MODELING PANELIST CONSISTENCY IN BOOKMARK STANDARD-SETTING STUDIES

by

Joseph Fitzpatrick

Submitted to the graduate degree program in Educational Psychology and Research and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson Dr. William Skorupski

Dr. Neal Kingston

Dr. Vicki Peyton

Dr. Jonathan Templin

Dr. Lisa Wolf–Wendel

Date Defended: December 11th, 2017

The Dissertation Committee for Joseph Fitzpatrick

certifies that this is the approved version of the following dissertation:

MODELING PANELIST CONSISTENCY IN BOOKMARK
STANDARD-SETTING STUDIES

Chairperson Dr. William Skorupski

Date approved: December 12th, 2017

Acknowledgements

I owe thanks to so many people for the completion of this dissertation. First, I am forever grateful to Dr. William Skorupski, my academic advisor, for his excellent guidance, invaluable suggestions, and hours and hours of patient and insightful discussion. I could not have asked for a better teacher and mentor, and I know that I will continue to consider his teaching and advice throughout my professional career.

I would also like to thank the members of my dissertation committee. Dr. Neal Kingston's lectures changed the way I thought about testing, and I am also grateful for the support I received from the Center for Educational Testing and Evaluation, of which Dr. Kingston was director when I began my graduate study. Dr. Vicki Peyton is a wonderful instructor, and her advice and support in helping me deliver my first college lecture made me a better speaker and—I hope—a better teacher. Dr. Jonathan Templin provided expert advice and posed challenging questions as we began thinking about this study, and I know the final work is much better for his input. I am grateful to Dr. Lisa Wolf–Wendel for devoting her time and expertise to my committee, and to Dr. Kelli Thomas for her help during the proposal phase of this study.

I would also like to thank Dr. Marianne Perie, who provided me not only with means of financial support for much of my time at the University of Kansas, but also with some of the most valuable professional experiences of my graduate career.

Most importantly, I would like to thank my family, and in particular my wife, Jennifer. She was enthusiastic in encouraging me to begin this journey, and lovingly supportive in helping me to finish it. Our children, Frances and Murphy, make me so happy, and inspire me every day. Finally, thanks to my parents, for all their love, support, and patience.

Abstract

The purpose of this study is to develop a psychometrics of standard-setting for the Bookmark standard-setting procedure. Using simulated and real data, the error variance of Bookmark cutscores is modeled as a function of both within-group and between-group variance. Fully Bayesian methods are then used to estimate the total error variance associated with Bookmark cutscores. The results of the study suggest that the estimates of error variance for the proposed method tend to be larger than those calculated using traditional methods. Consequently, the confidence intervals were more likely to include the “true” cutscore in the simulation study. Because the result of a standard-setting study is almost always a recommendation rather than a final cutscore, one goal is to provide policymakers with a more accurate representation of the amount of uncertainty around a panel’s recommendation.

Table of Contents

List of Tables

List of Figures

Chapter I: Introduction

Statement of Problem

The Bookmark standard-setting method has become an increasingly popular method for setting cutscores on large-scale educational tests in the United States. Federal policy under the Elementary and Secondary Education Act of 1965, recently reauthorized as the Every Student Succeeds Act (ESSA), requires that each state's academic standards include at least three performance levels into which students may be classified by the statewide assessment system. Those assessment systems are almost always asked to serve multiple purposes, including providing measures of student growth and allowing for differentiated feedback to be provided at the school, district, and state levels. In this testing environment, the Bookmark method has been attractive to states in part because it can easily incorporate multiple cutscores as well as a high degree of input from the teachers expected to be most familiar with student performance.

Although the Bookmark method has been in use for two decades, relatively little research exists to support the inference that the cutscores recommended by panelists truly represent the expected scores of examinees at each intended performance level. The Bookmark method requires panelists to make estimates of estimates of performance for examinees within a very narrow range of ability, but few studies exist to demonstrate that panelists are actually able to do so with much accuracy (and the studies that do exist tend to suggest the opposite).

Purpose

The purpose of this study is to develop a psychometrics of cutscores that are set using the Bookmark method. A model-based approach to evaluating Angoff standard-setting results was introduced by Skorupski and Fitzpatrick (2014) and Skorupski, Zhao, Fitzpatrick, and Chen

(2015). Their general framework, which the authors termed “Cutscore Distribution Theory” (CDT), enabled more accurate estimates of the error associated with recommended cutscores than have been calculated using traditional approaches. CDT models the uncertainty associated with cutscores as a function of both panelist consistency and accuracy, while traditional approaches tend to focus only on consistency.

The results of standard-setting studies nearly always serve as a recommendation to a policy-making body (such as a state department of education) rather than the final cutscore(s). The goal in modeling the error around cutscores, therefore, is to estimate more accurate confidence intervals around panelists’ recommendations. Because the Bookmark method has become increasingly popular in large-scale educational testing, this study aims to extend CDT to Bookmark standard-setting studies.

Data and Variables

This study uses both simulated and real data. The real data were collected as part of an operational standard-setting study for a statewide general summative assessment program in English Language Arts (ELA) and Mathematics. In both the simulation and real data studies, the dependent variables are the standard errors of measurement of the recommended Bookmark cutscores calculated using different model of error. The primary independent variables are the amounts of within-group and between-group variance of the panelist ratings.

Research Questions

The following research questions are addressed in this study:

1. How will the width and location of the 95% credible intervals of the model-based compare with the 95% confidence intervals from the traditional method?

2. How similar will cutscore estimates from the model-based method be to those calculated using the traditional method?
3. How does changing the total number of panelists or the number of groups affect the estimates from the model-based and traditional methods?

Hypotheses

Modeling both within- and between-group variance will allow more accurate estimation of the error associated with cutscore recommendations than can be estimated using only the overall variance. Based on the results of previous studies with the Angoff standard-setting method (Skorupski & Fitzpatrick, 2014; Skorupski et al., 2015) cutscore error is expected to increase as panelists are less consistent, and as the number of panelists increases.

Research Expectations

Although most methodological research on standard-setting over the past several decades points out how difficult it is for panelists to estimate conditional item difficulties accurately, few established standard-setting methods attempt to account for panelist inaccuracy when calculating and reporting recommended cutscores. The confidence intervals around cutscore recommendations are typically viewed as a window within which a policymaking body may set the final cutscore. This research is expected to show how that window may narrow or widen as Bookmark panelists are more or less accurate, respectively, in their estimates of performance for borderline examinees. The traditional method of calculating standard errors of Bookmark cutscores uses only the consistency among the full sample of panelists' Bookmark placements. This study attempts to capitalize on the fact that Bookmark panelists typically work in small groups or tables in order to model the overall variance as coming from two sources: variation among panelists at the same table, and differences between tables.

Summary

This research addresses the problem of estimating the error variance associated with cutscores set using the Bookmark standard-setting method. Current methods of estimating this error account for only within-sample variability. As a result, the confidence intervals within which final cutscores are selected may be too narrow, thereby reducing the chance that the final cutscores are closely aligned with the actual level of performance the cutscores are intended to reflect. The aim of this study is to take advantage of the structure of a typical Bookmark sample in order to construct a model of cutscores that allows for more accurate estimation of their uncertainty. One primary goal of this research is to allow practitioners to establish cutscores more closely aligned with panelists' estimates of borderline performance.

Chapter II: Literature Review

Standard-Setting Background

Purpose of standard-setting. Standard-setting in psychological measurement typically refers to setting cutscores on tests for the purposes of making decisions (e.g., Cizek, Bunch, & Koons, 2004). Loomis defines standard-setting as “a process of translating a performance standard to a score scale” (2012, p. 108). Generally speaking, standard-setting methods describe a set of procedures for defining construct-based performance categories and then translating those to actual scores on the test (Kane, 1994). The categorization may be as simple as a pass/fail decision, or it may involve placing examinees into one of several classifications (e.g., “Below Basic,” “Basic,” “Proficient,” or “Advanced”). Cizek (1996) emphasizes that standard setting is a research endeavor that should be evaluated with the same rigor applied to any other research study, including the choice of method and its implementation. Unlike many other methodological questions in psychological measurement, however, research on standard-setting methods is complicated by the fact that researchers tend not to expect much consistency across implementations of a particular method. Indeed, much of the existing work on standard-setting explicitly states that there is no “true” cutscore to be discovered by a standard-setting panel. Rather, standard-setting panelists are expected to use their individual opinions and expertise to inform their judgments. As Cizek et al. (2004) explain:

> It is now a widely accepted tenet of measurement theory that the work of standard-setting panels is not to search for a knowable boundary between categories that exist. Instead, standard-setting procedures enable participants to bring to bear their judgments in such a way as to *translate* policy decisions…into locations on a score scale; it is these translations that create the effective performance categories. This translation and creation

> are seldom, if ever, purely statistical, impartial, apolitical, or ideologically neutral activities. (pp. 32–33, emphasis in original)

The idea of translation underscores the need to differentiate terminology related to cutscores. Kane (1994) carefully distinguishes “passing scores” from “performance standards.” The passing score (i.e., cutscore) is on the scale of the test; it is intended to correspond with some adequate level of the construct being measured (i.e., the performance standard). In other words, “the performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version” (Kane, 1994, p. 426). This distinction appeared to be blurred in early tests that claimed to be “criterion-referenced,” in which acceptable performance was defined, for example, as getting two-thirds of the items correct (Glass, 1978).

Arbitrariness of cutscores. There can be a great deal of variability across different standard-setting methods and panels. As Linn (2003) concluded:

> The variability in the percentage of students who are labeled proficient or above due to the context in which the standards are set, the choice of judges, and the choice of method to set the standards is, in each instance, so large that the term proficient becomes meaningless. (p. 12)

Linn (2003) argued against the “standards-based reporting” that has become commonplace in educational measurement. He maintained that we should reserve standard-setting for situations in which the need for performance categories is clear (e.g., licensure or certification). However, Linn acknowledges that as long as laws dictate the setting and reporting of performance categorizes, they will continue to be used. As such, many researchers have proposed ways of evaluating the standard-setting process to ensure that it is as defensible as possible (e.g., Engelhard, 2011; Hambleton, 2001; Hambleton & Pitoniak, 2006; Jaeger, 1988; Kane, 1994; Lee & Lewis, 2008; Reckase, 2006).

Kane (1994) identifies two assumptions related to passing scores and performance standards: the descriptive assumption and the policy assumption. Evidence supporting these two assumptions is part of the interpretive argument for the validity of any interpretations made on the basis of the passing scores. The descriptive assumption—that the cutscore corresponds to some specified level of performance—is not inherently arbitrary. To the extent that the test is reliably measuring a construct on which there is a well-defined performance standard, it may be reasonable to assume that a given performance standard can be tied to a specific point on the test score scale. In contrast, the policy assumption claims simply that "the performance standard is appropriate, given the purpose of the decision" (p. 435). Kane argues that the "unavoidable arbitrariness in standard setting" (p. 427) stems from this second assumption. For many psychological tests there is not a good reason to value one performance standard over another that is slightly higher or slightly lower. Kane defines this source of arbitrariness as separate from the actual translation of a performance standard to the scale of the test. To say that a cutscore is arbitrary is essentially to say that the policy assumption is not well met.

Accordingly, much of the criticism of standard-setting is criticism of the policy assumption itself. Glass (1978) argues that it is impossible to specify any criterion-based threshold that makes a meaningful distinction between performances. According to Glass, the fields of psychology and education have not adequately defined the concept of "mastery" or expertise in a way that allows us to tell masters or experts from non-masters or non-experts reliably. Setting standards on tests, in Glass's opinion, implies precise measurement of a construct that is vague at best and certainly not well understood. Therefore, such categorizations are inherently arbitrary.

Glass (1978) made the same distinction that Kane (1994) did much later: that there is a difference between the cutscore and the performance standard. He felt that the imprecision of language about criterion-referenced measurement at the time created the impression that cutscores are more grounded than they really are. Glass's point was not that it is impossible to make comparisons among test performances, but simply that it is impossible to place a cutscore or threshold at a point that differentiates acceptable from unacceptable performance in anything but an arbitrary way. To place a cutscore is to dichotomize (what we usually assume to be) a continuum. If we are unsure that clearly defined categories exist, according to Glass, it makes little sense to search for their demarcation line. Glass argues that human behaviors are more complex than can be accounted for by our construct definitions, and the fact that there will always be people who are misclassified by an established test threshold poses a serious threat to the validity of cutscore interpretations.

Other experts, however, do not believe standard-setting is necessarily undermined by the fact that performance standards are often arbitrary (e.g., Hambleton, 1978; Popham, 1978; van der Linden, 1982). In a response to Glass (1978), Popham (1978) points to wine tasting and art criticism as just two examples of judgmental tasks that can still be accomplished with a high degree of consistency and defensibility. Popham argues that standard-setting is judgmental but not "mindless and capricious" (p. 298). Hambleton (1978), too, argues that the term "arbitrary" need not have a negative connotation as it pertains to standard-setting. As van der Linden (1982) summarized:

> There are many other instances in which arbitrary choices have to be made in which deliberate, defensible results are obtained. What should be avoided is capricious standard setting, that is, standard setting in which the learning objectives are inconsistently

> translated into the cutoff score and, in fact, erratic standards of mastery are obtained. (p. 295)

In addition, Glass's (1978) argument that categorization is arbitrary does not make it unnecessary. Mehrens and Cizek (2012), for example, claim that many of the decisions that are informed at least in part by test scores are inherently categorical. They point to college admissions, employee selection, and professional licensure as just a few examples of cases when categorical decisions must be made. "Proponents of setting performance standards accept the position," they argue, "…that for a given purpose a level of knowledge, skill, or ability exists that is too low to tolerate" (p. 35). The judgmental, arbitrary aspect of standard-setting is in determining "how low is too low?"

Defensibility of cutscores. Reasonable people can have differing definitions of a performance standard, so there is no reason to expect them to agree on precisely the same test score (Kane, 1994). The purpose of any standard-setting method is not to locate the "true" cut score; in most cases no such score exists. Instead, we might simply hope "to show that the proposed standard is reasonable" (Kane, 1994, p. 457). Perhaps the most important product of any standard setting study, therefore, is a compelling body of evidence that the chosen cutscore is reasonably appropriate given the inferences we wish it to inform.

The evidentiary burden for a given cutscore depends to some degree on the stakes of the test. Tests that are used to deny employment or admission are more likely to face legal challenges, for example, and will require an extensive body of evidence that will stand up to rigorous examination in court.

In addition to considering issues of fairness and opportunity to learn, U.S. courts have examined whether professional standards (such as the *Standards for Educational and*

Psychological Testing (AERA, APA, & NCME, 2014; hereafter *Standards*) or the Equal Employment Opportunity Commission *Guidelines* (EEOC, 1978)) were adhered to when evaluating the defensibility of cutscores (Mehrens & Popham, 1992; Phillips, 2012). Thorough documentation of the standard-setting plans, processes, and decisions can help to show that cutscores (and the inferences they are used to inform) are legally and psychometrically defensible (Phillips, 2012). Contrary to Glass's (1978) argument, the courts have generally regarded judgmental standard-setting methods as acceptable (Mehrens & Popham, 1992).

It is important to point out that although standard-setting methods are not designed to "discover" the cutscore, this paper adopts the position of Cutscore Distribution Theory (CDT; Skorupski et al., 2015) that there is a parameter to be estimated:

> In standard setting, it is often assumed that there is no "true" cutscore; we rely on the informed judgment of expert panelists, but we don't necessarily conceive of a cutscore that "should" be found. CDT proceeds by rejecting that notion in favor of an asymptotic definition of a "true cutscore." The development of this idea is conceptually very similar to assumptions made in deriving Classical Test Theory. (p. 5)

In this sense, a panel's recommended cutscore may be viewed as an estimate of the cutscore that would be set by the population of qualified panelists (see Cizek (1993), however, for a criticism of this "parameter estimation paradigm"). Although not always highlighted by standard-setting research, the result of most standard-setting studies is a *recommended* cutscore that is presented to the policymaking body that makes the final decision. For example, the results of NAEP standard-setting studies are recommended cutscores taken to the National Assessment Governing Board, which decides the final cutscores (Loomis, 2012). The policymaking body typically considers several pieces of information in addition to the recommended cutscore itself, including impact data, performance on other similar tests, and panelists' self-evaluations of how well they understood the process (Loomis, 2012). Part of establishing defensible cutscores, therefore, is developing methodology that quantifies and minimizes error in the panelists' recommendations in order to provide the final decision-makers with the most useful information.

Standard-Setting Methods

Jaeger (1989) described standard-setting methods as either "test-centered" or "examinee-centered," a classification that Kane (1994; 1998) also incorporated[1]. Test-centered methods ask panelists to focus on the test or items on the test, while examinee-centered methods ask panelists to make decisions about examinees (Kane, 1994). This schema was expanded by Hambleton, Jaeger, Plake, and Mills (2000) to account for newer methods that were not easily classifiable as either test- or examinee-centered. Because the focus of the current study is on one particular test-centered method (Bookmark), this paper will use the Jaeger (1989) classification system for convenience, but see Hambleton and Pitoniak (2006) for a list of methods classified using one dimension of the Hambleton et al. (2000) schema.

Although Kane (1998) recommends using examinee-centered methods for performance tests and test-centered methods for more objectively scored tasks, he acknowledges that little research has been conducted to support this belief.

Examinee-centered methods. Examinee-centered methods ask standard-setting panelists to make judgments about the performance of examinees. After examinees have been classified, the cutscore is placed at the point that maintains the classifications. Because the focus of the

[1] Although Jaeger's classification of methods as either examinee- or test-centered is still widely cited in standard-setting research, Cizek and Bunch (2007) point out that all judgmental methods require panelists to consider both examinees and test items, and "methods differ primarily in the degree to which one or the other focus is explicitly emphasized" (p. 105). This paper will adopt Jaeger's classification scheme for the sake of custom.

current study is on the Bookmark method (a test-centered approach), this section will be limited to brief descriptions of a few of the most common examinee-centered procedures. The reader is referred to Cizek and Bunch (2007) and Cizek (2012) for more details on examinee-centered methods.

Body of Work method. The Body of Work (BoW; Kingston, Kahl, Sweeney, & Bay, 2001) method was developed specifically to be used for assessments with constructed-response or other complex tasks that are not well suited to traditional item-rating procedures. BoW panelists are asked to evaluate and make a holistic judgment about an examinee work sample (such as an essay or portfolio). According to Kingston and Tiemann (2012), the BoW method was developed to elicit judgments that teachers are already comfortable making (as opposed to the often-unfamiliar task of judging item difficulty). Zieky, Perie, and Livingston (2008) note that although the BoW method can work with mixed-format tests that include some multiple-choice items, it is ill-suited for tests with many such items.

