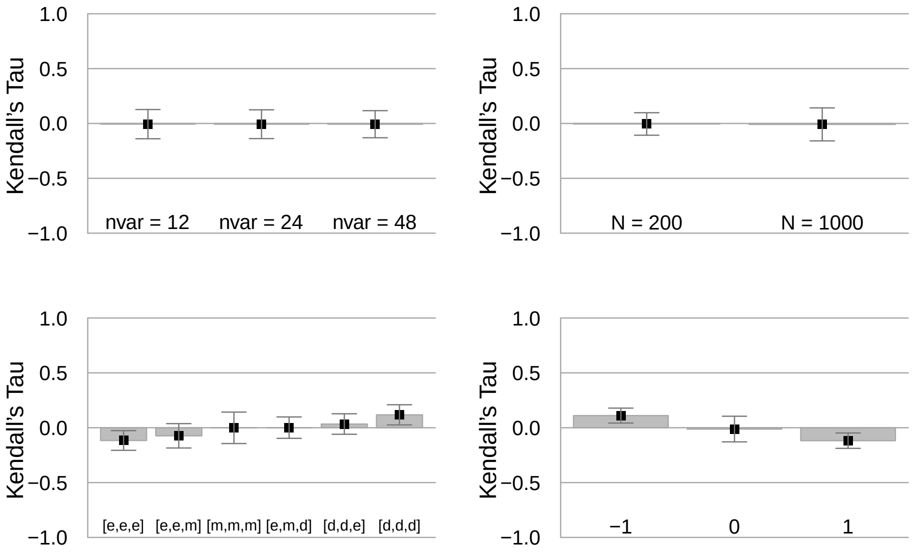


Figure 86. Five-categorical data: Main effect of independent variables on Kendall's τb (NRMV2). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

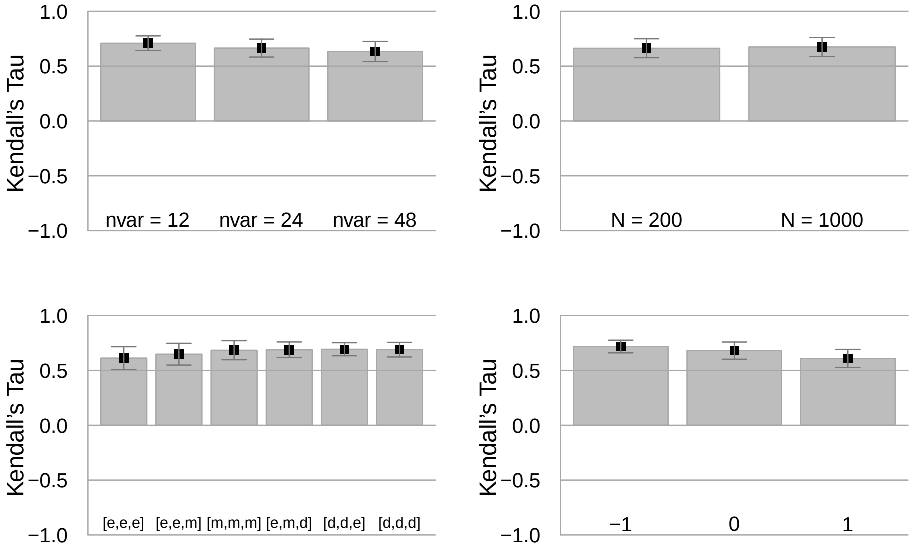


Figure 87. Five-categorical data: Main effect of independent variables on Kendall's τb (NRMV3). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

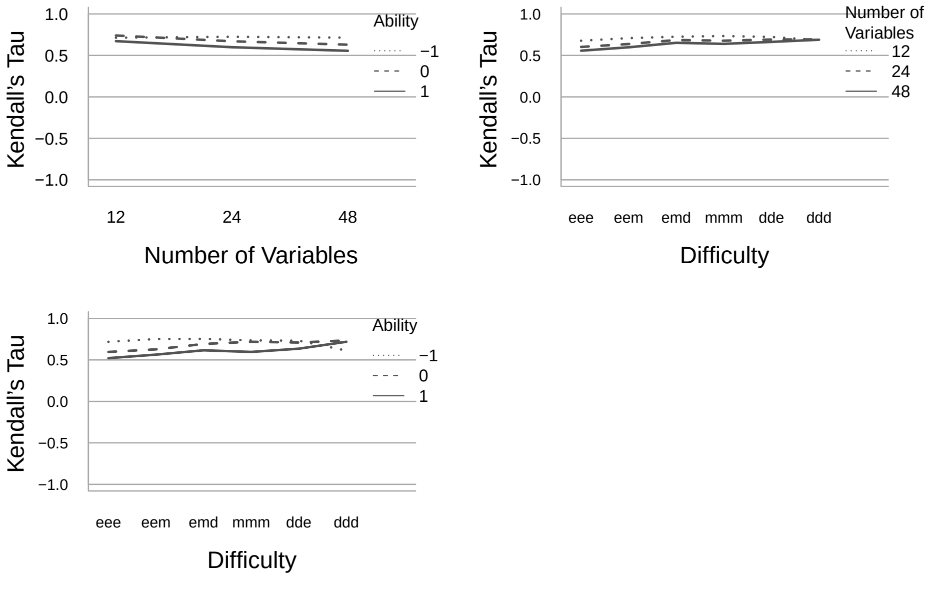


Figure 88. Five-categorical data: Interaction effects of independent variables on Kendall's τb (NRMV3). Top left panel: Interaction of ability and number of variables; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

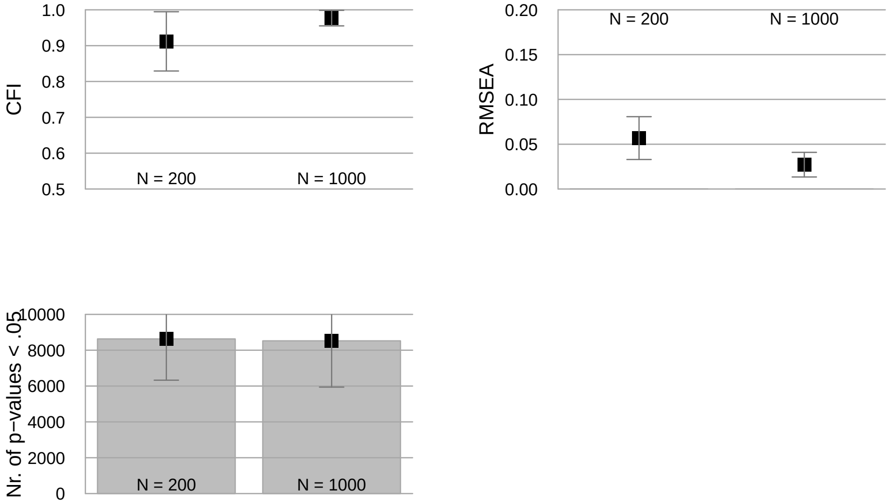


Figure 89. Five-categorical data: Main effect of number of respondents on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV1).

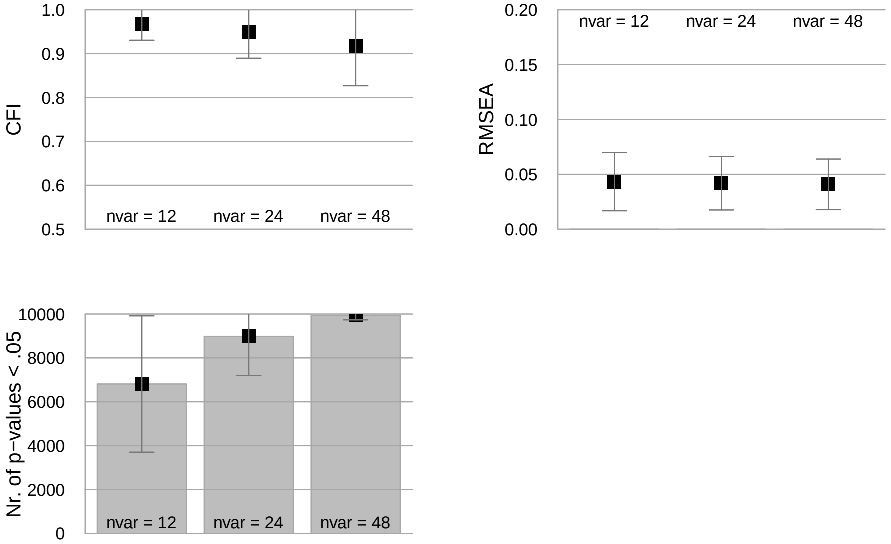


Figure 90. Five-categorical data: Main effect of number of variables on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV1).

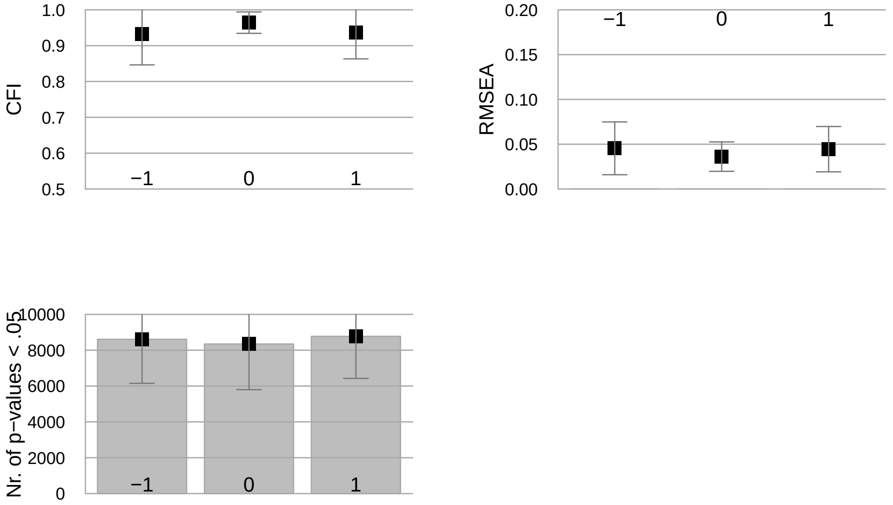


Figure 91. Five-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV1).

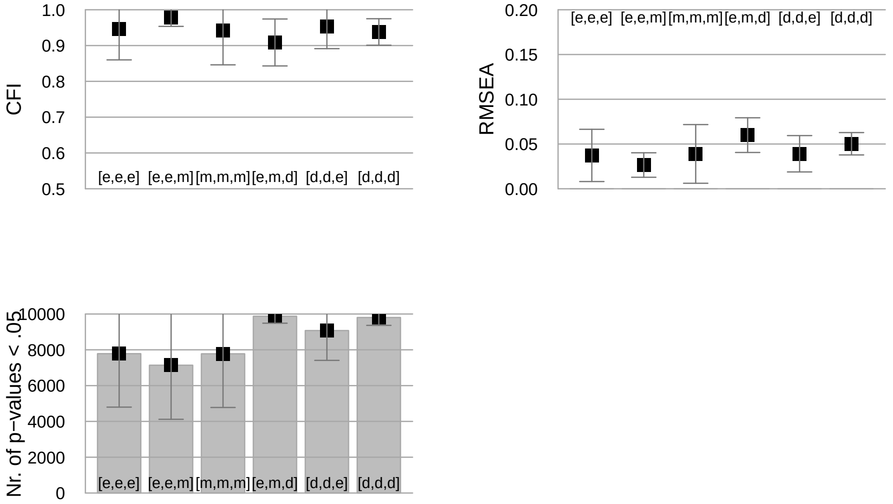


Figure 92. Five-categorical data: Main effect of difficulty on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV1).

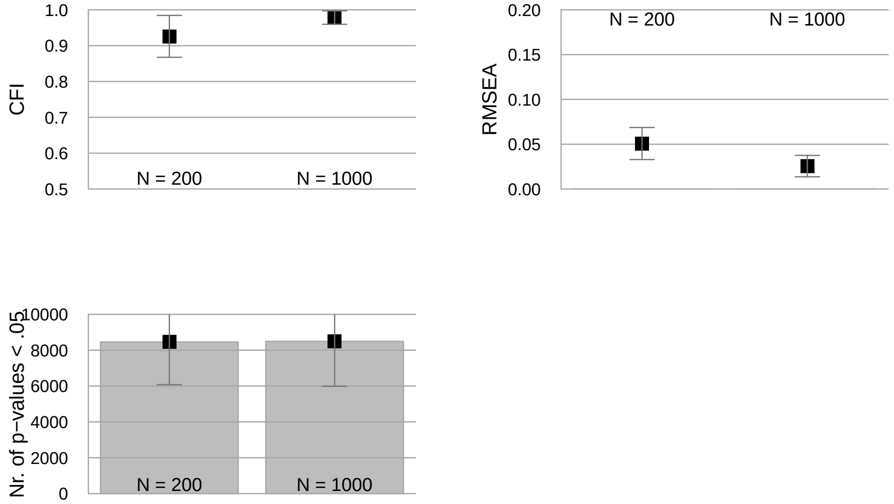


Figure 93. Five-categorical data: Main effect of number of respondents on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV2).

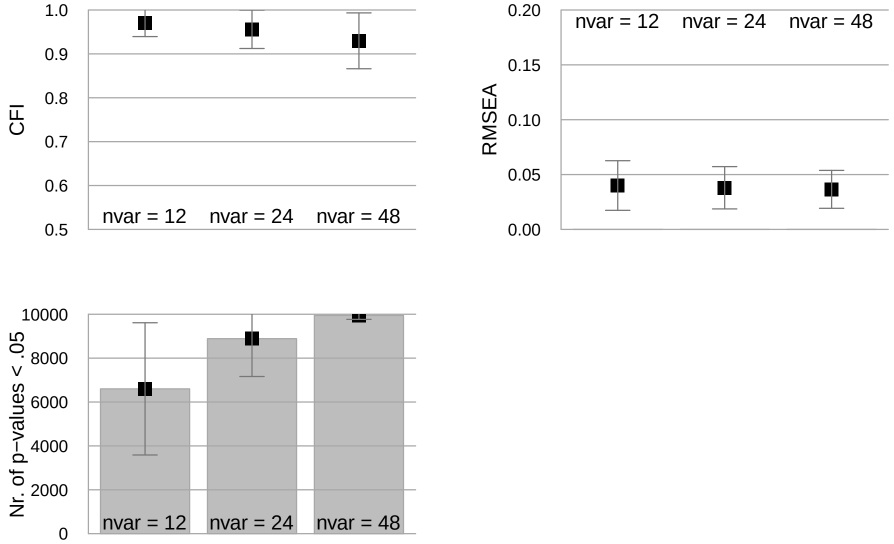


Figure 94. Five-categorical data: Main effect of number of variables on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV2).

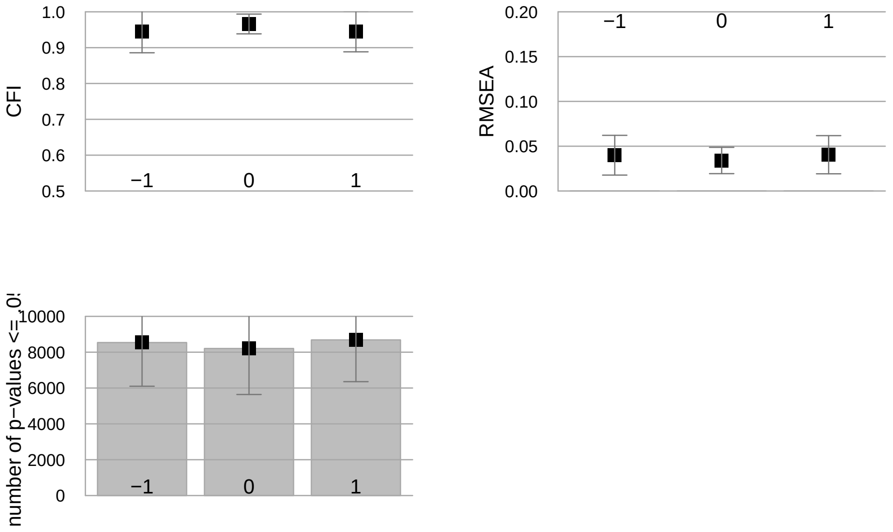


Figure 95. Five-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV2).