The BoW method is an iterative procedure. In addition to a training round, the process includes a “range-finding” round and a “pin-pointing” round (e.g., Cizek & Bunch, 2007; Kingston & Tiemann, 2012; Zieky et al., 2008). BoW panelists work with entire samples of examinee work (e.g., essays or work products), assigning each sample to a performance level. In the range-finding round, panelists work independently. A range of possible cutscores is identified by reviewing actual test scores of the samples placed into each category. In the pin-pointing round, panelists repeat the process, but evaluate only samples with actual scores in the identified range. Kingston & Tiemann (2012) and Zieky et al. (2008) advise using new samples for the pin-pointing round, although Cizek and Bunch (2007) state that range-finding samples may be included. For both the range-finding and pin-pointing rounds, the panelists should not be aware of any of the actual scores of the work samples. Once BoW data are collected, logistic regression appears to be the preferred method for calculating actual cutscores (e.g., Cizek & Bunch, 2006; Kingston & Tiemann, 2012; Zieky et al., 2008).

Contrasting Groups method. The Contrasting Groups method (Berk, 1976) asks panelists first to classify examinees (typically a small subset of all test-takers) into groups (e.g., "masters/nonmasters," or "passers/failers"). The classification could presumably be done before or after the examinees take the test, so long as the panelists do not know the examinees' test scores when they make their judgments. Those classifications are then used to create empirical score distributions for each group. A cutscore is calculated that minimizes the misclassifications of those same examinees, assuming that the judged classifications are accurate.

As noted by Cizek and Bunch (2007), the described procedure is not likely to be so straightforward in practice. Limited sample sizes and messy raw score distributions may necessitate some sort of smoothing procedure before selecting a cutscore. Furthermore, Berk's (1976) original description of the method described creating mastery/nonmastery groups on the basis of whether students had received instruction in the domain to be tested. Although the method can be (and has been) extended to rely on panelist judgment to classify students, doing so is certain to strain our assumption that the classification is accurate. Hambleton, Swaminathan, Algina, and Coulson (1978) pointed out that the Contrasting Groups procedure is only useful to the extent that this assumption is met.

Borderline Group method. The Borderline Group method (Zieky & Livingston, 1977) is similar to the Contrasting Groups method, except that panelists identify one group of "borderline" examinees rather than classifying each examinee into one of two groups. The cutscore is the median test score for the group of borderline examinees. Although the cutscore depends only on the borderline group, panelists are usually asked to classify examinees into three groups (e.g., adequate, inadequate, and borderline; Cizek & Bunch, 2007; Livingston & Zieky, 1982).

As with the Contrasting Groups method, the Borderline Group method still relies heavily on the ability of the panelists to identify borderline examinees accurately. The effect of misclassified examinees—whose scores may lie at the extremes of the group distribution—is mitigated by using the median rather than the mean as the cutscore (Livingston & Zieky, 1982).

Test-centered methods. Test-centered methods ask standard-setting panelists to make judgments about aspects of the test itself (usually the items on the test). Typically, panelists review the test and try to predict the performance of a hypothetical examinee. This paper will not attempt to mention the many different test-centered methods and modifications to those procedures, but will rather describe a few of the most widely cited methods. For a more extensive listing and description of methods, see Cizek and Bunch (2007) and Cizek (2012).

Nedelsky method. In the procedure proposed by Nedelsky (1954), panelists are shown a multiple-choice item and asked to estimate how many of the options a minimally proficient examinee will *know* are incorrect. The reciprocal of the number of remaining options is summed across all items on the test to determine the cut score. That is, if a minimally proficient examinee will know one option is incorrect on a four-option multiple-choice item, then the reciprocal of the remaining options (1/3) represents the probability that he will answer the item correctly (and the sum of these reciprocals for the entire test represents his probable score). The technique is limited in that it can be applied only to multiple-choice items and that it assumes examinees will guess randomly among all plausible options. In a comparison of the Nedelsky and Angoff methods, van der Linden (1982) argued that this assumption of Nedelsky is unlikely to be met in all circumstances. Also, because there are a limited number of values a four- or five-option item can take, inconsistency among panelists can substantially affect the variability of the resulting cut scores.

Ebel method. The Ebel (1972) method first asks panelists to make a judgment about both the difficulty and the relevance of each item. The panelists categorize each item on each dimension such that every item is cross-classified in a table. In other words, each item would be placed in one of three difficulty categories (easy, medium, and hard) and one of four relevance categories (essential, important, acceptable, and questionable). The judgments are summarized in a table (in the current example, the table would have 12 cells: three difficulty categories by four relevance categories). Each individual item judgment is tabled; that is, if 5 panelists classify 50 items, the sum across all 12 cells in the table will be 250. After tabulating the judgments, the panelists are asked to judge what percentage of items in each cell the borderline examinee should answer correctly. For each cell in the table, that percentage is multiplied by the number of items in the cell, and that product is summed over all cells. Finally, the cutscore is calculated by dividing that sum by the total number of judgments (in this case, 250) to yield a percentage correct score.

Despite the ease of calculations with the Ebel method, it has several disadvantages. For example, Poggio (1984) reports that the process is very time-consuming, and that the inclusion of the “questionable” category can make panelists uneasy. As Cizek (2007) explains, it is reasonable for panelists to wonder why items they deem as having questionable relevance should be included on the exam, which may weaken the overall validity argument for using the test to classify examinees. However, as Meskauskas (1976) noted:

> One gathers the impression that Ebel is not committed to the particular descriptors used along the two dimensions. Many test constructors may wish to use somewhat different

> descriptors, as the inclusion, in a test, of a category of items judged of questionable relevance seems hard to defend. (p. 138)

Meskauskas also points out, though, that the relevance and difficulty dimensions are likely to be correlated, and that requiring separate judgments about the percentage of items in each category necessary to pass the test is “entirely arbitrary” (p. 138) without some evidence to support the practice. Cizek and Bunch (2007) report that the method is used mainly for classroom assessments in the fields of medicine and healthcare.

Angoff method. One of the most commonly applied standard-setting methods was proposed by Angoff (1971)[2], in which panelists are shown items on a test and asked to estimate whether a minimally proficient person would get that item correct. In a footnote, Angoff proposed a modification to this procedure, in which the panelists are asked to estimate the *probability* with which a minimally proficient person would select the correct answer. The sum of these conditional probabilities for all items on the test equals the estimated number of items correct for the minimally proficient examinee, which is then averaged across all panelists to determine a mean estimate. This probability-based modification has received the most attention in the literature and seems to be the most common in practice; this paper will use the term “Angoff method” to refer to the probability estimate method.

Angoff’s footnote describing the probability-based procedure provides a general outline but contains little detail:

> In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer the item correctly. The sum of these

[2] Although the procedure bears his name, Angoff attributed it to Ledyard Tucker (Cizek & Bunch, 2007).

> probabilities, or proportions, would then represent the minimally acceptable score. (Angoff, 1971, p. 515)

Because Angoff did not provide a more detailed description of the procedure, subsequent researchers have modified or clarified aspects of it to increase the consistency of the ratings and to help ensure the accuracy of the results. In a review of empirical studies on Angoff methods, Brandon (2004) identified a five-step process common among published Angoff-based methods:

1. select panelists,
2. train panelists,
3. define the performance criterion,
4. estimate performance of examinees at this level for each item, and
5. review empirical item information and discuss ratings.

In most studies, panelists work independently, assigning judgments without feedback from the researcher or other panelists. Brandon notes that the completion of step 5 typically concludes one “round” of standard setting, and that many researchers have implemented a two- or three-round process. Incorporating such an iterative procedure allows panelists to review and revise their original ratings after examining previous item performance data and/or discussing their ratings as a group.

Bookmark method. The Bookmark standard-setting method was developed as an IRT-based procedure that was particularly well-suited to the mixed format tests with multiple cutscores that became common in educational assessment in the 1990s (Lewis, Mitzel, Mercado, & Schulz, 2012; Mitzel, Lewis, Patz, & Green, 2001). In a Bookmark standard-setting study, panelists are presented with a booklet of items ordered from easiest to hardest and then asked to place a marker (or bookmark) to separate the items that minimally proficient examinees (MPEs) likely will have mastered from those that MPEs likely will not have mastered. The cutscore is then calculated using the IRT scale location of the item or items near the bookmark (Lewis et al., 2012).

The ordered item booklet. A primary feature of the Bookmark method is the ordered item booklet (OIB). The OIB contains one item per page, with the items ordered from easiest to hardest. Item difficulty is usually determined by IRT estimates (such as item *b*-parameters), but the Bookmark method can also be implemented without the use of IRT. For example, Buckendahl, Smith, Impara, and Plake (2002) described a Bookmark study that ordered items according to their *p*-values. No matter the metric used to estimate difficulty, the items within the OIB should be carefully selected to span the range of both content and difficulty that are represented by the test (Cizek & Bunch, 2007). The OIB may be a single complete test form or a representative sample of items from a larger item pool. Because cutscores are calculated by mapping item difficulty onto examinee ability, the OIB should be constructed so that no large “gaps” in difficulty exist between adjacent items. For this reason, items from an intact test form might need to be augmented with additional items to bridge gaps in item difficulty in an OIB (Cizek & Bunch, 2007). Mitzel et al. (2001) state that OIBs usually span a range of 80 to 110 score points, noting that this would make the OIB longer than most educational tests.

One advantage of the Bookmark method is that it can easily incorporate a variety of item types beyond traditional selected response items. Dichotomously scored items appear only once in the OIB, but polytomous items will appear once for each possible score point above zero (Cizek & Bunch, 2007). For example, an item with possible scores of 0, 1/3, 2/3, and 1 would appear as three separate pages in the OIB, with their locations within the booklet determined by the difficulty of obtaining each non-zero score point (estimated using a polytomous IRT model, for example).

Cizek and Bunch (2007) provide a detailed description of the OIB and the kind of information that is typically provided by the OIB. In addition to the item itself, each page in the OIB may also provide the key to the item and statistical information (such as the scale value associated with the item's difficulty). If some items within the OIB have a common stimulus (e.g., a reading passage), then the stimulus material is provided to panelists as well.

A typical round of a Bookmark procedure. As is the case with many other standard-setting methods, a Bookmark study is an iterative process typically comprising three rounds (Mitzel et al., 2001). Although some activities may be included only in specific rounds (e.g., presentation of impact data) or vary slightly by round (e.g., discussion of bookmark placements with small table groups vs. with the whole sample), the general procedure of panelists placing their bookmarks is very similar in each round. Before panelists begin the actual Bookmark process, they should be very familiar with both the performance level descriptors that define typical students in each category, and they should also carefully develop "borderline performance" descriptors that further help them to define the target population (Zieky et al., 2008).

In each round, panelists review each item in the OIB and are encouraged to consider and discuss the knowledge and skills that are required to answer each item correctly. Mitzel et al. (2001) recommend that discussions in round 1 be limited to the concept of "mastery" rather than on actual performance levels in order "to make the first round of judgments as uncontaminated as possible by others' opinions" (p. 253). For polytomously scored items, panelists consider what is required to attain the given score point.

As panelists work through their OIBs, they are asked to place a bookmark to separate items that a borderline examinee would probably have mastered from items a borderline examinee probably would not have mastered. The notion of “probable mastery” should be carefully defined by the meeting facilitators to ensure panelists have a common definition (Zieky et al., 2008; see the next section for a discussion of response probability values). After panelists have placed their initial bookmarks, they should review at least a few more items beyond that point to ensure the items after the bookmark are generally not expected to be mastered by borderline examinees (Zieky et al., 2008). If panelists simply stop at the first encounter of a “non-mastery” item, the cutscore may be biased downward by a specific item. This is especially true if panelists view items within the OIB as being “out of order” in some regions (e.g., Cizek & Bunch, 2007).

After panelists have placed their bookmarks, they review and discuss their placements (either in small groups or as a whole panel) and are encouraged to build consensus about the knowledge and skills required of the items in the OIB. Impact data (i.e., the percentage of examinees in each category) are usually presented after round 2, based on the mean or median bookmark location of the group after the second round of bookmark placement (Mitzel et al., 2001). For larger studies, the discussion in round 3 shifts from small-table discussions to whole-panel discussion. After the round 3 discussion, panelists make their final bookmark placements, and the final recommended cutscores and impact data are presented to the panel (Mitzel et al., 2001). In each round, panelists may change the location of their bookmark placement or maintain the same placement from the previous round (Karantonis & Sireci, 2006).

Selection of a response probability value. Much of the research on the Bookmark method has focused on the idea of “mastery” as it is presented to Bookmark panelists. In a typical

Bookmark standard-setting study, panelists are asked to review items in the OIB and, for each item, ask themselves whether an MPE would have a 0.67 probability of answering it correctly (or, in the case of a polytomous item, obtaining at least that particular score). Lewis et al. (2012) argue that although any probability value could be used in theory, a response probability of 0.67 (RP67) is relatively easy for panelists to understand (e.g., “a two-in-three chance”) and is similar to the way teachers tend to think about mastery of a given task. However, those authors note that the choice of RP value has been a point of much discussion in the literature. Huynh (2006) argues that the RP value should correspond with the scale location of an examinee who would be expected to answer the item correctly, or the point at which the information of a correct response is maximized. For the Rasch and 2PL IRT models, that point corresponds to a probability of 0.67. Wang (2003) argued that a value of 0.5 (RP50) is more appropriate in a Rasch context, although Huynh (2006) explained that their results were not actually at odds and instead depend on the type of information one wishes to maximize.

In theory, the selection of an RP value should not affect the cutscore. Panelists using a higher RP value would place their bookmarks sooner in the OIB than panelists using a lower one, but the performance standard (and the test score to which it is translated) are theoretically identical (Hambleton & Pitoniak, 2006). In practice, however, differences can occur that demand careful consideration. For one thing, unless the Rasch model is used, the rank order of items by difficulty may not be the same at RP50 as at RP80, for example (Beretvas, 2004). Such differences may have an effect on where panelists place their bookmarks in the OIB. On a related note, the differences in scale locations between adjacent items in the OIB may not be the same for different RP values even if their difficulty ordering is the same, which could also affect the resulting cutscore calculation. Another problem is that Bookmark panelists may behave or view their task differently with different RP values. In order to place their Bookmarks in a location consistent with the desired performance standard, panelists need to understand and internalize the RP value being applied. To that end, using RP67 seems to be most appropriate. For example, panelists in a Mapmark standard-setting (a procedure similar to Bookmark) study felt more comfortable using RP67 than RP50 (Williams & Schulz, 2005). Similarly, a study by the National Research Council (2005) found that panelists preferred RP67 to either RP50 or RP80. After reviewing relevant literature, Lewis et al. (2012) conclude that "the once controversial question related to RP criteria seems to be reduced to considering values anchored for the most part between RP65 and RP70" (p. 235). For the sake of easier understanding by panelists, those authors recommend RP67 (i.e., 2/3 probability).

Methodological Research on Standard-Setting.

Regardless of the method selected, the results of a standard-setting study may still vary from study to study. The following sections highlight some of the methodological research factors that can influence the defensibility of cutscores.

Panelist selection. This paper will follow the lead of Jaeger (1991) and exclude standard-setting panelists "whose authority to establish test standards derives from position rather than qualification" (p. 3). As Jaeger explains, standards set by such panelists tend to be established normatively or simply by tradition. Criterion-referenced tests, on the other hand, are "deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (Glaser & Nitko, 1971, p. 653). Most high-stakes educational and professional tests require standards that are explicitly linked to performance standards. We rely on the judgment of experts to establish such links (Jaeger, 1991).

In selecting panelists for standard-setting, then, it becomes important to consider what makes one an expert. Jaeger (1991) reviewed standard-setting literature and identified several characteristics of expertise that may be required in various standard-setting applications. For example, in most educational situations, standard-setting panelists should be very familiar with both the content of the test and the population of students who will take it. In that case, experienced teachers of a particular subject at a particular grade-level may be most qualified to identify minimally competent performance. For a professional licensure examination, experts may be those professionals who already hold the license themselves but who are still familiar with the skills and abilities of entry-level practitioners (Jaeger, 1991).

In addition to considering whom to select, it is also important to consider how many experts will be on a standard-setting panel. As Jaeger (1991) points out, this is essentially a question of sample size. As such, he recommends that "the number of judges sampled should be sufficiently large so that the magnitude of most estimation errors will be tolerable" (p. 5). Using a variety of criteria and existing test scales, Jaeger shows that this number may range from around 13 to as many as 87. It may be unfeasible, of course, to use very large numbers of panelists, and the exact number of panelists required will depend somewhat on the stakes of the decision being made and the precision of the test itself (with higher-stakes tests likely requiring larger samples of panelists than lower-stakes ones). Berk (1996) reports that most standard-setting panels comprise between 5 and 20 panelists. Raymond and Reid (2001) reviewed much of the same literature and concluded that panels of 10 to 15 panelists would be acceptably dependable in most situations, although they admit that there will be exceptions to that rule. For example, standard-setting panels for NAEP increased in 1993 from approximately 20 to 30 panelists per grade level for operational tests, although that panel was then split into two panels of 15 (Loomis, 2012).

Panelist training. Standard-setting can be a cognitively challenging process that is often completely unfamiliar to the panelists we rely upon. Unfortunately, there is relatively little research on how well panelists understand what is being asked of them (e.g., Plake, 2008; Skorupski, 2012). The research that does exist on panelists' understanding of the standard-setting process tends to confirm that panelists do not fully understand or are uncomfortable with their task. Hein and Skaggs (2009) found that a group of Bookmark panelists often deviated from the procedure for placing their bookmarks and tended to rely on external factors rather than their individual judgment. In addition, many of the panelists in that study expressed discomfort with the task of defining a single cut-off point, especially when they perceived items in the OIB to be out of order in terms of difficulty. In another qualitative study, McGinty (2005) found that Angoff panelists were also influenced by factors external to the test, and "felt a tension between the desire to set high standards and the desire to be viewed by the public as doing a good job [of teaching]" (p. 280). The goal of training activities is to ensure that panelists feel comfortable with their task and that their judgments are not unduly influenced by extraneous factors.

Hambleton and Skorupski (2005) found that panelists' understanding of their task increased as the study proceeded, but that most panelists seemed to be confused in the early rounds. In a review of literature on panelists' understanding of standard-setting, Skorupski (2012) concludes that more time should be devoted to training panelists to ensure that they are adhering to the method being used.

Training for standard-setting panelists typically consists of activities that take place before the actual standard-setting process itself and that are intended to orient the panelists to their task (Reid, 1991). The amount and type of training required can depend on the standard-setting method being used as well as the make-up and qualifications of the panelists themselves (Raymond & Reid, 2001). Although panelist training is intended in part to increase the precision of cutscores by increasing panelist consistency (e.g., Clauser, Swanson, & Harik, 2002), the goal is “not to make panelists reach consensus and not to bias their judgments” (Loomis, 2012, p. 109). Rather, the goal of training is to ensure that panelists understand the standard-setting method and are making well-informed judgments.

Panelists surveyed before a standard-setting study using the item-mapping method showed little agreement about what was expected of them or even what the performance standards should be (Skorupski & Hambleton, 2005). Although those results come from a single sample, it is clear that adequate training is crucial if we wish to assume that panelists are attending to the same factors to make their judgments or are operating from the same definition of the performance standard. Even after training, panelists may differ in the way they allow external considerations (e.g., impact data or the performance of their own students) to affect their ratings (Ferdous & Plake, 2005).

Training activities are typically conducted on-site immediately before the standard-setting study itself, but can also include materials and information provided to panelists in advance of the meeting. Raymond and Reid (2001) describe examples of panelist training used for an educational test and a professional licensure test, and identify the following broad categories of training activities: (a) contextualizing the standard-setting; (b) defining a reference group on which to base judgments; (c) practice activities. Those authors also note that providing various types of feedback (e.g., item performance or impact data) may be considered part of panelist training as well.

Training activities that help provide context to panelists may include describing the background, purpose, and consequences of the test itself (Loomis, 2012; Raymond & Reid, 2001). They might also include asking panelists to take the test themselves to familiarize themselves with the content and format of the assessment as it will be encountered by examinees.

Defining the reference group usually involves the thorough review of the performance level descriptors (PLDs) that describe the knowledge and skills of typical students in each performance category (e.g., Loomis, 2012). Berk (1996) proposed asking panelists to build on existing PLDs to construct detailed operational definitions of the PLDs that will help them to interpret their final cutscores. (Although it is possible that PLDs could be developed in conjunction with a standard-setting meeting, Egan, Schneider, and Ferrara (2012) point out that federal peer review guidance under NCLB required PLDs to be developed before standard-setting.) Panelists' perception of the reference group is affected by the descriptions that are developed as part of training, with more detailed descriptions leading to increased agreement in the panelists' own descriptions of borderline examinees (Giraud, Impara, & Plake, 2005). A panelist in one study, for example, reported she was able to internalize and make judgments using the group definition of borderline proficiency, despite the fact that she disagreed with the performance standard (Giraud & Impara, 2005). Ensuring that panelists are all using the same definition of "borderline proficiency" will help to reduce error in their recommendations and lead to stronger interpretations of the cutscore.

Reid (1991) proposed three criteria for evaluating whether training has been effective: 1) ratings produced by individual panelists should be stable over time, 2) ratings should be consistent with relative item difficulties, and 3) ratings should reflect realistic expectations about the performance of examinees.

Panelist accuracy. Van der Linden (1982) used the term "intrajudge inconsistency" to describe "when the judge specifies probabilities of success on the items which are incompatible with each other and, consequently, imply different standards" (p. 296). To avoid confusion with consistency across panelists, Skorupski and Fitzpatrick (2014) refer to intrajudge consistency as "judge accuracy." Unless otherwise stated, this paper will follow their lead and use the term "panelist accuracy" to describe how closely panelists' estimates of examinee performance align with actual examinee performance.

The ability of standard-setting panelists to predict item performance for groups of examinees has been shown to improve by providing training, allowing the panelists to discuss their ratings, and providing normative data (e.g., item *p*-values). Teachers who were not provided training or allowed to discuss their ratings were more accurate at predicting overall item difficulty than at predicting item difficulty for a group of "borderline" students, and they systematically underestimated the performance of borderline students (Impara & Plake, 1998). Van de Watering and van der Rijt (2006) also found that teachers were mostly poor judges of item performance for borderline examinees, although the teachers in that study tended to overestimate student performance. As in the Impara and Plake (1998) study, the teachers in the van de Watering and van der Rijt (2006) study were not provided training. However, panelists who were provided with focused training and allowed to discuss their ratings were quite accurate in estimating the performance of borderline examinees (Plake & Impara, 2001). Accordingly, the covariance between empirical performance and estimated performance has been shown to increase substantially when panelists are allowed to review performance data and discuss their ratings between rounds (Busch & Jaeger, 1990). Showing panelists the proportion of examinees (or a group of examinees) that responded correctly to an item may decrease the cognitive demand of estimating probability for a hypothetical examinee.

Panelist consistency. Several studies have evaluated different modifications to the Angoff method that are intended to increase panelist consistency. Among the modifications studied are accounting for the effects of panelist expertise level (Norcini, Shea, & Kanya, 1988), training the standard-setting panelists (e.g., Clauser, Swanson, & Harik, 2002), and using a common definition of minimum competence (Fehrmann, Woehr, & Arthur, 1991).

The consistency of panelists in a standard-setting study is of primary importance to the current study because the traditional measure of error variance of a cutscore is calculated as the standard error of the mean cutscore recommendation. That is, more consistent the recommendations among the panel lead to smaller estimates of error variance of the cutscore.

Many of the procedures implemented in standard-setting studies (e.g., multiple rounds of ratings, providing feedback to panelists, etc.) are therefore intended to increase the consistency of judgments across a panel. However, even a highly reliable cutscore can be inappropriate for several different reasons (Kane, 1994). For example, if one influential panelist dominates the discussion, panelists' final ratings may be quite consistent with one another while still being inappropriate given the performance standard.

Modeling Standard-Setting Data

Standard-setting studies may be evaluated on a number of criteria to support the validity of how we interpret the resulting standards. Kane (1994) groups the evidence required for such support into three categories: a) procedural, b) internal, and c) external. Internal validity evidence is the type most relevant to the proposed study, although evidence in all three categories is critical to document. According to Kane,

> The internal checks on validity focus on the consistency of the results of the standard-setting study, in particular the consistency of the judges in translating the performance standard into a passing score. Therefore, they provide an empirical check mainly on the descriptive assumption, which posits a correspondence between the performance standard and the passing score. (p. 448)

Calculating the standard error of cutscores. The focus of internal validity checks described by Kane is the precision of the cutscore, which can be quantified with an estimate of its standard error (e.g., Kane, 1994). A small standard error—although no guarantee that a cutscore is appropriate—at least suggests that we would be more likely to get a similar result if we were to repeat the study. A relatively large standard error, on the other hand, will almost always undermine our confidence in the cutscore.

As mentioned previously, the result of a standard-setting study is usually a recommendation to the policymaking body that ultimately decides the cutscore. An estimate of the standard error of the recommended cutscore can help establish a confidence interval within which the policymaker may adjust the recommended cutscore (e.g., Cizek & Bunch, 2007). The *Standards* (AERA, APA, & NCME, 2014) also specify that the variability of the cutscore itself should be reported:

> **Standard 5.21**
>
> When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.
>
> Comment: … Where applicable, variability over participants should be reported. Whenever feasible, an estimate should be provided of the amount of variation in cut

> scores that might be expected if the standard-setting procedure were replicated with a comparable standard-setting panel. (pp. 107–108)

Kane mentions two methods by which the standard error of a cutscore may be estimated: 1) conducting multiple standard-setting studies, and 2) conducting a generalizability study to estimate variance components for different facets of a single standard-setting study. A hybridization of those two approaches in which two groups of panelists receive orientation and training together but complete the rest of the process separately (Hambleton & Pitoniak, 2006) is also possible and appears to be the procedure used for NAEP assessments (Loomis, 2012). Conducting two completely separate studies is likely to be cost- and resource-prohibitive for most testing programs, and so this study will consider only efforts to estimate the standard error resulting from a single standard-setting study.

Hambleton and Pitoniak (2006) classify most internal evidence of validity for standard-setting as falling into one of three categories: 1) consistency within method, 2) intrapanelist consistency, and 3) interpanelist consistency. The first of these is a more holistic estimate focused on the results of a standard-setting study (or what those authors termed “across-panel consistency” (p. 458)). Intra- and interpanelist consistency, on the other hand, refer to the reliability of the panelists themselves, individually (intrapanelist consistency) and as a panel (interpanelist consistency).

Several different approaches to calculating the standard error of cutscores have been developed, focused mainly on the notion of intrapanelist consistency, interpanelist consistency, or both. The traditional approach to calculate the standard error of judgment is to use the standard error of the mean for the individual panelists’ standards (e.g., Jaeger, 1991). A generalizability theory (G-Theory; Cronbach, Glaser, Nanda, & Rajaratnam, 1972) approach developed by Brennan and Lockwood (1980) provides three different estimates of the standard error, depending on whether the results are intended to generalize over samples of panelists, items, or both. It should be noted that when generalizing over samples of panelists for a fixed set of items, the G-Theory standard error is equivalent to the traditional measure. Kane and Wilson (1984) extended the G-Theory approach to account for non-zero correlation between measurement error and standard-setting error. A G-Theory approach has also been applied to both item-rating (e.g., Angoff) and item-mapping (e.g., Bookmark) standard-setting methods to include generalization over rounds of the study as well as designs with panelists nested within table groups or items nested within item “pools” (Yin & Sconing, 2008).

Models for standard-setting. A number of researchers have proposed frameworks or models through which to evaluate the data provided by standard-setting panels. Such models can help researchers and practitioners to assess the defensibility of cutscores and may help them to know where to focus efforts to improve the standard-setting process. For example, van der Linden (1982) incorporated a Rasch model analysis to evaluate intrapanelist consistency. Such an approach could help to identify panelists whose judgments do not seem to be consistent with their overall recommended passing score. Kane (1987) presented a statistic for evaluating the fit of an IRT model to standard-setting data and introduced three methods for calculating cutscores using IRT, assuming that the data fit the model. Kane cautioned, however, that there is no inherent reason to assume standard-setting data fit an IRT (or any other) model, and so this assumption must be tested empirically before using such models to calculate a cutscore. As mentioned in the previous section, several authors have applied G-Theory to standard-setting data in order to account for and quantify the possible sources of error that affect the recommended cutscore (e.g., Brennan, 1995; Brennan & Lockwood, 1980; Kane & Wilson, 1984).

In addition to modeling overall error, several studies have specifically addressed the variability of standard-setting panelists themselves. Both Jaeger (1988) and Engelhard and Anderson (1998) proposed methods for evaluating standard-setting judgments that can identify panelists with aberrant patterns of judgment. Engelhard has applied a many-faceted Rasch (MFR) model to evaluate both Angoff data (2009) and Bookmark data (2011). The MFR approach allows the evaluation of judgments on a number of factors, including panelist severity and presence of a halo effect. Although these methods can provide feedback about the variability of the panelists themselves—which can be valuable information to the body responsible for setting the final cutscores, they do not directly address the appropriateness of the cutscore itself.

Reckase (2006) proposes a theory of standard-setting and describes what he calls a standard-setting panelist's "intended cut score" (ICS). The ICS is the cutscore that a panelist would recommend if he were perfectly consistent in the way he translated the policy definition of a performance level to a score on the reporting scale. The estimation of the ICS is, in Reckase's view, the goal of standard-setting. His proposed framework is intended to allow evaluation of the effectiveness of a standard-setting method at recovering a panelist's ICS. Although standard-setting is typically a group process, Reckase focuses on the recovery of the ICS for an individual panelist. The ICS is assumed to be the driving factor of the standard setting process, regardless of the specific procedure employed. The closeness of the ICS to the actual policy definition, despite being an essential question of validity, is not considered. Reckase proposed three criteria for evaluating a standard-setting methodology: 1) whether the ICS is recovered if a panelist is perfectly consistent in making the required judgments, 2) whether the estimate of the ICS is unbiased over many theoretical replications of the standard-setting process, and 3) whether the standard deviation of this theoretical distribution of estimated ICSs is “as small as possible” (p. 6), with the acceptable magnitude of such variability depending on how the cut score is used. Of the many potential sources of error that might influence cutscore recommendations, Reckase considers only errors in judgments when panelists are working independently.

One limitation of Reckase’s framework is that it assumes standard-setting panelists understand their task (Schulz, 2006). The question of whether panelists understand what is being asked of them is a primary validity concern in standard-setting work. Schulz argues that the relative ease of understanding by panelists using one method may be an important advantage over competing methods, but Reckase’s framework does not permit such distinctions. Schulz also points out that a small standard deviation of estimated ICSs might suggest, for example, that panelists are not using their individual judgment, or that discussions are being dominated by only a few panelists. Evaluating standard-setting methods by this third criterion, in Schulz’s opinion, will not adequately take into account how a method might rely on or encourage individual variability. On a related note, Schulz points out that although Reckase’s framework provides a model for an individual panelist’s ICS, the actual cut score recommendation resulting from a standard setting procedure is a group mean or median. Schulz believes that a model of standard setting must take into account the (intentional) dependence among ratings (e.g., through subsequent rounds of rating) for an accurate estimation of the standard error of the final recommendation.

More recently, a psychometrics of standard-setting, termed “Cutscore Distribution Theory” (CDT; Skorupski & Fitzpatrick, 2014; Skorupski, Zhao, Fitzpatrick, & Chen, 2015), has been proposed for the Angoff method. CDT considers two separate sources of error for Angoff panelists: accuracy and consistency. Accuracy is defined as the correlation between a panelist's item ratings and the true conditional *p*-values of the items for a borderline examinee. Consistency is the correlation among panelists' item ratings. The traditional measure of standard error in an Angoff study considers only consistency, but Skorupski and his colleagues point out that panelists may be consistent in their ratings but still not very accurate in terms of estimating the true conditional probabilities of success. Simulated results show that the size of standard error can be over- or underestimated as panelists are more or less accurate in estimating these conditional probabilities. Using real data, Skorupski et al. (2015) also demonstrated how accuracy may be estimated in an operational setting using empirical item performance and the Rasch model to estimate the vector of true conditional probabilities.

Bayesian Estimation

Because the methods in this study employ Bayesian methods, a brief review of their history is given below, along with a discussion of the differences between Bayesian and more traditional "frequentist" approaches.

Bayesian data analysis was developed in large part on the basis of writings by the Reverend Thomas Bayes and Pierre Simon Laplace, in the 18th and early 19th century (Gelman, Carlin, Stern, & Rubin, 2004). Early work on probability treated statistical parameters as given and focused on determining the probability of observing specific data, given those parameters. In independent work, both Bayes and Laplace are credited with "inverting" this probability statement, or focusing on the probability distribution of the parameters, given the observed data (Gelman et al., 2004).

Bayes' Rule. The foundation of Bayesian statistical inference is Bayes' Rule or Bayes Theorem, presented in Gelman et al. (2004) as:

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta) p(y \mid \theta)}{p(y)} \ ,$$

where

- $p(\theta \mid y)$ is the posterior density, or the probability of the parameters *θ* given observed data *y*;
- $p(y \mid \theta)$ is the likelihood function (i.e., the sampling distribution), or the probability of *y*, given *θ*;
- $p(\theta, y)$ is the joint probability distribution of *θ* and *y*;
- $p(\theta)$ is the prior distribution of the parameters;
- $p(y)$ is the marginal distribution of the data, or the distribution of *y* over all possible values of *θ*.

Having the marginal distribution in the denominator makes the posterior a proper density (i.e., its sum is 1). However, as Gelman et al. (2004) explain, the marginal distribution may be considered a constant and, therefore, omitted, yielding an alternative form of Bayes' rule:

$$p(\theta \mid y) \propto p(\theta) p(y \mid \theta) ,$$

which states that the posterior is proportional to the product of the prior and the likelihood.

Bayesian vs. Frequentist approaches. Bayesian statistical analysis and more typical, so-called "frequentist" methods are fundamentally different in their approach. Sir Ronald Fisher, who developed the analysis of variance (ANOVA), wrote in his 1925 book *Statistical Methods for Research Workers* that "the theory of inverse probability is founded upon an error, and must be wholly rejected" (Fisher, 1925, p. 10). Under the frequentist paradigm, a researcher's question can be framed as "How likely are my data, given certain parameters?" For example, an ANOVA from the frequentist perspective might be employed to determine how likely an observed group difference is, given that their true difference is 0. If the F-test result suggests the observed difference is sufficiently unlikely (say, a 5% chance), then the researcher rejects the null hypothesis and concludes that the groups are, indeed, different. Among the underlying assumptions for this particular technique is that the residual variance is normally distributed with an expected value of 0. Although different statistical methods are more or less sensitive to violations or such assumptions (see, for example, Glass, Peckham, and Sanders, 1972, for a review and discussion of the practical consequences), the point here is that they are implicit, typically accepted as true when deciding to use a particular analytic method.

In contrast, a Bayesian researcher approaching the same problem might ask, "How likely are true group differences to have a certain value, given the data?" To answer this question, the Bayesian analyst specifies both a prior distribution and a likelihood function. The prior represents the researcher's belief about how the parameters are distributed in the population. In the absence of data, the posterior distribution is defined by the prior; as observations accumulate, the influence of the prior is reduced. As Box and Tiao (1973/1992) explain:

> A prior distribution, which is supposed to represent what is known about unknown parameters before the data is available, plays an important role in Bayesian analysis. Such a distribution can be used to represent prior knowledge or relative ignorance. In problems of scientific inference we would usually, were it possible, like the data "to speak for themselves." Consequently, it is usually appropriate to conduct the analysis as if a state of ignorance existed *a priori*. (p. 2)

The likelihood function "is *the* function through which the data **y** modifies prior knowledge of **θ**; it can therefore be regarded as representing the information about **θ** coming from the data," (Box & Tiao, 1973/1992, p. 11, emphasis in original).

Chapter III: Methods

This study extends Cutscore Distribution Theory (Skorupski & Fitzpatrick, 2014) to the Bookmark standard-setting method. The error variance of cutscores set using the modified Angoff method was shown to be a function of panelist consistency, panelist accuracy, and sample size. The goal of this study is to apply that concept to the Bookmark method. Because an advantage of the Bookmark method is to order items by difficulty (thereby increasing the accuracy of panelists), this study focuses on consistency and sample size, with consistency partitioned into both within-group and between-group components. Study 1 will use simulated data to determine the relationship between the standard error of Bookmark cutscores and panelist consistency and sample size. Study 2 will apply the same methodology to real standard-setting data to compare the CDT error variance to that calculated using traditional methods.

Study 1

Data generation. All data for Study 1 were generated and analyzed in *R* (R Core Team, 2014).

A 100-item test was simulated by generating 100 true *a*-, *b*-, and *c*-parameters under the 3PL IRT model (Birnbaum, 1968). Generating distributions were Uniform (lower bound=0.5, upper bound=2.0) for the *a*-parameters and Beta (α=5, β=20) for the *c*-parameters. To ensure an even distribution of item difficulty values (as in an OIB consisting of an augmented test form), the *b*-parameters were an evenly spaced sequence of 100 numbers from -3.0 to 3.0. True item parameters were fixed across all conditions and replications.

A Bookmark standard-setting study was simulated to establish three cutscores on the test. The true cutscores on the latent scale were assumed to be known, and were fixed at -0.5, 0.5, and

1.5 across all conditions. For each condition, a total sample of N_t panelists were simulated by generating N_i intended cutscores (ICSs) in each of N_g groups for each of the three cuts. ICSs are simulated to account for two sources of variability: group-level variation, which will be referred to as bias, and individual variation within a group, referred to as inconsistency, in the following manner:

$$Bias_{g,k,r} \sim N(0, \sigma_b)\ ,$$

$$ICS_{i,g,k,r} \sim \begin{cases} N([\theta_k^* + Bias_{k,g,r}], \sigma_c), & k = 1 \\ N_{trunc}([\theta_k^* + Bias_{g,k,r}], \sigma_c, a = [\theta_{k-1}^* + Bias_{g,k-1,r}], b = \infty), & k = 2, 3 \end{cases}$$

where

- σ_b is the SD of true bias,
- σ_c is the SD of true consistency,
- θ_k^* is the true latent ability value associated with cutscore *k*,
- $ICS_{i,g,k,r}$ is the intended cutscore for panelist *i* in group *g* at cutscore *k* within replication *r*, and
- *a* and *b* are the lower and upper bounds, respectively, of the truncated normal distribution.