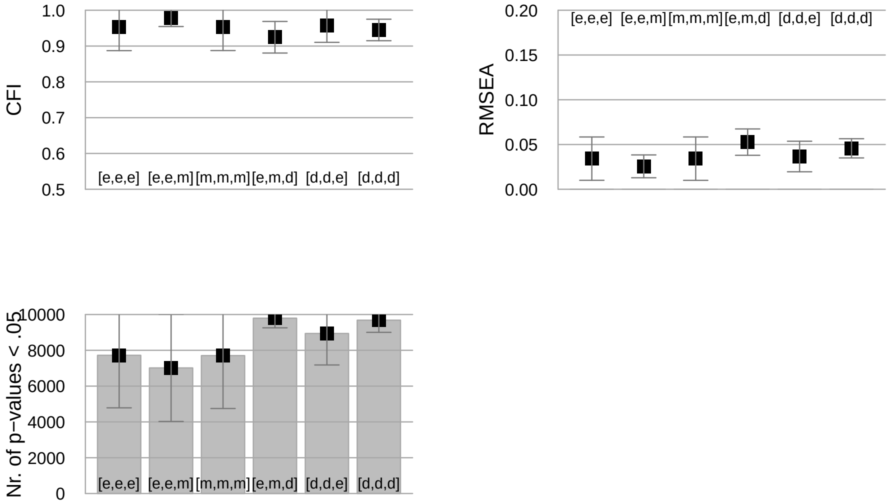


Figure 96. Five-categorical data: Main effect of difficulty on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV2).

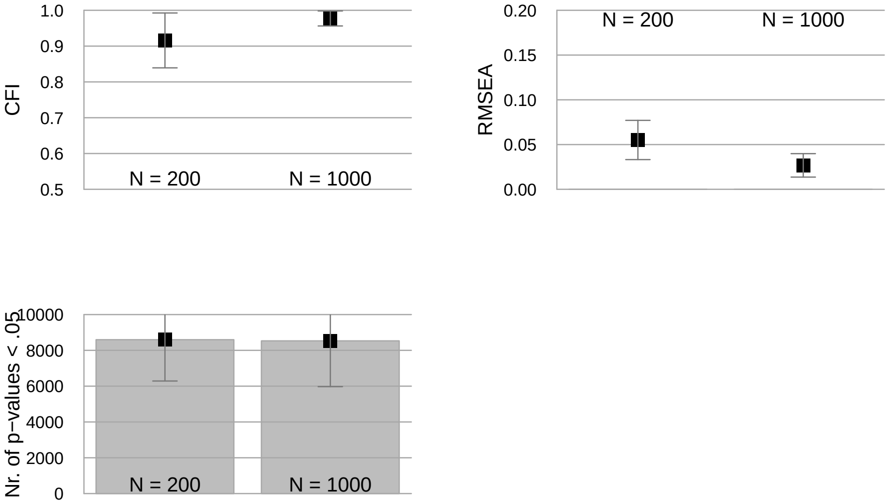


Figure 97. Five-categorical data: Main effect of number of respondents on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV3).

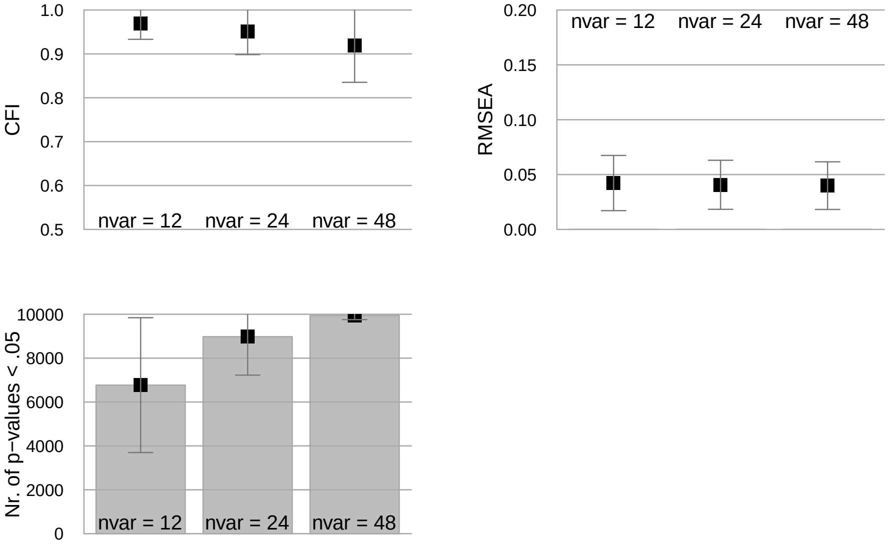


Figure 98. Five-categorical data: Main effect of number of variables on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV3).

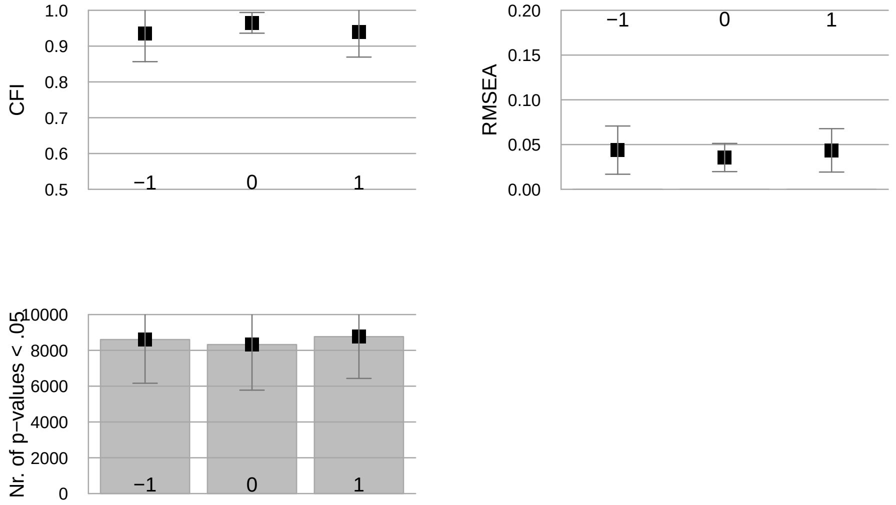


Figure 99. Five-categorical data: Main effect of ability on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV3).

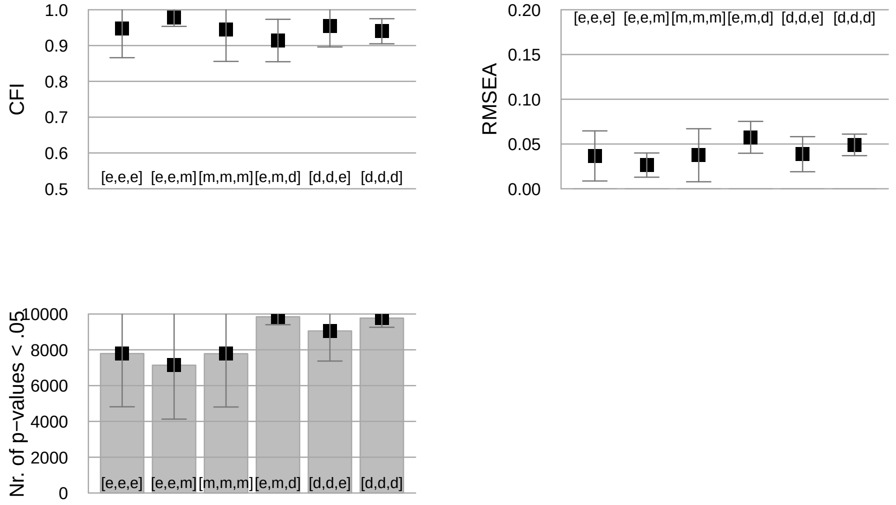


Figure 100. Five-categorical data: Main effect of difficulty on CFI (top left panel), RMSEA (top right panel) and number of p-values < .05 (bottom left panel) (NRMV3).

F Comparison of Empirical Scoring Methods (Graphics)

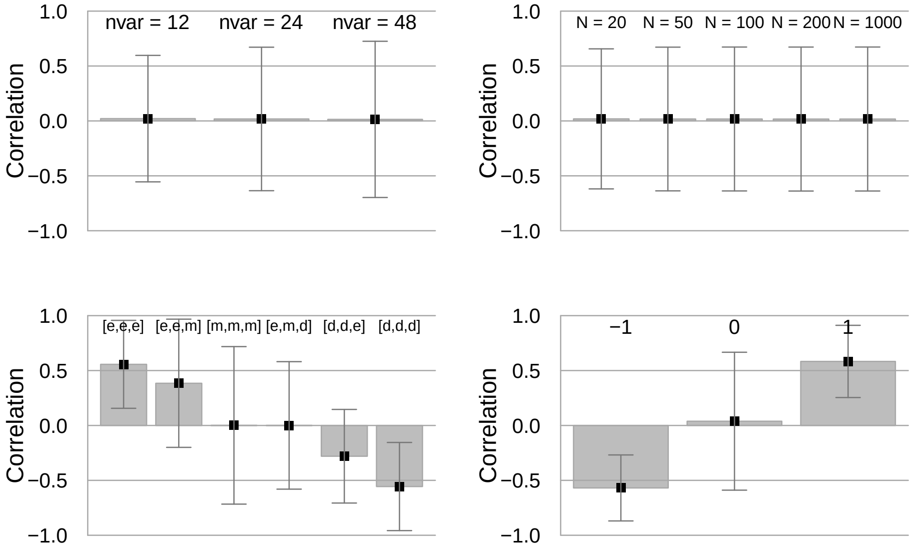


Figure 101. Two-categorical data: Main effects of independent variables on correlation (Mode CBM). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

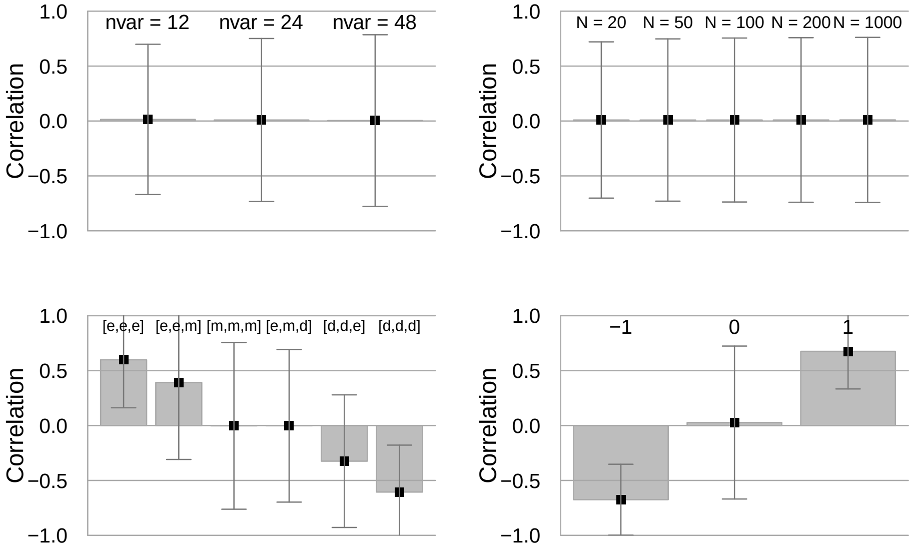


Figure 102. Five-categorical data: Main effects of independent variables on correlation (Mode CBM). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

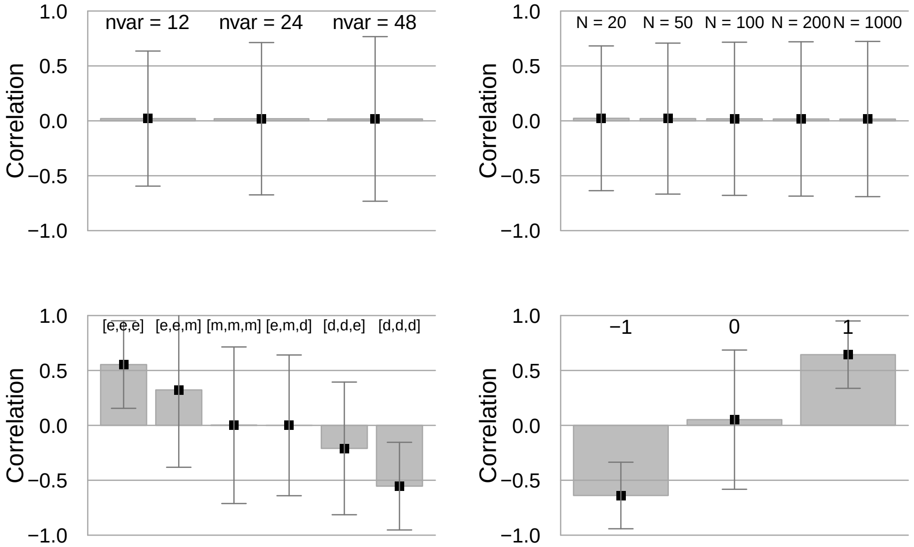


Figure 103. Two-categorical data: Main effects of independent variables on correlation (Proportion CBM). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

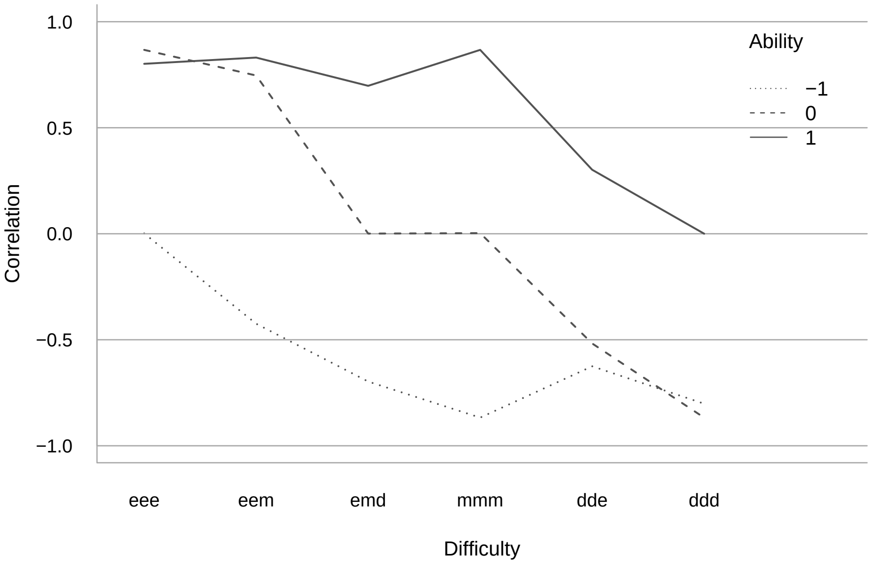


Figure 104. Two-categorical data: Interaction effect of ability and difficulty on correlation (Mode CBM).