In other words, for each replication, the ICSs were drawn randomly from a normal distribution centered at the true cut value plus the group-level bias, with a variance of σ_c^2. To ensure that ICS values would be generated such that $ICS_{i,g,k,r}$ was always less than $ICS_{i,g,(k+1),r}$, the two higher cuts were drawn from truncated normal distributions, with a lower bound at the preceding cutscore. Additionally, $Bias_{g,k,r}$ was regenerated if $(\theta_k^* + Bias_{g,k,r})$ is less than or equal to $(\theta_{k-1}^* + Bias_{g,k-1,r})$.

The process of panelists working through OIBs was simulated by first re-ordering the items. For IRT models other than the Rasch model, the ordering of items by difficulty may vary by examinee ability. As in operational Bookmark studies, the items in this study were ordered by their Bookmark difficulty level (BDL) values, which are the θ values that correspond with a response probability of 2/3 (RP67). The BDL for each item *j* was calculated using the true item parameters and RP = 2/3:

$$BDL_j = b_j - \frac{\log\left(\frac{1-RP}{RP-c_j}\right)}{a_j}$$

After reordering the items by their BDL values, a true conditional probability value *P* was calculated for each panelist in each group, for each item, at each of the panselist's three ICS values. Next, for each $P_{i,g,j,k}$ value, a vector of 30 random uniform draws between 0 and 1 was generated. If at least 20 (i.e., two-thirds) of those values are less than $P_{i,g,j,k}$, then the panelist received a 1 ("yes") for that item for that cutscore; if fewer than 20 of the values are less than $P_{i,g,j,k}$, the panalist received a 0 ("no"). This procedure is intended to simulate the typical instructions to Bookmark panelists, namely to imagine a group of 30 borderline examinees and decide whether at least two-thirds of them would answer the item correctly. The result of this process is a set of three matrices per group (one for each cutscore) of 1s and 0s, with a row for each panelist and a column for each item, in BDL order.

Next, bookmark placement locations were calculated for each panelist. Because Bookmark panelists are typically instructed to continue past the first "no" item, the bookmark locations were determined starting with the first item and looking for the earliest sequence beginning with 0 in which the subsequent four items contained no more than a single 1. Such a procedure placed a bookmark at the first position of any of the following four sequences: 01000, 00100, 00010, or 0000X (where X can equal 0 or 1). For example, consider the following vector for first 10 items for a particular panelist:

1111100100

The bookmark location assigned to this panelist would be 6, because the earliest sequence 00100 begins at item 6. In the case that such a sequence does not occur, the values for the last three items were examined. If the last three items were all 0s, the bookmark location was the third-to-last item. If not, but the last two items were 0s, the bookmark location was the second-to-last item. Likewise, if the last item was 0 but the second-to-last item is not, the bookmark location was at the last item. Finally, if a bookmark location still did not exist (meaning that no such sequence existed and the last item had a value of 1), then the bookmark location was the last item plus one (i.e., 101). Thus, the lowest possible bookmark location was 1 (e.g., if the vector for a panelist was all 0s) and the highest bookmark location was 101 (indicating that a panelist believes the borderline examinees should have mastered all items in the OIB).

Traditional bookmark cutscore calculation. Traditional bookmark cutscores were calculated by converting each bookmark placements to corresponding values on the latent ability scale, $\theta^{*}_{i,g,k}$, for each panelist *i* in group *g* for each cutscore *k*. Unless the mean bookmark location was the first or last item, the latent value was calculated as the average of the BDL values for the bookmark item and the item immediately preceding it in the OIB (i.e., the next easiest item). If the mean bookmark location was 1 or 101, the latent cutscore was simply the BDL for the first or last item, respectively. The mean bookmark location across all N_t panelists was the traditional bookmark cutscore.

The traditional standard error and 95% confidence intervals of the bookmark cutscores were calculated as

$$SE_{trad_k} = \frac{SD(\theta^{*}_{\bullet,\bullet,k})}{\sqrt{N_T}} \text{ and}$$

$$\text{95\% Confidence Interval} = (\bar{\theta}^{*}_{\bullet,\bullet,k} - 1.96 * SE_{trad_k},\ \bar{\theta}^{*}_{\bullet,\bullet,k} + 1.96 * SE_{trad_k})$$

Cutscore Distribution Theory bookmark cutscore estimation. CDT bookmark cutscores were estimated using a fully Bayesian approach to model two main sources of uncertainty: between-group bias and within-group inconsistency. Each panelist's bookmark placement was converted to an implied vector of item responses for an examinee with ability at the cutscore. In other words, each panelist's bookmark placement implies a Guttman vector, with 1s for the first item through the item immediately preceding the bookmark location, and 0s for the bookmark item and all subsequent items. The likelihood for each panelist's implied item response *x* was specified as a Bernoulli trial, and the model for their probability was modified 3PL IRT, as follows:

$$x_{i,g,j,k} \sim Bernoulli(p_{i,g,j,k})$$

$$p_{i,g,j,k} = c_j + \frac{1 - c_j}{\left(1 + e^{-a_j(\theta_{i,g,k} - b_{RP_j})}\right)}$$

where

- $x_{i,g,j,k}$ is the implied item response for panelist *i* in group *g* for item *j* at cutscore *k*,
- $p_{i,g,j,k}$ is the probability of a correct response for each item, with subscripts as above,
- a_j and c_j are the known 3PL *a*- and *c*- parameters for item *j*, and

- b_{RPj} is the bookmark difficulty level (BDL) for item *j*, calculated from known 3PL item parameters and a response probability (RP) value of 2/3.

The model here is a modified 3PL IRT, using item BDLs instead of their *b*-parameters. The item parameters provided to OpenBUGS are the same parameters used to generate the data, and are treated as known. In a real-data application (as in Study 2), parameter estimates from a previous calibration will be treated as known values. An alternative method would be to estimate new item parameters using panelists' implied item response vectors; however, scaling the test in this way would place panelists' ability estimates on a separate scale from the one used for examinees, making direct comparisons difficult. By using a modified IRT model for panelist estimates, the CDT methodology assumes that panelists interact with items in the same way that examinees do, and the assumption that Bookmark data fit an IRT model is untested (for example, see Kane, 1987, for a similar discussion). However, the model used here was chosen in part to ensure that panelists' ability estimates would be on the same scale as examinee ability estimates.

The effect of the modified IRT model is to shift each item's item characteristic curve to the right while maintaining its shape. The primary estimand of interest is $\theta_{i,g,k}$, which is interpreted as the ability level a panelist has in mind when making bookmark placements. The priors for the ability parameters are as follows:

$$\theta_{i,g,k} \sim N(\vec{C}, \mathbf{\Sigma_g}),$$

$$C_k \sim N(\bar{\mu}_{\bullet,k}, 0.001)$$

$$\mu_{g,k} = \bar{\theta}^*_{\bullet,g,k}, \text{ and}$$

$$\mathbf{\Sigma_g} \sim Wishart^{-1}(\rho, K),$$

where

- C_k is the CDT estimate for cutscore *k*,
- ρ is a *K*×*K* identity matrix,
- *K* is the total number of cutscores, and
- $\bar{\theta}^{*}_{\bullet,g,k}$ is the mean ability estimate from group *g* for cutscore *k*, calculated using the traditional method.

In other words, the prior for individual panelists' ability estimates is multivariate normal, with a mean vector equal to the *K* CDT cutscore estimates, and a separate variance/covariance matrix for each group. The prior for the cutscore estimate C_k is intentionally uninformative, with its mean determined by the data and a precision of 0.001.

Model parameters for each of the 100 replications within each study condition were estimated in OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009) using the R2OpenBUGS package (Sturtz, Ligges, & Gelman, 2005). For each replication, the posterior mean of C_k was saved as the CDT cutscore estimate, and the 2.5th and 97.5th percentiles of the posterior were saved as the lower and upper bounds, respectively, of the 95% credible interval.

Condition summaries. For each condition, the traditional and CDT cutscores and error estimates were compared with the true cutscore values of -0.5, 0.5, and 1.5. The "accuracy rate" for each method was calculated as the percentage of replications in which the true cutscore was included in the 95% confidence interval (for the traditional method) or 95% credible interval (for the CDT method).

Study 2

In Study 2, the estimation methodology of Study 1 was applied to data collected as part of an operational standard-setting study using a modified Bookmark procedure. The data are from an operational statewide testing program that set cutscores on English Language Arts (ELA) and Mathematics tests for grades 3 through 8 and high school during the summer of 2015. The standard setting study employed an extended Bookmark procedure, which was similar to the typical implementation of the Bookmark procedure except for the treatment of certain testlet-type groups of questions (e.g., a series of items about a single reading passage). There were also differences in the way items were scaled and OIBs were constructed, which are briefly described in the Results chapter.

Panelists in each grade-by-subject combination were tasked with setting three cutscores in order to define four performance levels. Rather than setting cutscores sequentially (as was simulated in Study 1), the operational study asked panelists to set Cutscore 2 first, followed by Cutscore 1 (the lowest cutscore), and the Cutscore 3 (the highest cutscore). Cutscore 2 was associated with the highest stakes for this testing program, and so starting with Cutscore 2 was intended to ensure that panelist fatigue would not influence their process. The study incorporated three rounds of the Bookmark process, in addition to a variety of training activities meant to orient panelists to the task and help define the “borderline” student at each cutscore. Panelists worked individually in round 1, followed by small group discussion and round 2 bookmark placements, followed by total sample discussion and round 3 (final) bookmark placements. The operational study also included a “vertical articulation” procedure at the conclusion in order to reconcile any potential discrepancies in cutscores across grade levels. For the purpose of this study, any changes after vertical articulation were not included.

The items that formed the OIBs were scaled using the 2-PL IRT model (Birnbaum, 1968). The data available for Study 2 included the item parameters, the item locations in the ordered item booklets, the panelist bookmark placements after each round, and the panelist identifiers from each table (i.e., small group) that could be used to determine which group each panelist belonged to. For the sake of simplicity, only the final (round 3) bookmark placements were used here. Results of Study 2 were compared with both the Study 1 results and the results using typical Bookmark standard errors.

Chapter IV: Results

Results of the simulation study are presented below, followed by results from the study using real data. For each study condition, results for cutscores calculated using the Traditional Mean method are compared with those calculated using the CDT method.

Study 1

Condition summaries. The primary outcomes of interest in Study 1 were the 95% confidence interval (for the Traditional method) and 95% credible interval (for the CDT method) around the cutscore estimates of the respective methods. (For the sake of simplicity, both will be referred to as 95% CIs.) Plots of the estimated cutscores and their 95% CI are presented in the Appendix, in Figures A.1 through A.204. The top figure on each page shows results for the Traditional method for a single condition, and the bottom figure shows results for the CDT method for a single condition. The x-axis in each plot is the ability metric, and the vertical dotted lines indicate the three true cutscores of -0.5, 0.5, and 1.5. The 100 replications within each condition are represented on the y-axis. For each cutscore, the dots represent the cutscore estimate for that replication, and the lines through it indicate the 95% CI around the estimate. The colors indicate whether the true cutscore is within the 95% CI for that replication. For Cutscore 1, blue means the CI includes the true cut, and orange means it does not; for Cutscore 2, black means the CI includes the true cut, and green means it does not; for Cutscore 3, purple means the CI includes the true cut, and red means it does not. Note that in order to keep the x-axes consistent across plots and still distinguishable among replications, a few data points fall outside the plotting area.

Summary statistics across all 100 replications for each condition are also presented in the Appendix, in Tables A.1 through A.18. The results under each condition for Cutscore 1 are given in Tables A.1 through A.6, those for Cutscore 2 in Tables A.7 through A.12, and those for Cutscore 3 in Tables A.13 through A.18. In each table, the first four columns list the number of panelists per group (N_I), the number of groups (N_G), and the data-generating values of the variance of bias and the variance of consistency. The Cutscore Mean columns give the mean and SD of the cutscores calculated using the Traditional Mean method and the CDT method. The next two columns provide the same information about the error variance estimates calculated under each method, and the final two columns provide the "accuracy rate," or the percentage of replications wherein the true cutscore was included within the 95% credible or confidence interval. Due to the high number of conditions and the size of these tables, the same data are also presented graphically in the sections that follow.

Recovery of true cutscores. The following sections summarize the recovery of each true cutscore by each estimation method. In general, the differences between the true and estimated cutscores depended more on the estimation method than the variance of consistency or bias.

Cutscore 1. Results for Cutscore 1 are given in Tables 1 through 6. The true value of Cutscore 1 was -0.5. For all conditions, the Traditional method overestimated the true cutscore by about 0.1, a difference that was fairly consistent as the total number of simulated panelists increased from 8 to 64.

Figure 1 presents plots of the mean cutscore estimates for each method against the variance of consistency with one group of eight panelists. (By definition, there is no bias in this condition because there is only one group.) The left plot shows that the Traditional estimates change very little as the variance of consistency increases. On the right, the CDT estimates get slightly worse as the variance of consistency changes from 0 to 0.3. Figure 2 plots the same data for one group of 16 panelists, with the general trends being very similar to those in Figure 1. With no group-level bias, the Traditional method appears less influenced by increasing levels of within-group consistency than the CDT method.

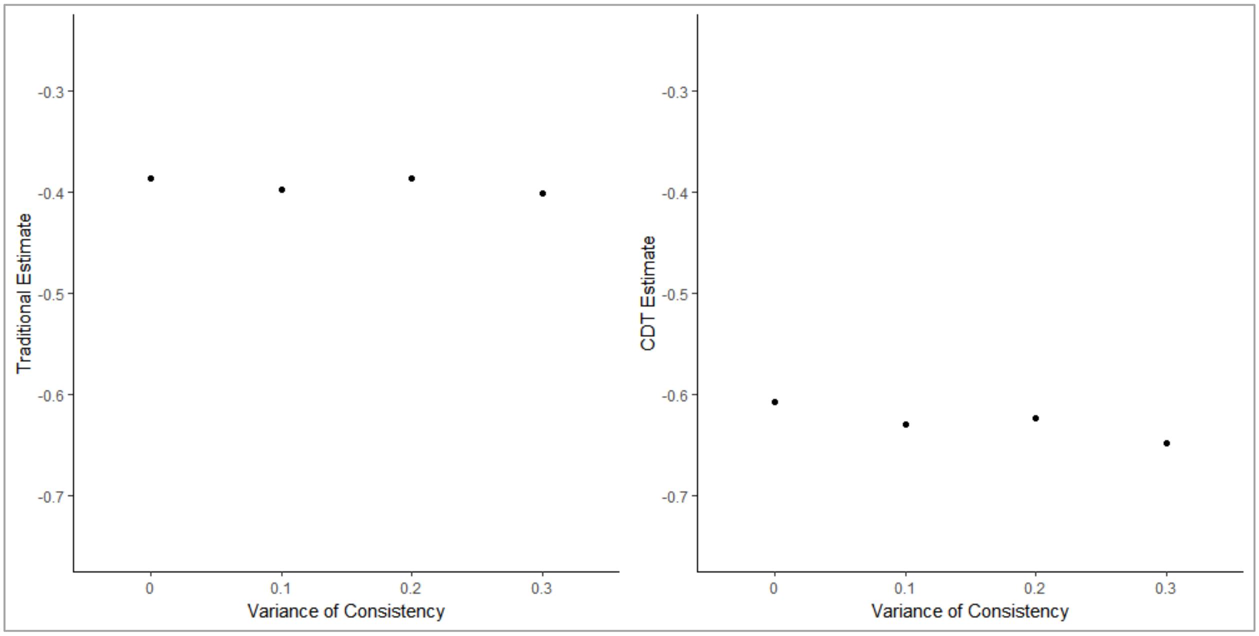

Figure 1. Mean Cutscore 1 estimates for each method with 1 group of 8 panelists.

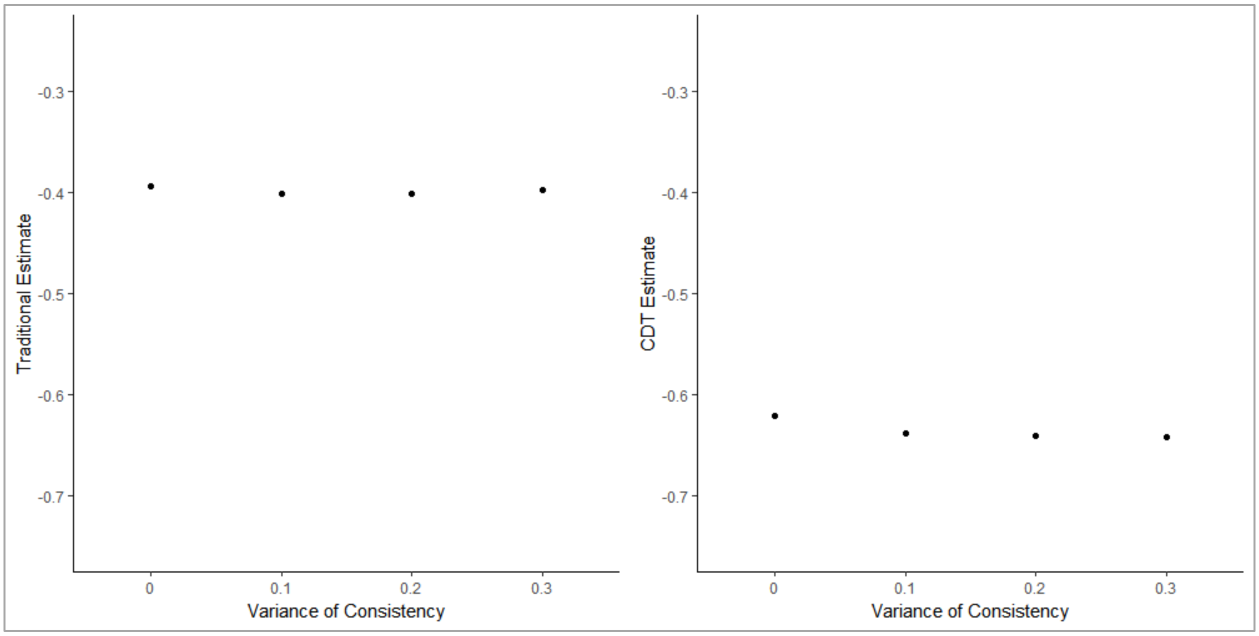

Figure 2. Mean Cutscore 1 estimates for each method with 1 group of 16 panelists.

Figures 3 through 8 present similar plots for each sample-size condition, except that the variance of bias is on the x-axis, and the data are grouped by the level of inconsistency. These plots help show that group-level bias has almost no effect on the mean estimates for the Traditional method, but that the CDT estimates tend to improve as bias, the number of groups, and the number of panelists within each group increase. Inconsistency also seems to have a greater effect on the CDT estimates, as they generally improved as the level of consistency variance increased.

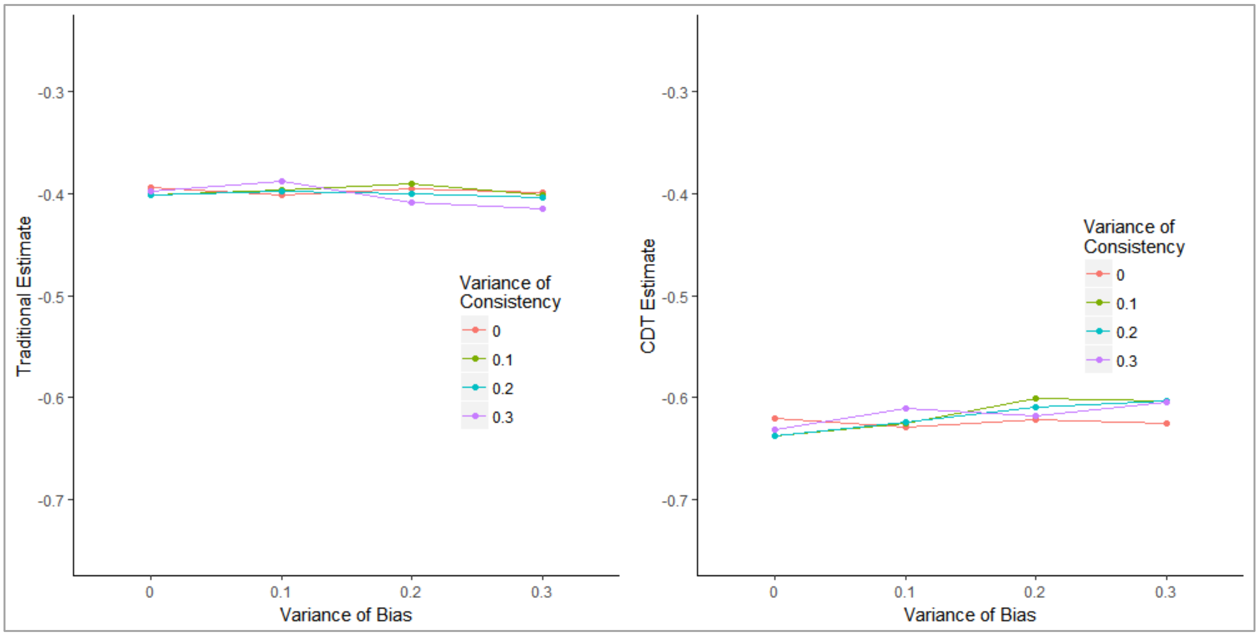

Figure 3. Mean Cutscore 1 estimates for each method with 2 groups of 8 panelists.

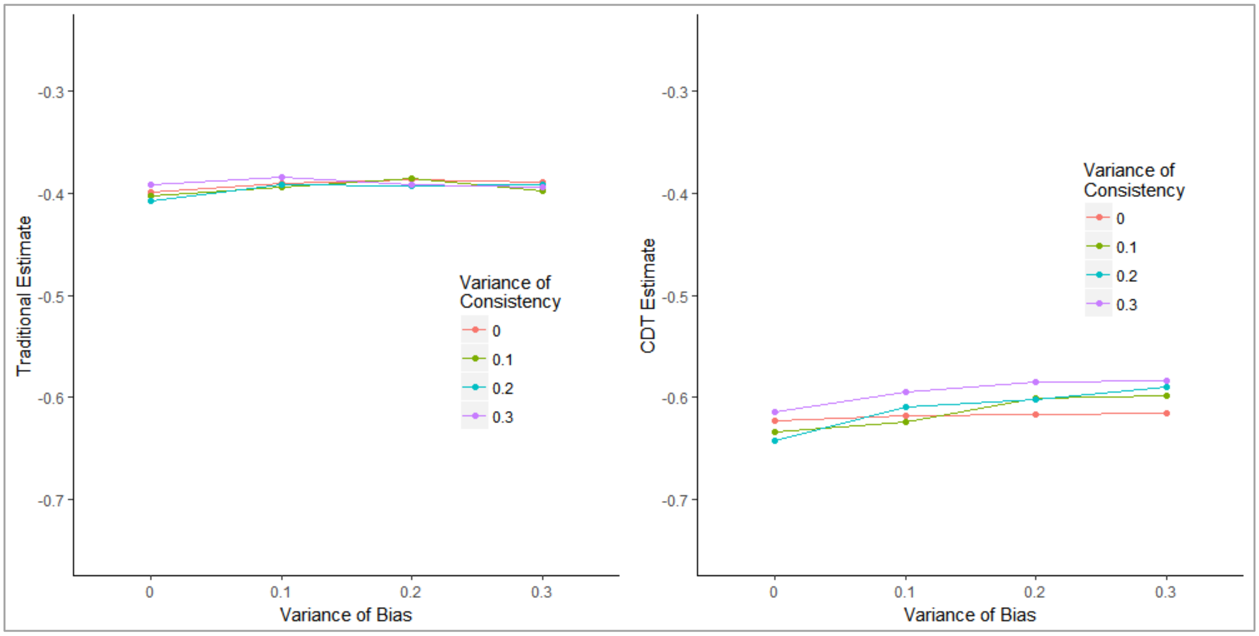

Figure 4. Mean Cutscore 1 estimates for each method with 3 groups of 8 panelists.

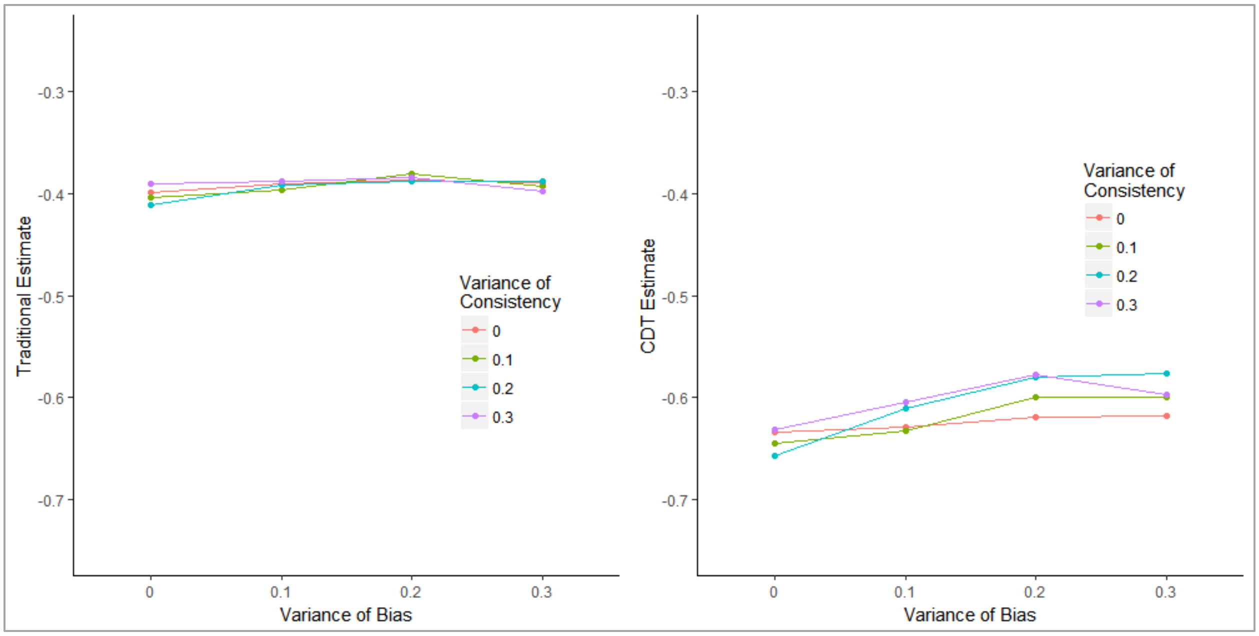

Figure 5. Mean Cutscore 1 estimates for each method with 4 groups of 8 panelists.

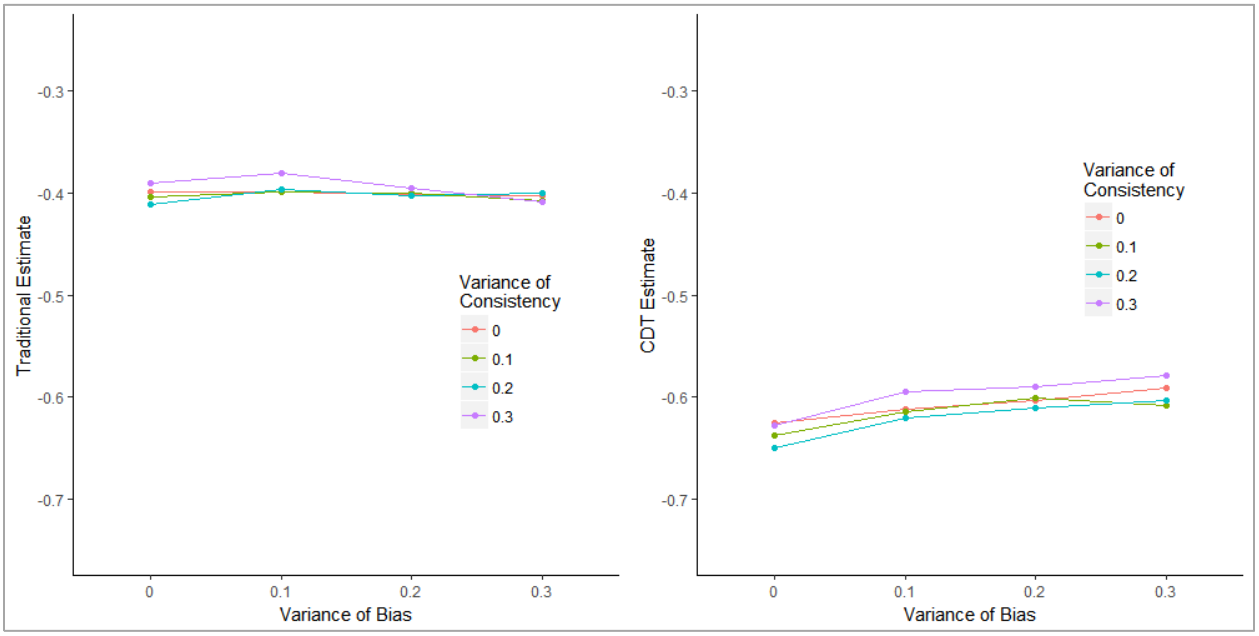

Figure 6. Mean Cutscore 1 estimates for each method with 2 groups of 16 panelists.

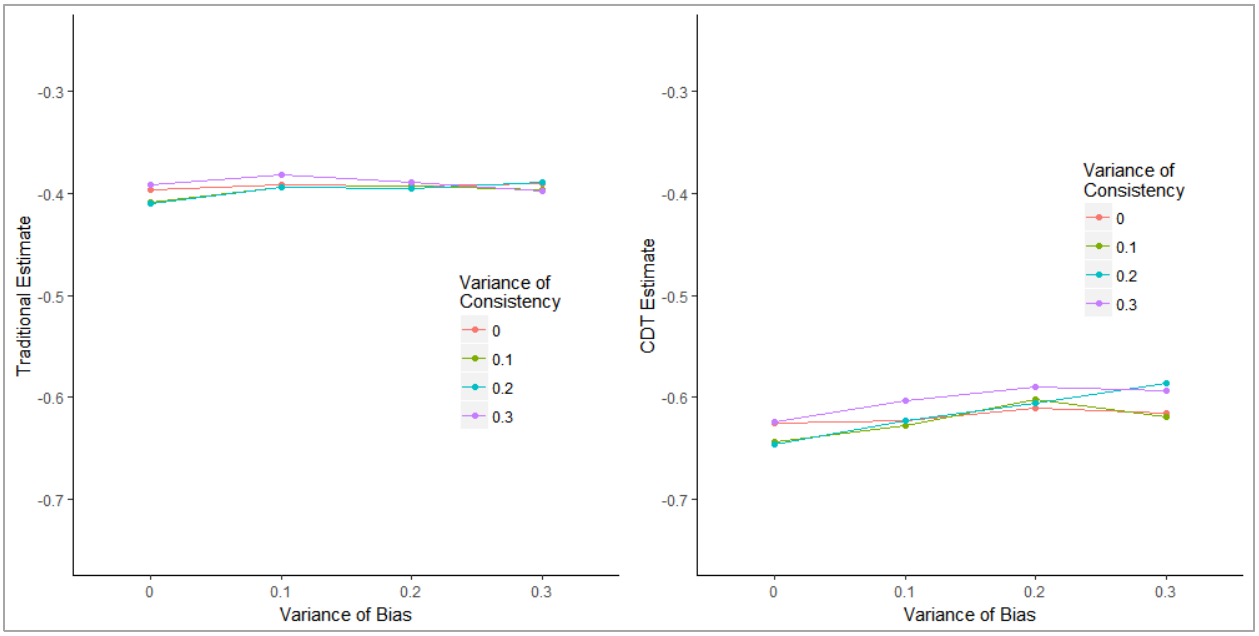

Figure 7. Mean Cutscore 1 estimates for each method with 3 groups of 16 panelists.

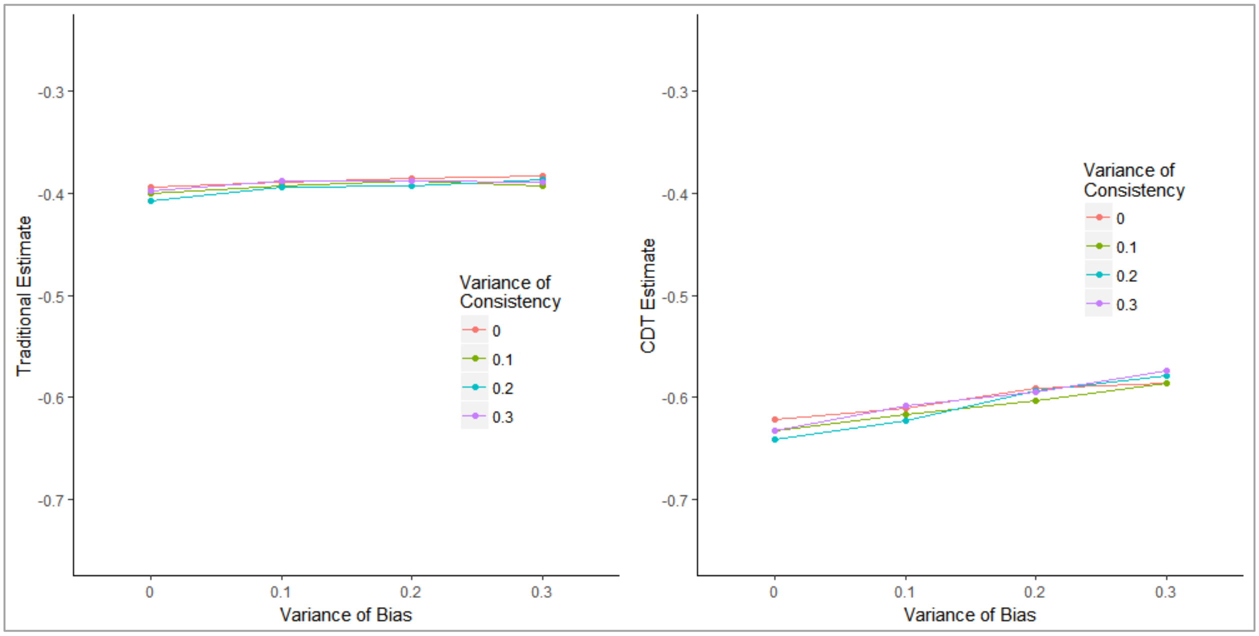

Figure 8. Mean Cutscore 1 estimates for each method with 4 groups of 16 panelists.

Cutscore 2. Results for Cutscore 2 are given in Tables 7 through 12. The true value of Cutscore 2 was 0.5. As with Cutscore 1, the Traditional method always overestimated the true cutscore. However, the CDT estimates for Cutscore 2 appeared to be more influenced by the levels of inconsistency and bias, and tended to be closer to the true cutscore.

Figures 9 and 10 present plots of the mean cutscore estimates for each method against the variance of consistency with one group of eight panelists, and then one group of 16 panelists. (By definition, there is no bias in this condition because there is only one group.) For both methods, the cutscore estimates increased as the variance of consistency increased. However, because the Traditional estimate was already overestimating the cutscore, the increase means that the Traditional estimates are worse as the panelists are less consistent. With one group of panelists, neither method was affected much by increasing the sample size from eight to 16.

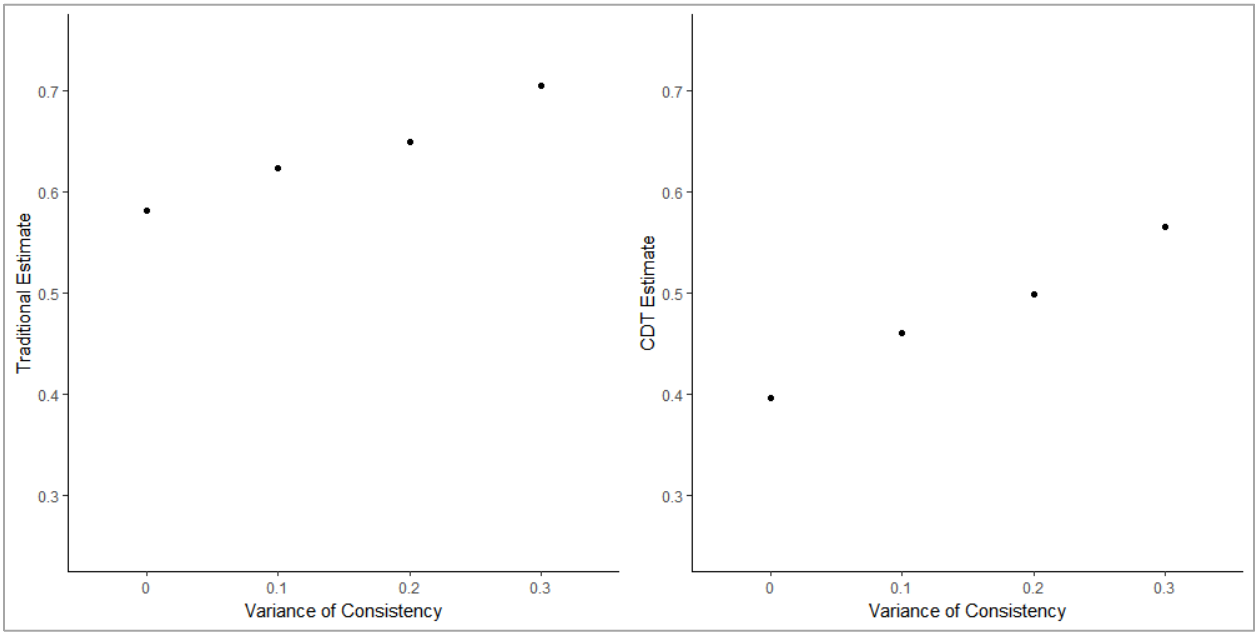

Figure 9. Mean Cutscore 2 estimates for each method with 1 group of 8 panelists.