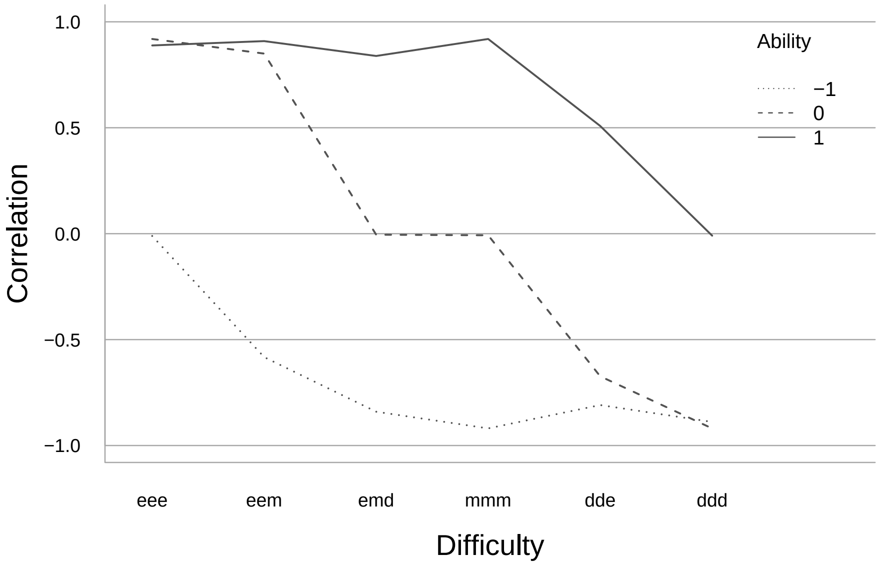


Figure 105. Five-categorical data: Interaction effect of ability and difficulty on correlation (Mode CBM).

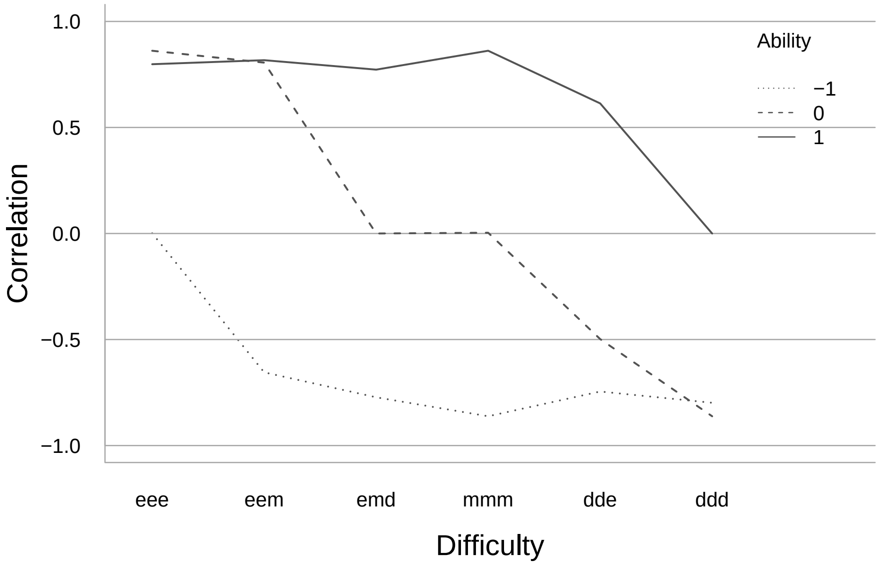


Figure 106. Two-categorical data: Interaction effect of ability and difficulty on correlation (Proportion CBM).

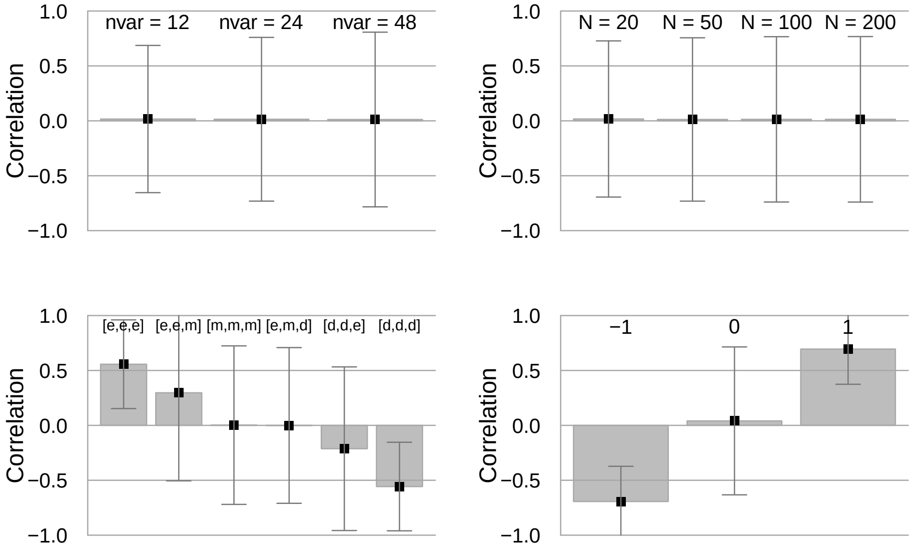


Figure 107. Two-categorical data: Main effects of independent variables on correlation (Consensus Analysis). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

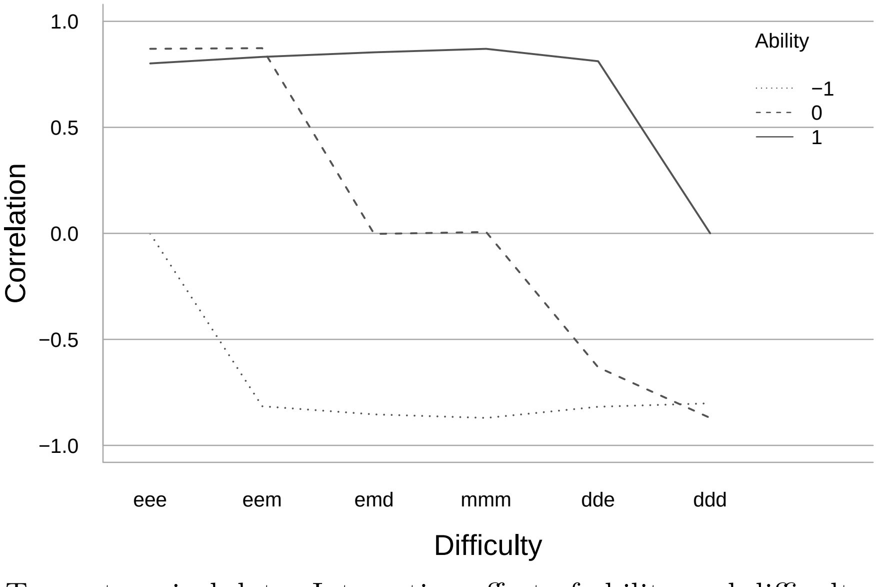


Figure 108. Two-categorical data: Interaction effect of ability and difficulty on correlation (Consensus Analysis).

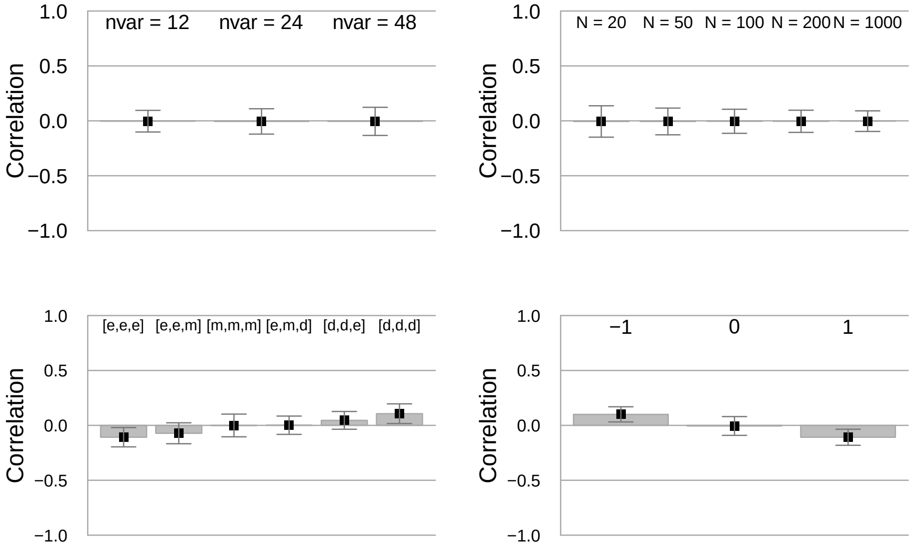


Figure 109. Two-categorical data: Main effects of independent variables on correlation (HOMALS). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

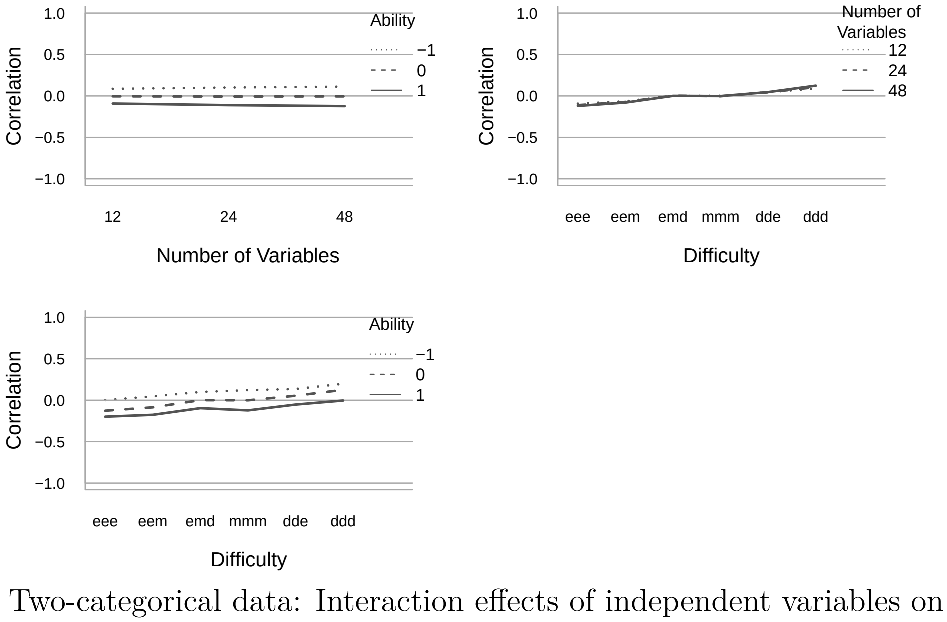


Figure 110. Two-categorical data: Interaction effects of independent variables on correlation (HOMALS). Top left panel: Interaction of ability and number of variables; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

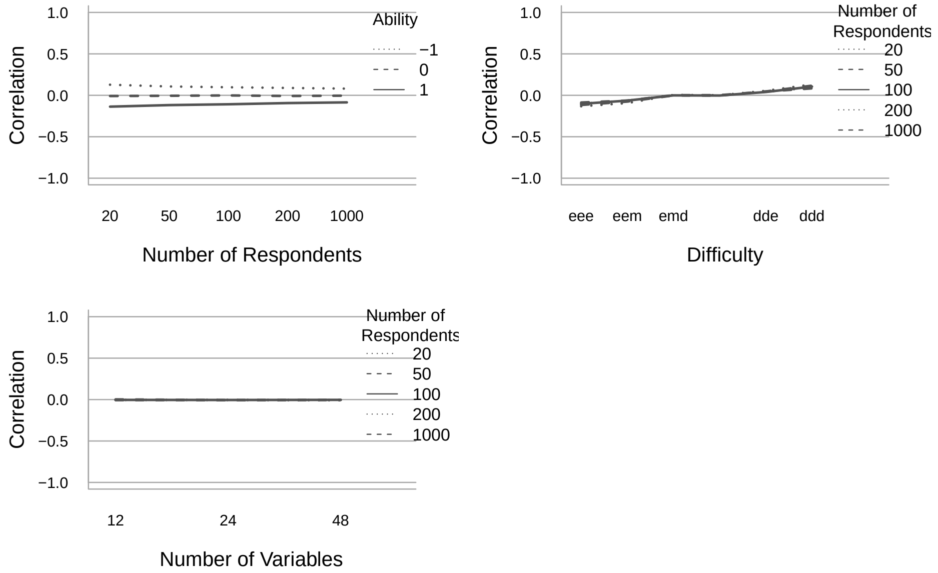


Figure 111. Two-categorical data: Interaction effects of independent variables on correlation (HOMALS). Top left panel: Interaction of ability and number of respondents; Top right panel: Interaction of number of respondents and difficulty; Bottom left panel: Interaction of number of respondents and number of variables.

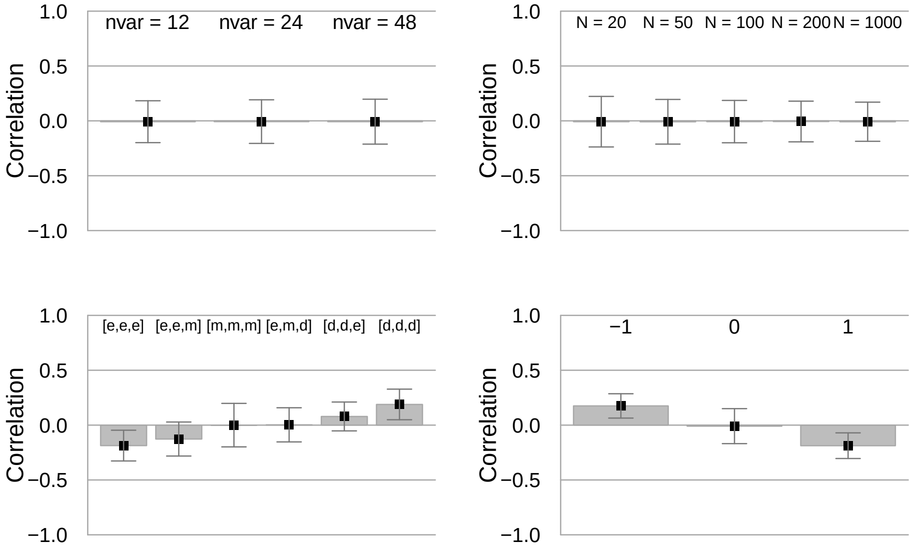


Figure 112. Five-categorical data: Main effects of independent variables on correlation (HOMALS). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

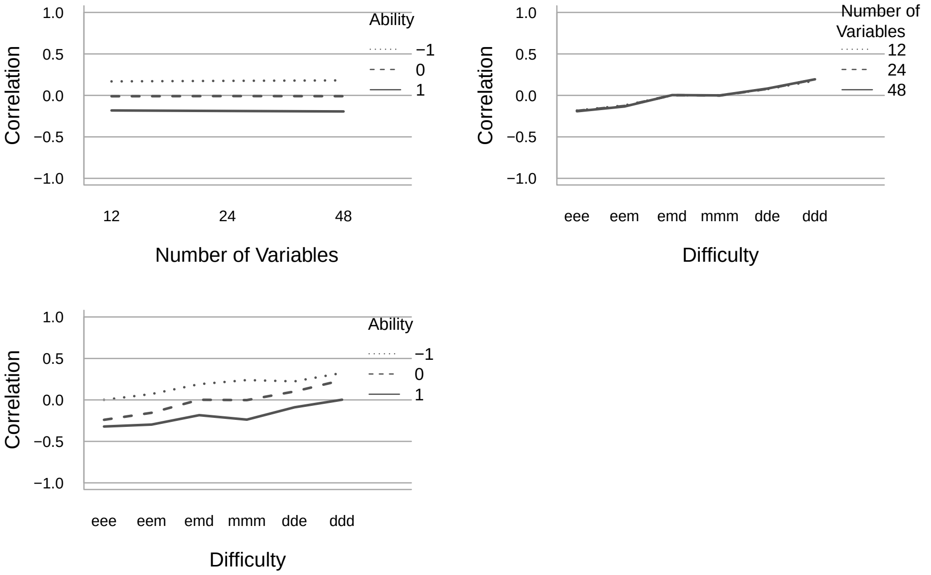


Figure 113. Five-categorical data: Interaction effects of independent variables on correlation (HOMALS). Top left panel: Interaction of ability and number of variables; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