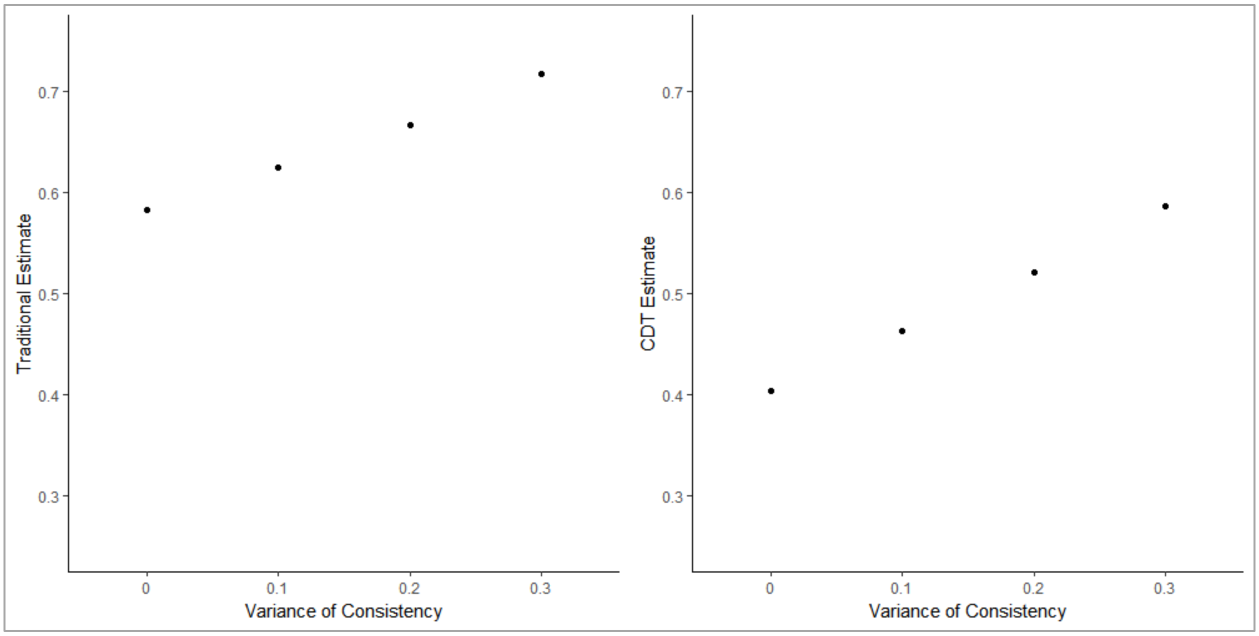

Figure 10. Mean Cutscore 2 estimates for each method with 1 group of 16 panelists.

Figures 11 through 16 present similar plots for each sample-size condition, except that the variance of bias is on the x-axis, and the data are grouped by the level of inconsistency. Unlike the results for Cutscore 1, the results for Cutscore 2 show that the effects of bias and inconsistency on mean cutscore estimates were very similar for each method, with the only real difference being their “starting point.” The lines in the plots are nearly parallel within each method and across method, suggesting no real method effect for on cutscore recovery for Cutscore 2.

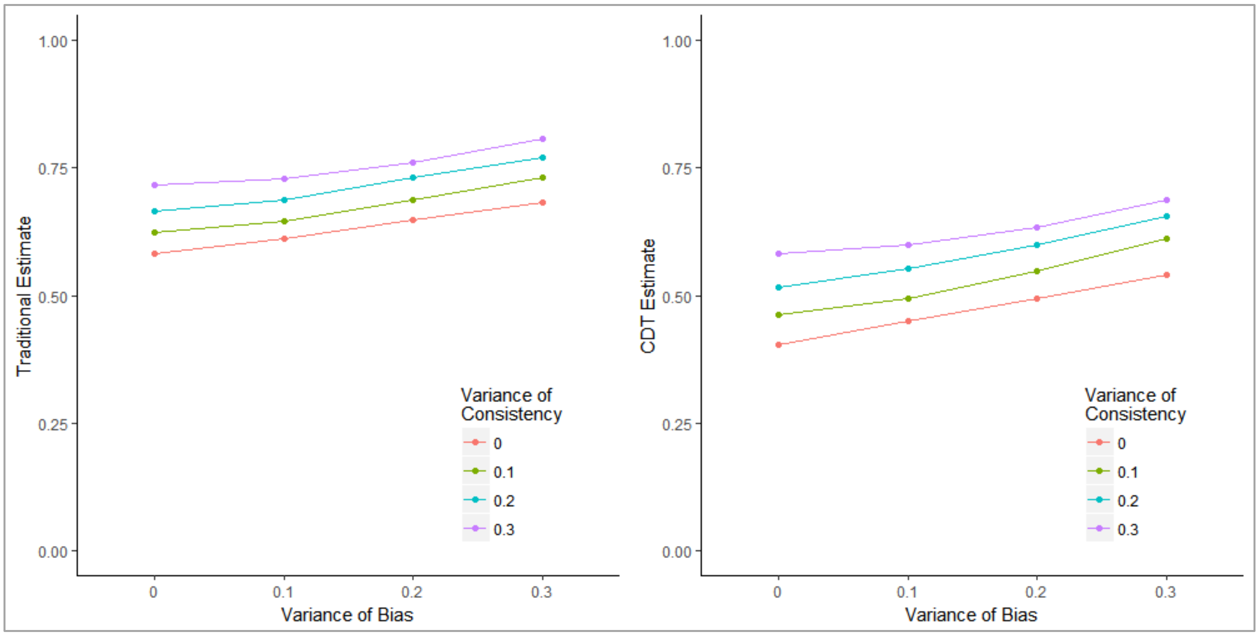

Figure 11. Mean Cutscore 2 recovery for each method with 2 group of 8 panelists.

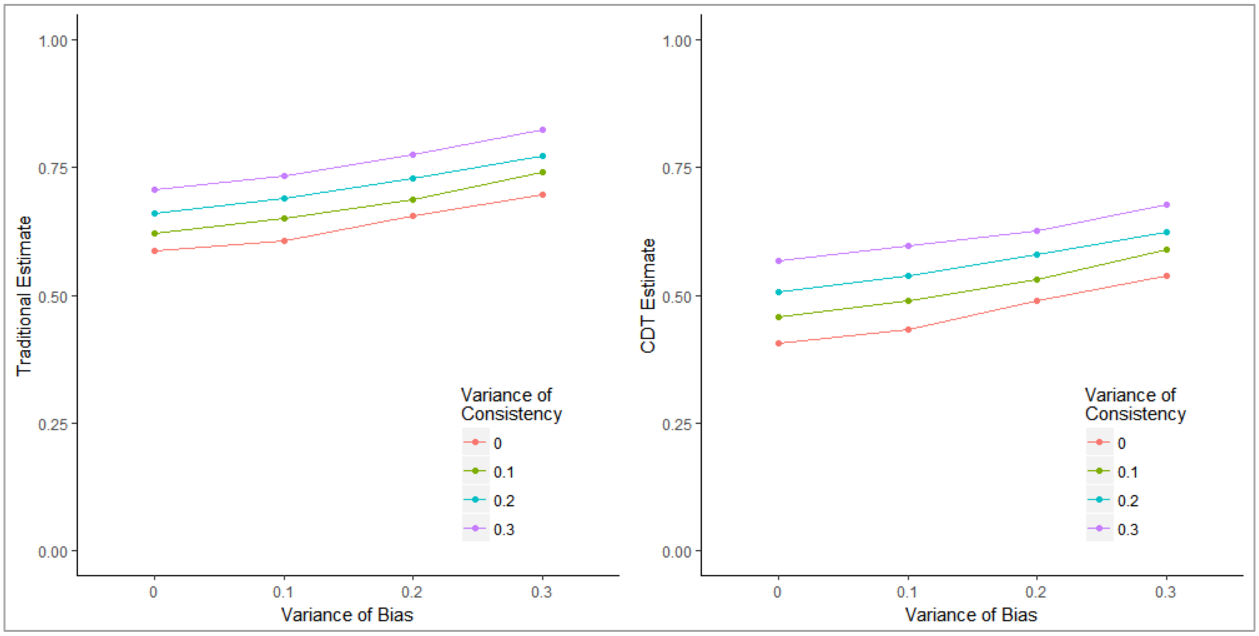

Figure 12. Mean Cutscore 2 recovery for each method with 3 groups of 8 panelists.

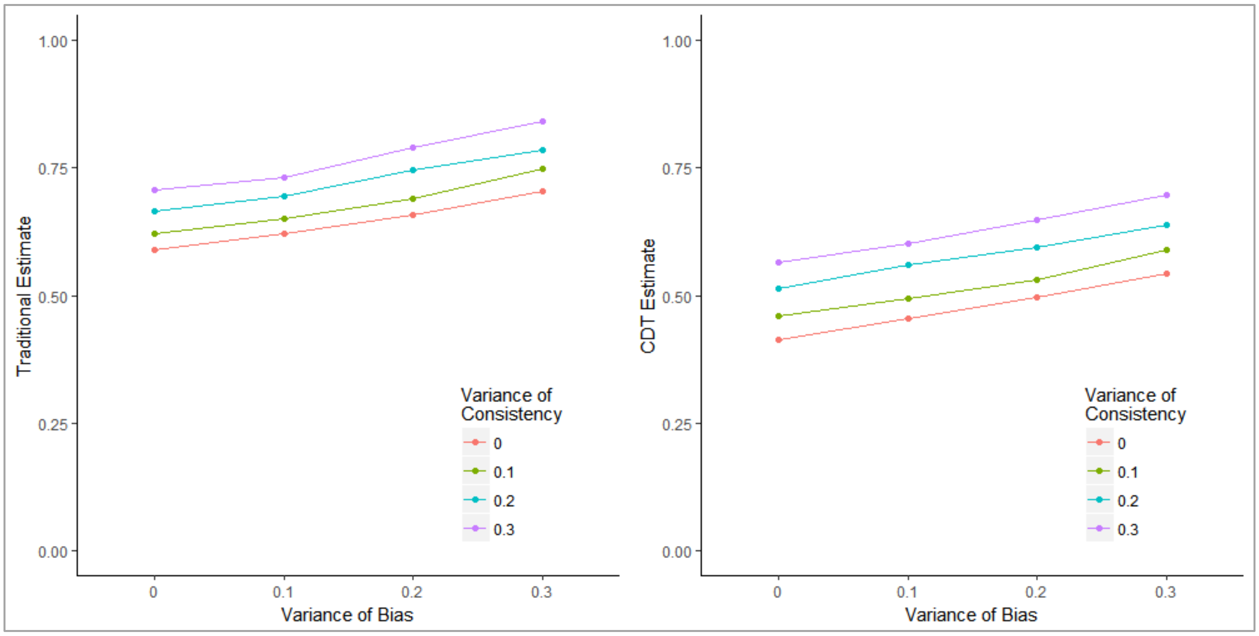

Figure 13. Mean Cutscore 2 recovery for each method with 4 groups of 8 panelists.

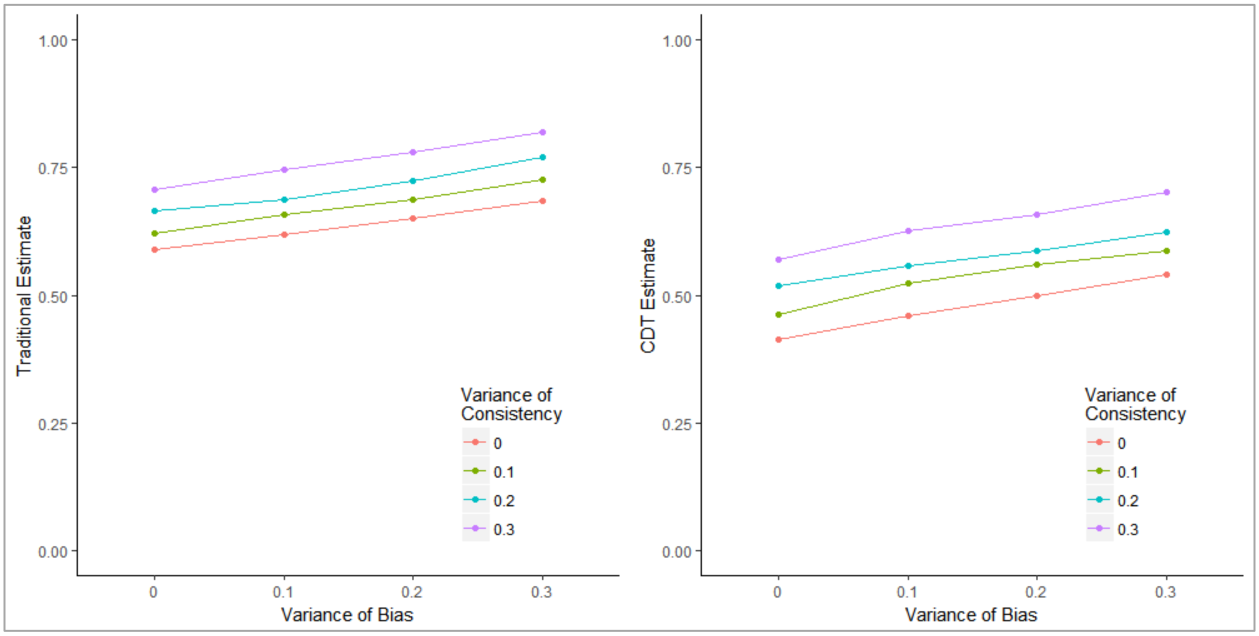

Figure 14. Mean Cutscore 2 recovery for each method with 2 groups of 16 panelists.

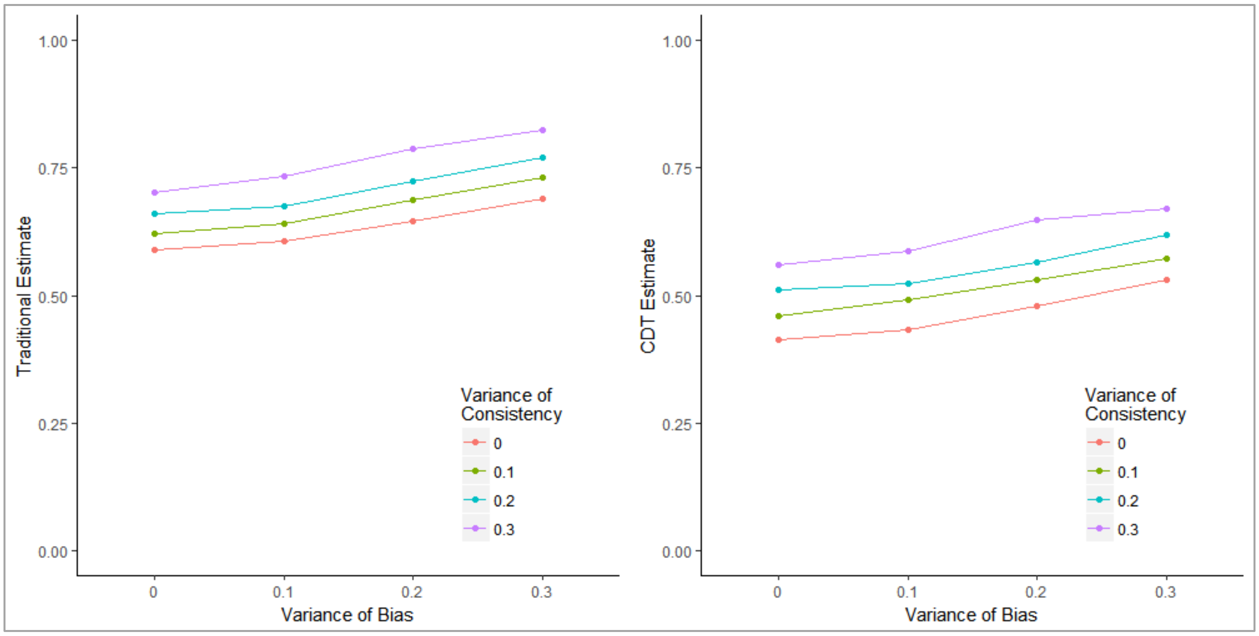

Figure 15. Mean Cutscore 2 recovery for each method with 3 groups of 16 panelists.

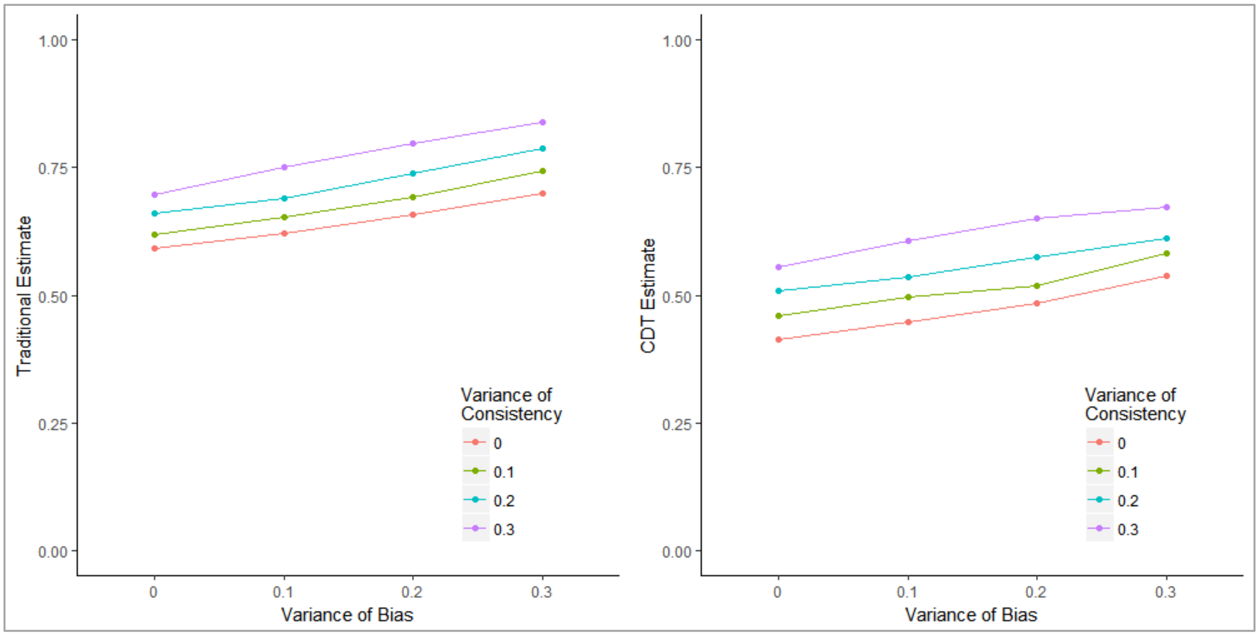

Figure 16. Mean Cutscore 2 recovery for each method with 4 groups of 16 panelists.

Cutscore 3. Results for Cutscore 3 are given in Tables 13 through 18. The true value of Cutscore 3 was 1.5. Unlike the results for Cutscores 1 and 2, the mean estimates of Cutscore 3 for each method were quite similar.

Figures 17 and 18 present plots of the mean cutscore estimates for each method against the variance of consistency with one group of eight panelists, and then one group of 16 panelists. (By definition, there is no bias in this condition because there is only one group.) These two plots show that the estimation methods produced very similar results for Cutscore 3, with the CDT estimates slightly more affected by the highest level of inconsistency than the Traditional method.