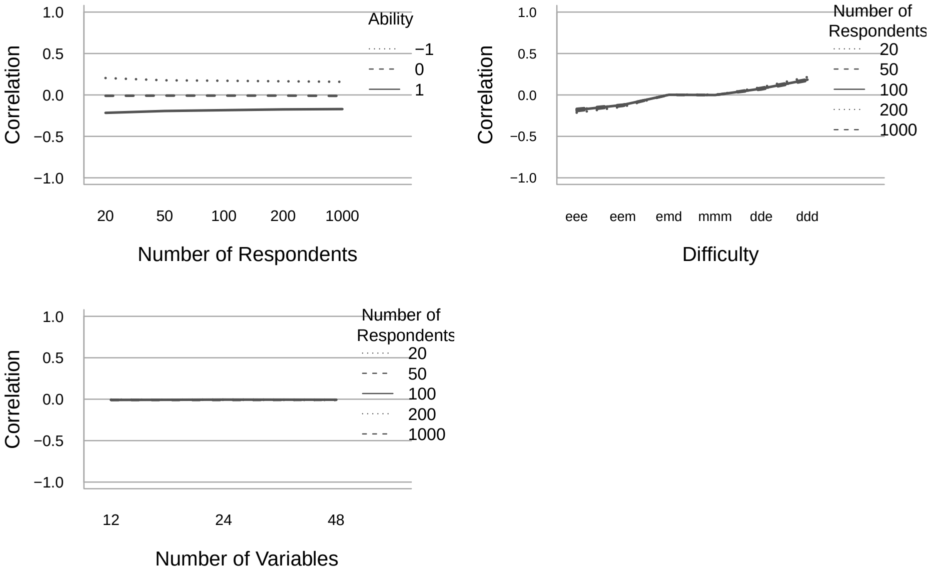


Figure 114. Five-categorical data: Interaction effects of independent variables on correlation (HOMALS). Top left panel: Interaction of ability and number of respondents; Top right panel: Interaction of number of respondents and difficulty; Bottom left panel: Interaction of number of respondents and number of variables.

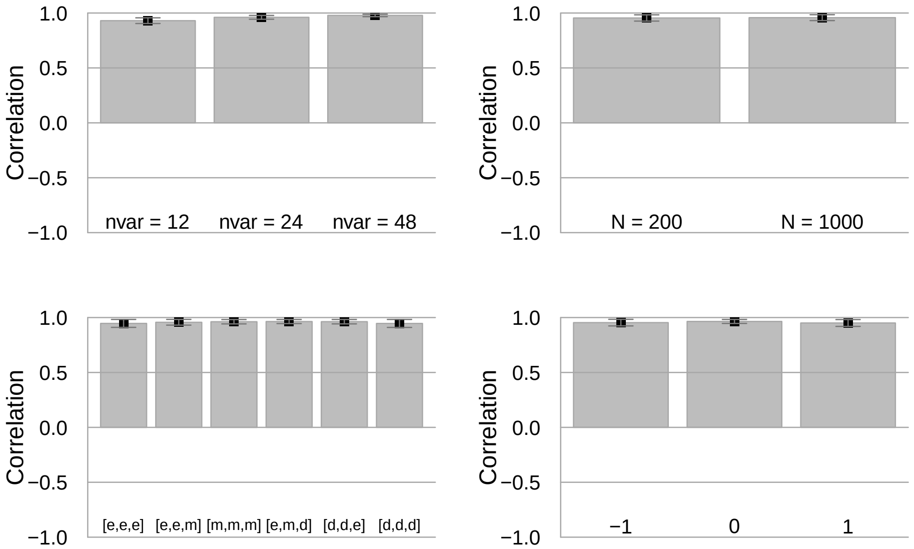


Figure 115. Five-categorical data: Main effects of independent variables on correlation (NRM V1). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

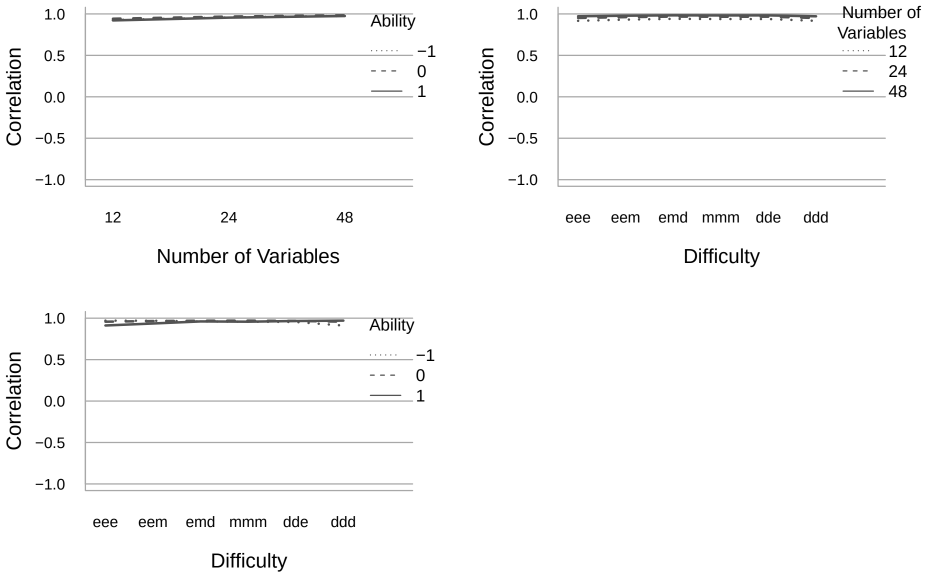


Figure 116. Five-categorical data: Interaction effect of independent variables on correlation (NRM V1). Top left panel: Interaction of ability and number of variables; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

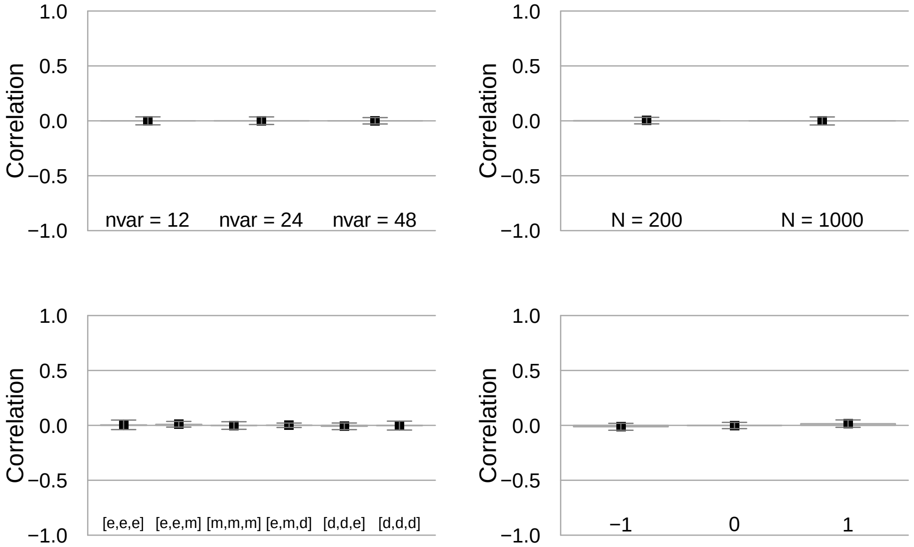


Figure 117. Five-categorical data: Main effects of independent variables on correlation (NRM V2). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

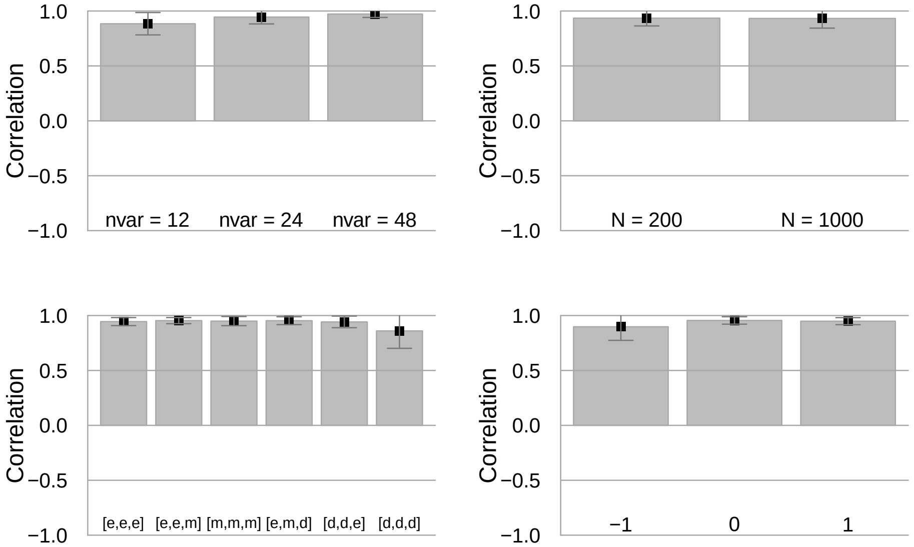


Figure 118. Five-categorical data: Main effects of independent variables on correlation (NRM V3). Top left panel: Number of variables; Top right panel: Number of respondents; Bottom left panel: Difficulty; Bottom right panel: Ability.

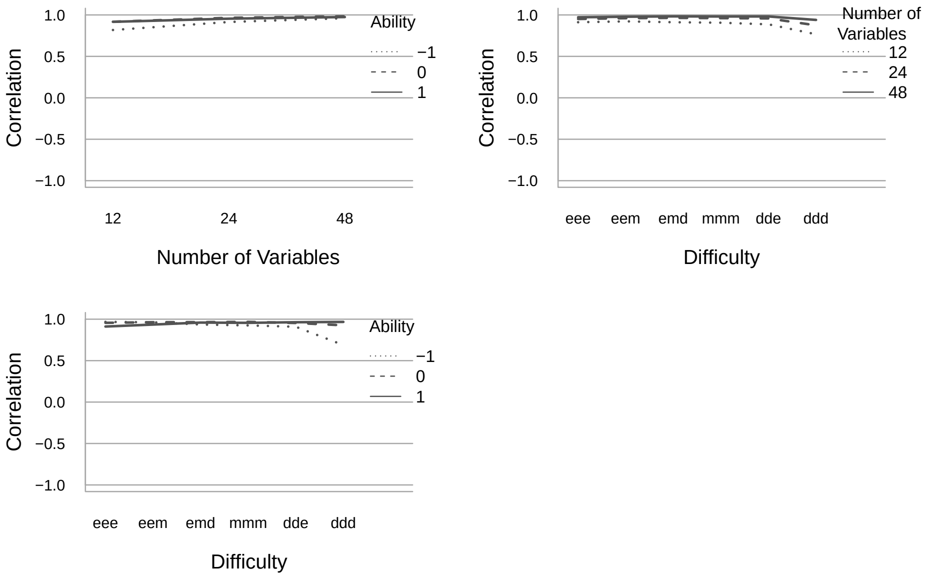


Figure 119. Five-categorical data: Interaction effect of independent variables on correlation (NRM V3). Top left panel: Interaction of ability and number of variables; Top right panel: Interaction of number of variables and difficulty; Bottom left panel: Interaction of ability and difficulty.

G TIMSS 2011 Study (Tables)

Table 38

Mean and Standard Deviation of Correlation Between Ability Estimates for Different Ability Levels for Sub Samples of Data Set MVI

	Mode CBM	Prop. CBM	CA
low ability	$M = -.13$ $(SD = .12)$	$M = -.13$ $(SD = .11)$	$M = .01$ $(SD = .14)$
medium ability	$M = .79$ $(SD = .08)$	$M = .83$ $(SD = .05)$	$M = .76$ $(SD = .13)$
high ability	$M = 1.00$ $(SD = .00)$	$M = .98$ $(SD = .02)$	$M = 1.00$ $(SD = .00)$

Note. Mode CBM = Mode Consensus Based Measurement; Prop. CBM = Proportion Consensus Based Measurement; CA = Consensus Analysis.

Table 39

Mean and Standard Deviation of Correlation Between Ability Estimates for Different Ability Levels for Sub Samples of Data Set MVD

	Mode CBM	Prop. CBM	CA
low ability	$M = -.11$ $(SD = .13)$	$M = -.17$ $(SD = .18)$	$M = .05$ $(SD = .20)$
medium ability	$M = .79$ $(SD = .01)$	$M = .84$ $(SD = .04)$	$M = .79$ $(SD = .06)$
high ability	$M = 1.00$ $(SD = .00)$	$M = .99$ $(SD = .01)$	$M = 1.00$ $(SD = .00)$

Note. Mode CBM = Mode Consensus Based Measurement; Prop. CBM = Proportion Consensus Based Measurement; CA = Consensus Analysis.

Table 40

Mean and Standard Deviation of Correlation Between Ability Estimates for Different Sample Sizes for Sub Samples of Data Set MVI

	Mode CBM	Prop. CBM	CA
$N = 20$	$M = .57$ ($SD = .47$)	$M = .58$ ($SD = .46$)	$M = .62$ ($SD = .41$)
$N = 50$	$M = .52$ ($SD = .56$)	$M = .53$ ($SD = .53$)	$M = .53$ ($SD = .47$)
$N = 100$	$M = .57$ ($SD = .52$)	$M = .55$ ($SD = .54$)	$M = .57$ ($SD = .51$)
$N = 200$	$M = .56$ ($SD = .53$)	$M = .57$ ($SD = .52$)	$M = .65$ ($SD = .39$)
$N = 1000$	$M = .56$ ($SD = .49$)	$M = .58$ ($SD = .51$)	

Note. Mode CBM = Mode Consensus Based Measurement; Prop. CBM = Proportion Consensus Based Measurement; CA = Consensus Analysis.

Table 41

Mean and Standard Deviation of Correlation Between Ability Estimates for Different Sample Sizes for Sub Samples of Data Set MVD

	Mode CBM	Prop. CBM	CA
$N = 20$	$M = .59$ ($SD = .51$)	$M = .56$ ($SD = .53$)	$M = .62$ ($SD = .48$)
$N = 50$	$M = .56$ ($SD = .51$)	$M = .53$ ($SD = .55$)	$M = .58$ ($SD = .48$)
$N = 100$	$M = .57$ ($SD = .48$)	$M = .54$ ($SD = .55$)	$M = .63$ ($SD = .41$)
$N = 200$	$M = .56$ ($SD = .50$)	$M = .56$ ($SD = .52$)	$M = .63$ ($SD = .39$)
$N = 1000$	$M = .54$ ($SD = .53$)	$M = .55$ ($SD = .55$)	

Note. Mode CBM = Mode Consensus Based Measurement; Prop. CBM = Proportion Consensus Based Measurement; CA = Consensus Analysis.