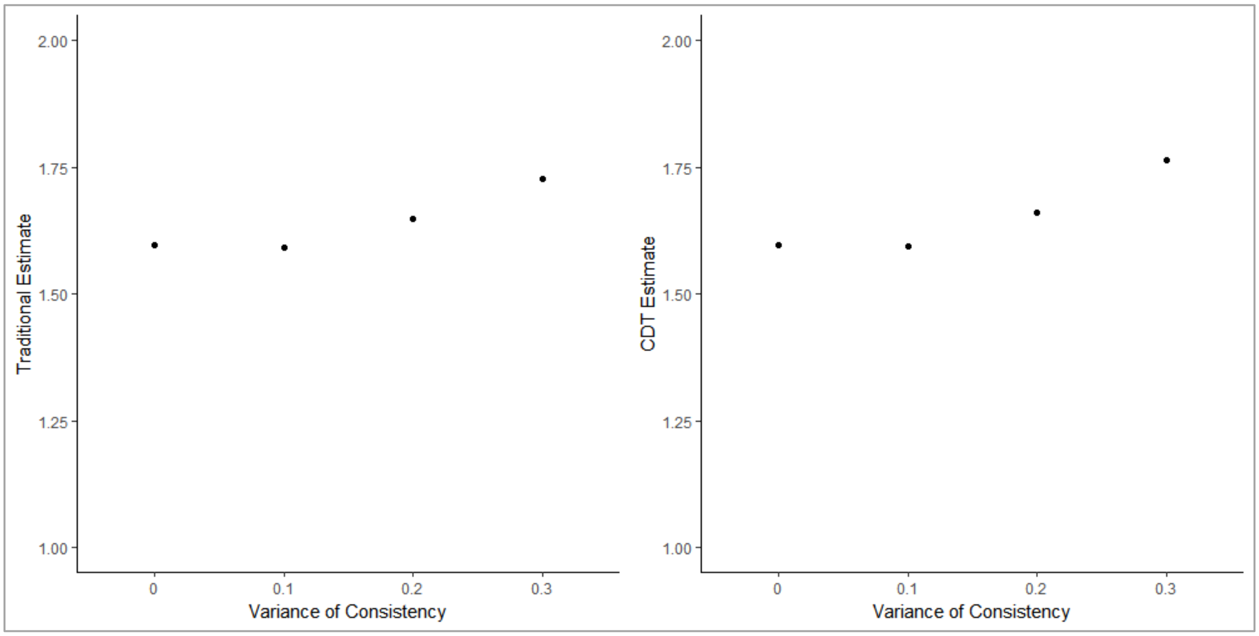

Figure 17. Mean Cutscore 3 recovery for each method with 1 group of 8 panelists.

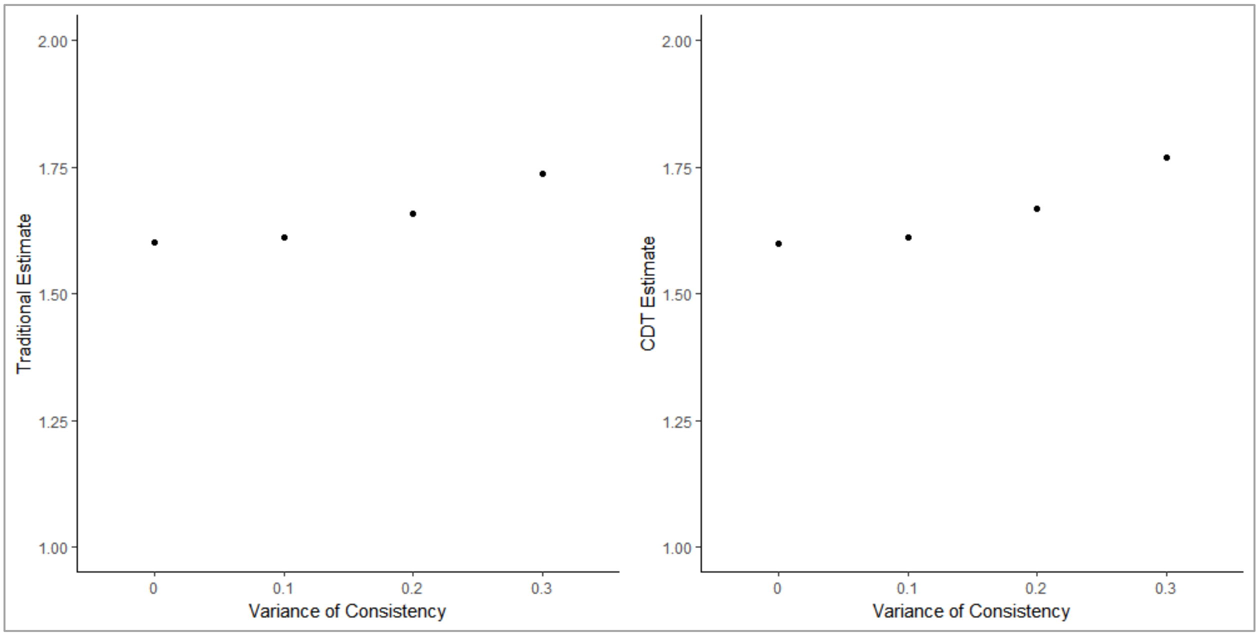

Figure 18. Mean Cutscore 3 recovery for each method with 1 group of 16 panelists.

Figures 18 through 24 present similar plots for each sample-size condition, except that the variance of bias is on the x-axis, and the data are grouped by the level of inconsistency. As with Cutscore 2, the results for Cutscore 3 show that the trends for both methods was very similar. In general, both methods produced better estimates of Cutscore 3 than Cutscores 1 or 2, which might suggest the presence of a ceiling effect, as Cutscore 3 is the furthest Cutscore from the mean item difficulty value. For both methods, the mean cutscore estimates got worse as the level of inconsistency increased.

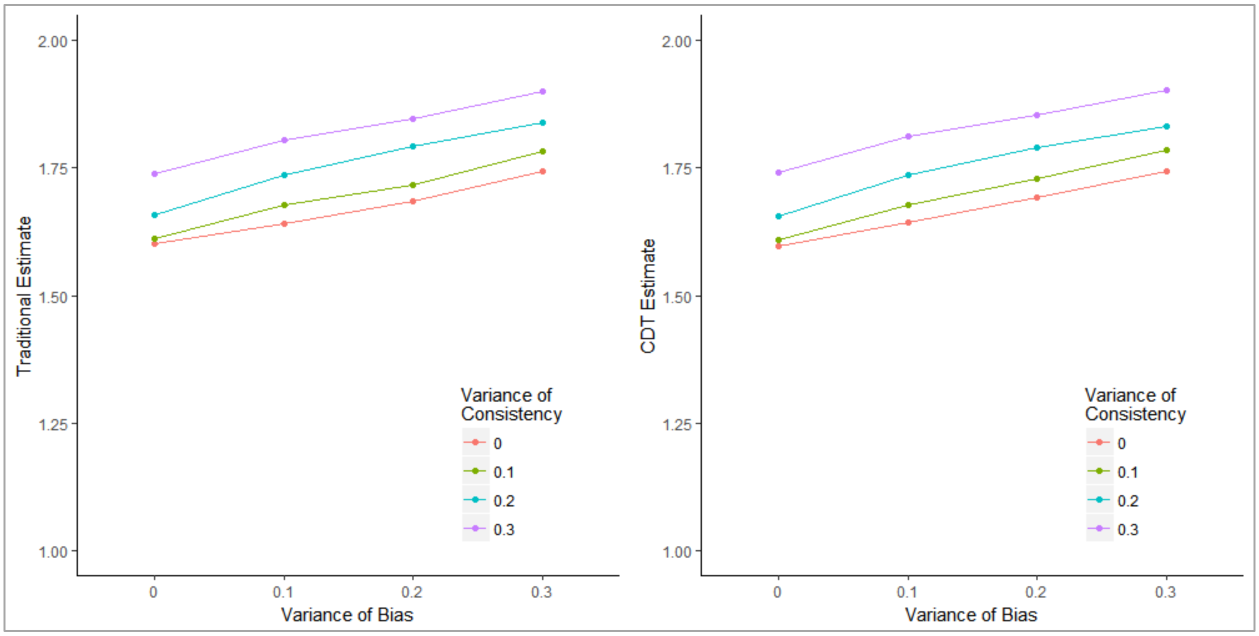

Figure 19. Mean Cutscore 3 recovery for each method with 2 groups of 8 panelists.

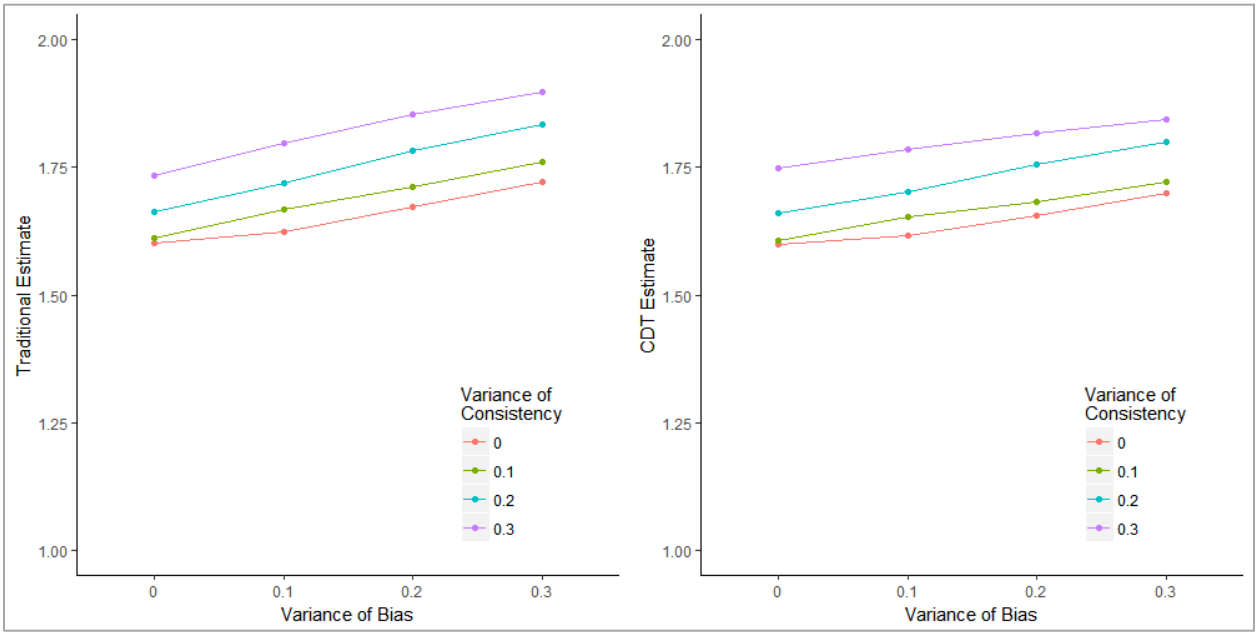

Figure 20. Mean Cutscore 3 recovery for each method with 3 groups of 8 panelists.

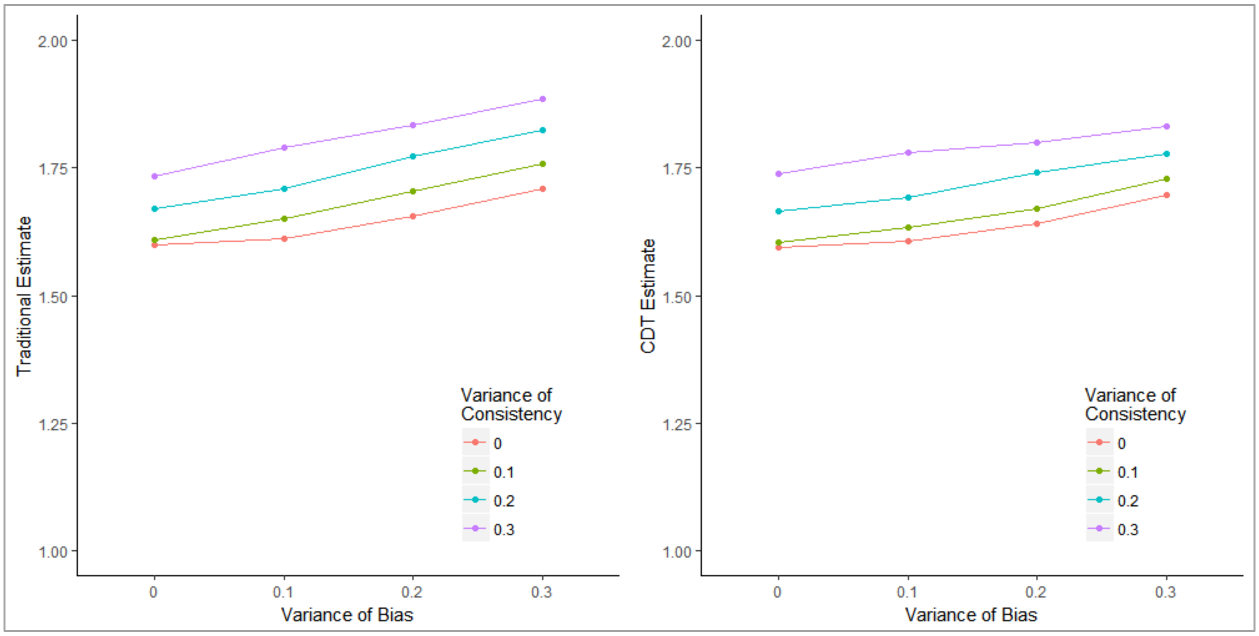

Figure 21. Mean Cutscore 3 recovery for each method with 4 groups of 8 panelists.

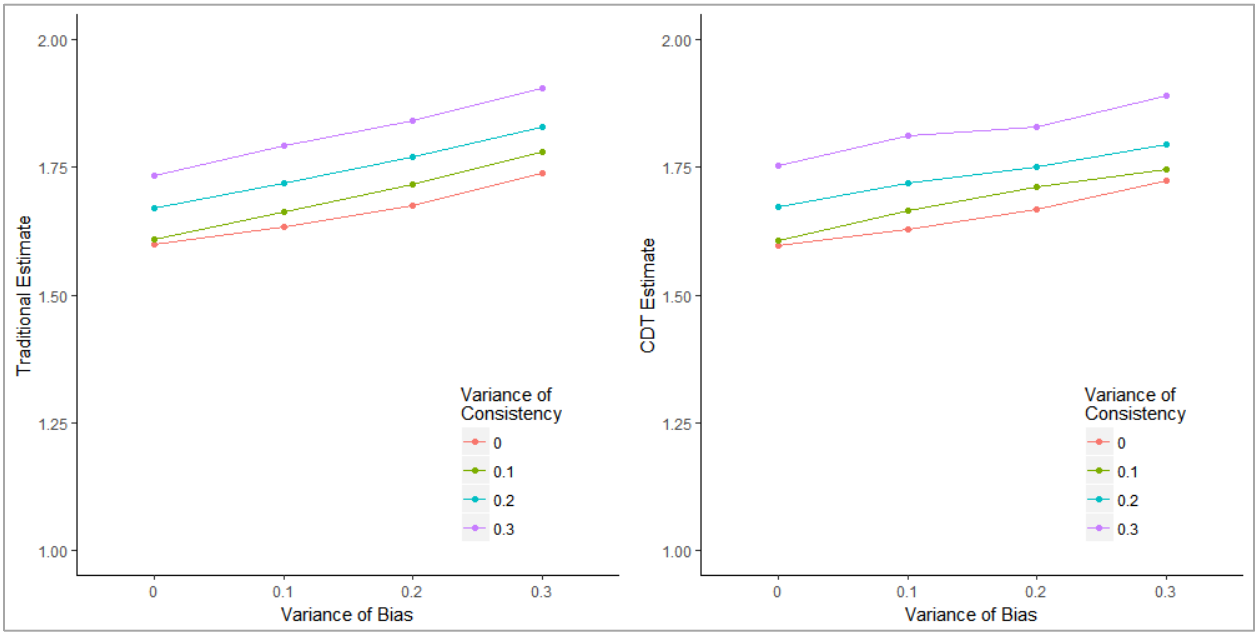

Figure 22. Mean Cutscore 3 recovery for each method with 2 groups of 16 panelists.

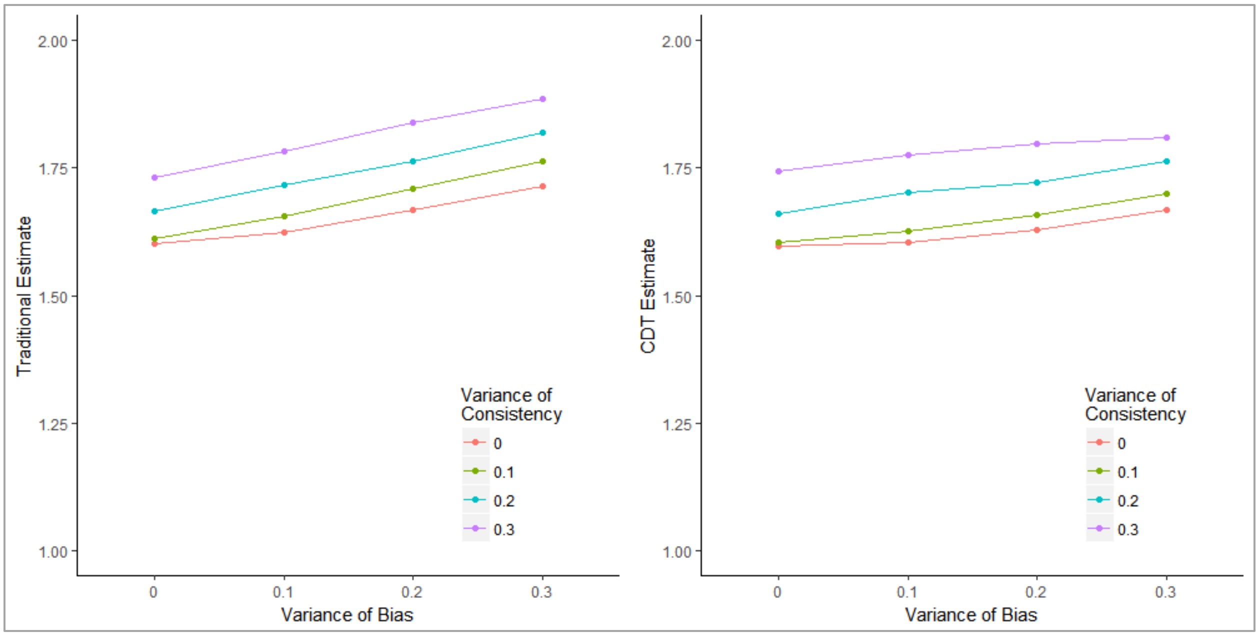

Figure 23. Mean Cutscore 3 recovery for each method with 3 groups of 16 panelists.

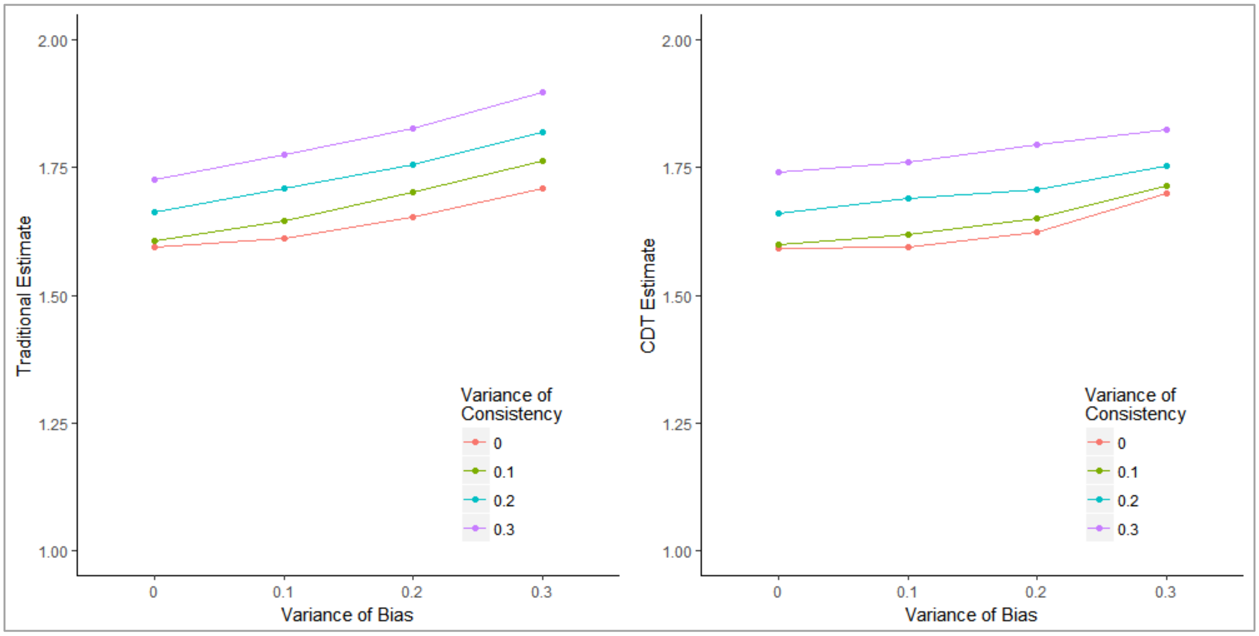

Figure 24. Mean Cutscore 3 recovery for each method with 4 groups of 16 panelists.

Cutscore error variance. The following sections summarize the estimates of error variance associated with the cutscore estimate for each method, for each cutscore.

Cutscore 1. Results for Cutscore 1 are presented in condition summary Tables 1 through 6, and results for each condition are shown in Figures 25 through 30. The data-generating model implies that the two non-random components of error variance (variance of bias and the variance of consistency) should be additive, so that the "true" error variance should be their sum. However, because the simulated bookmark placements were forced to be ordered, the actual error variance in each condition is smaller than the expected value of $\sigma_b^2 + \sigma_c^2$. This discrepancy is discussed further in Chapter V.

Figures 25 and 26 are plots of the error variance estimates under each method with only one group of judges. The plots in these two figures show a roughly linear increase in error variance estimates under each method as the variance of consistency increases. Also, the error variance estimates decreased by nearly 50% by doubling the sample size, from 8 to 16 panelists. The CDT estimates are always larger than the Traditional estimates, although both are much smaller than their expected values (which in this case are the consistency variance values themselves).

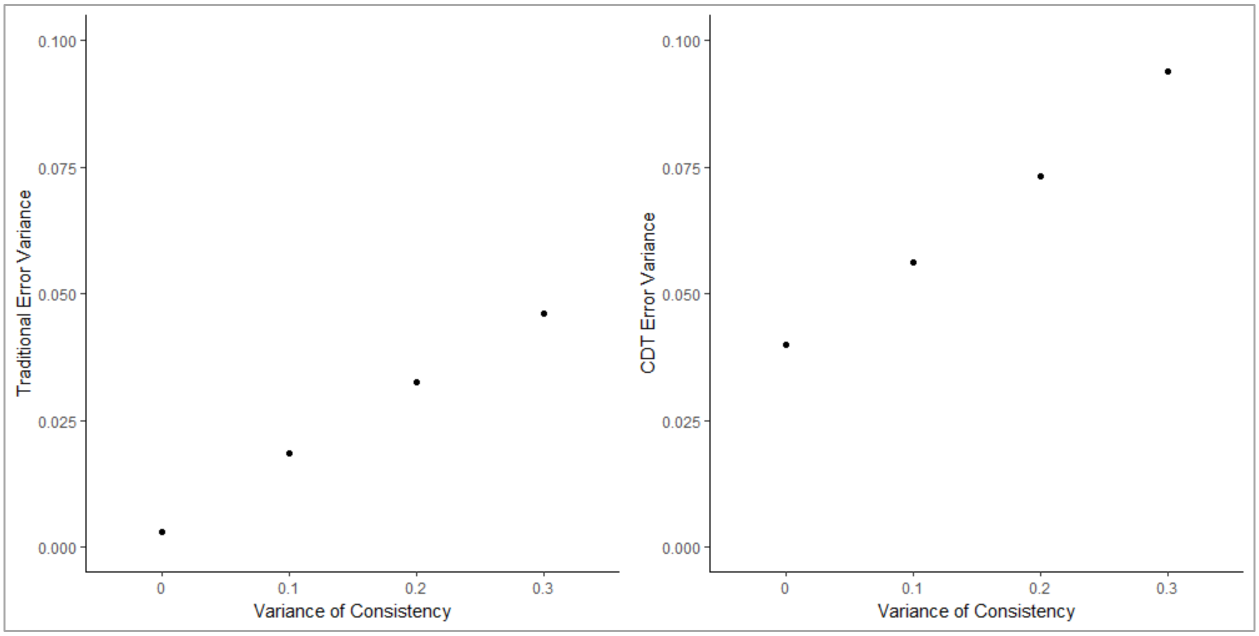

Figure 25. Cutscore 1 error variance estimates for each method, with 1 group of 8 panelists.

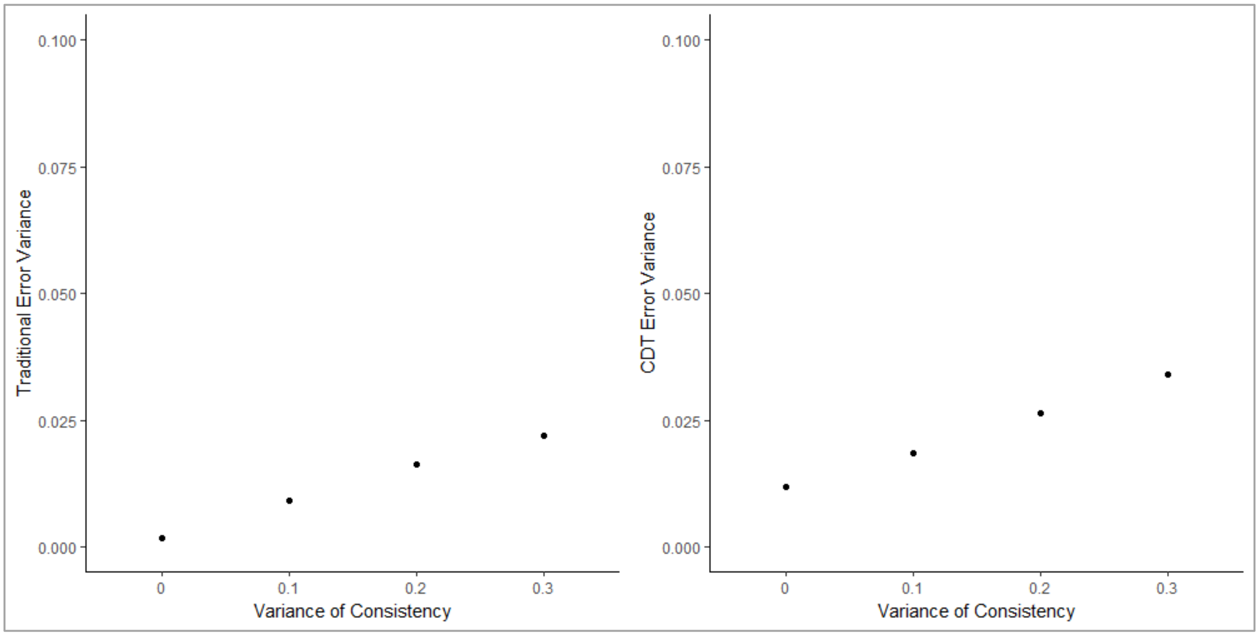

Figure 26. Cutscore 1 error variance estimates for each method, with 1 group of 16 panelists.

Figures 27 through 32 present the error variance estimates for the multiple-group conditions. With three or four groups of panelists, the results for both estimation methods were similar, although the CDT method produced larger estimates of error variance and appeared slightly more influenced by increased bias. With two groups, however, the CDT method was much more influenced by the group-level bias than the Traditional method. As was the case with a single group, the estimates of error variance decreased as the group size increased from eight to 16 panelists.

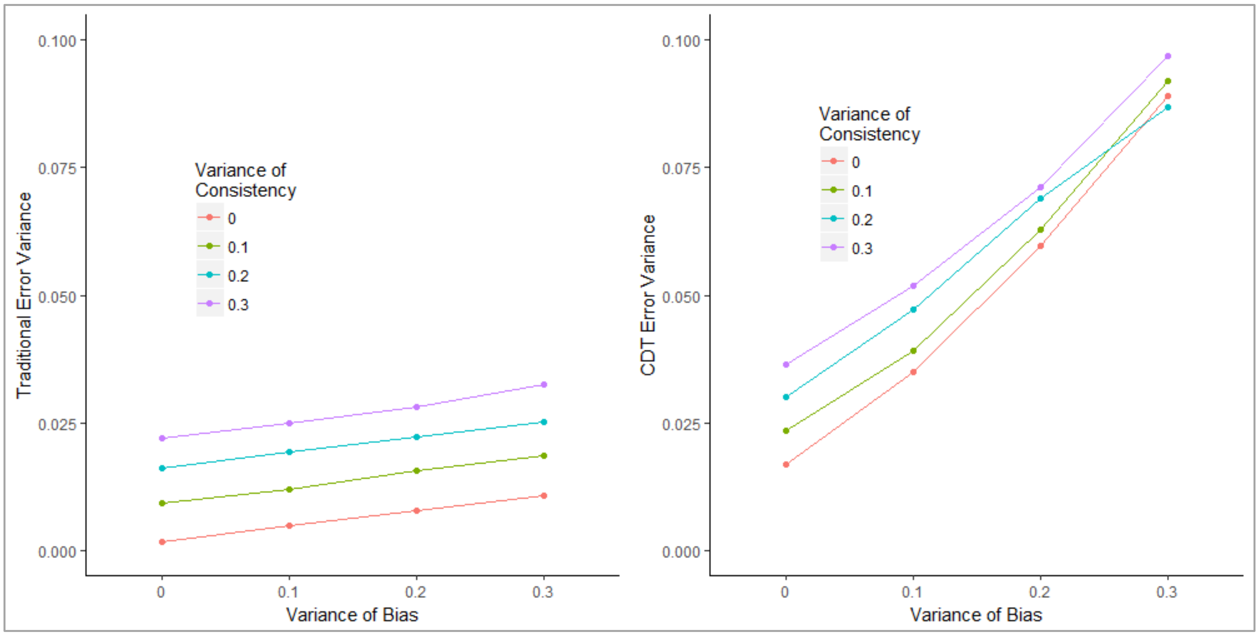

Figure 27. Cutscore 1 error variance estimates for each method, with 2 groups of 8 panelists.

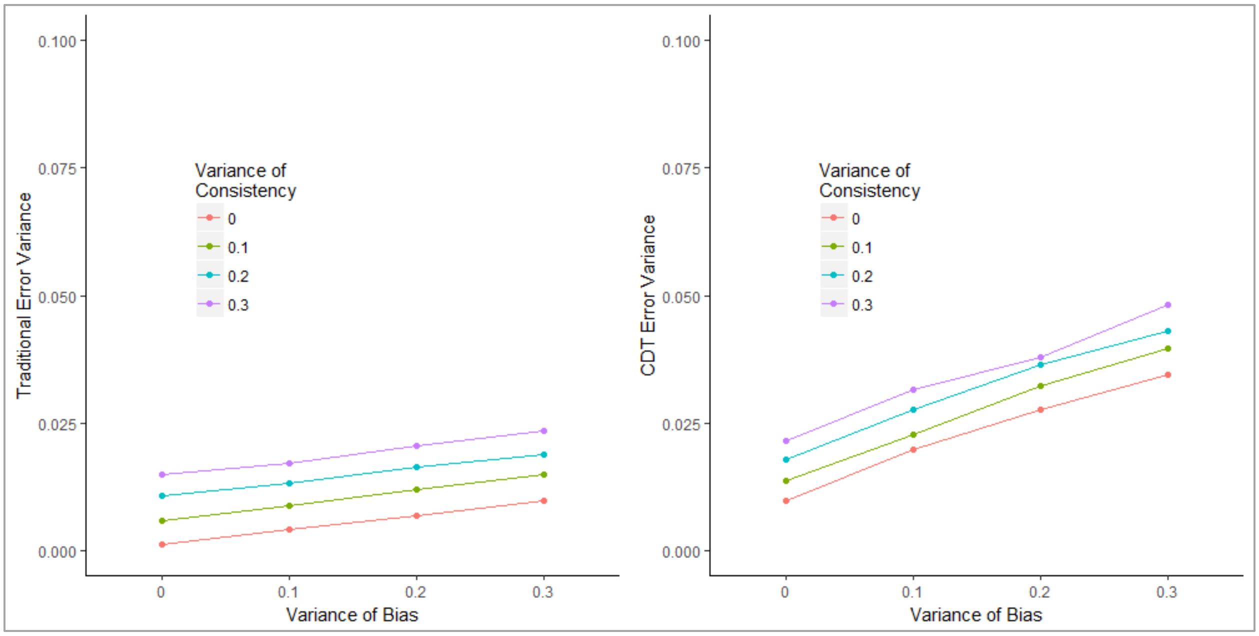

Figure 28. Cutscore 1 error variance estimates for each method, with 3 groups of 8 panelists.

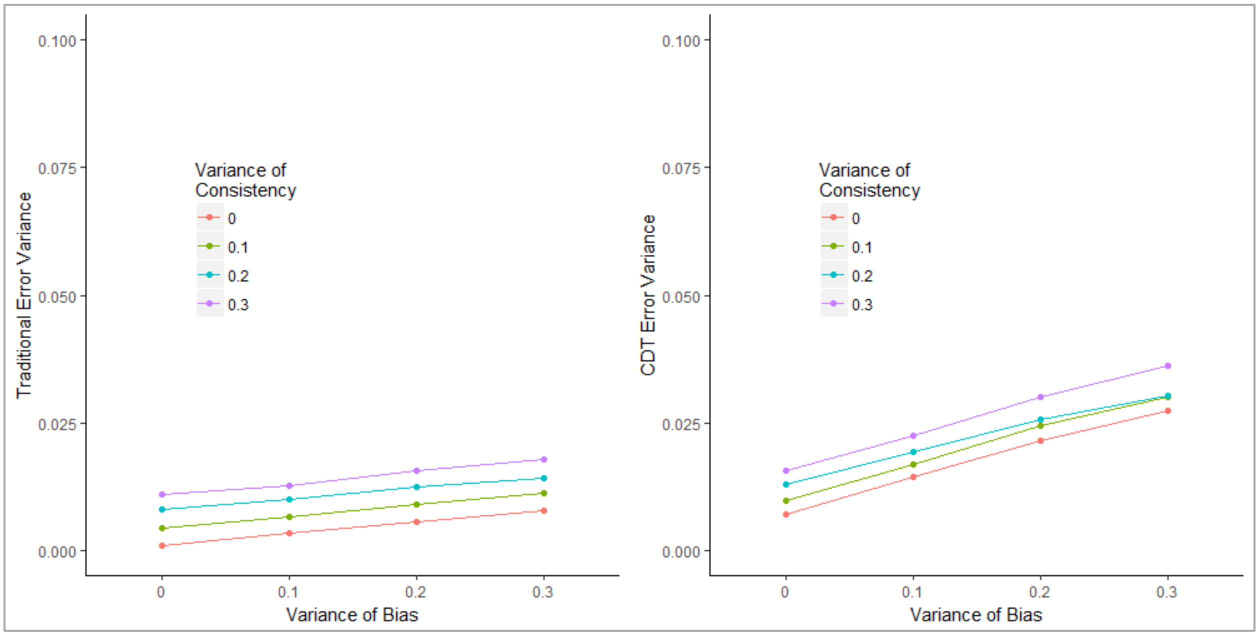

Figure 29. Cutscore 1 error variance estimates for each method, with 4 groups of 8 panelists.

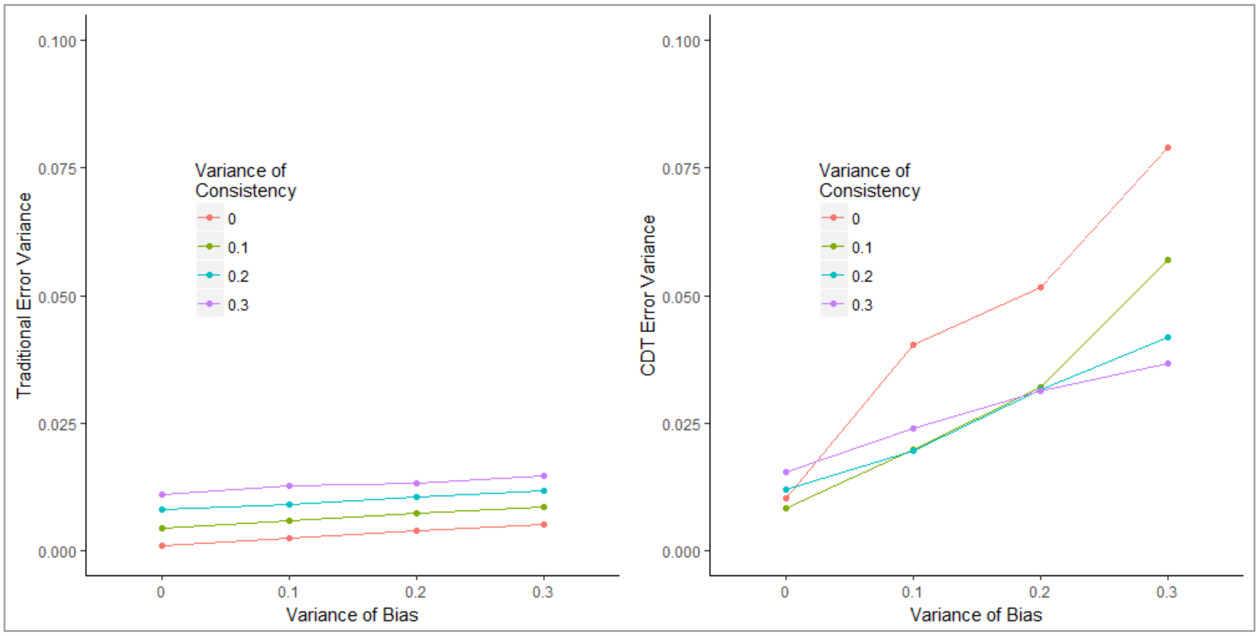

Figure 30. Cutscore 1 error variance estimates for each method, with 2 groups of 16 panelists.

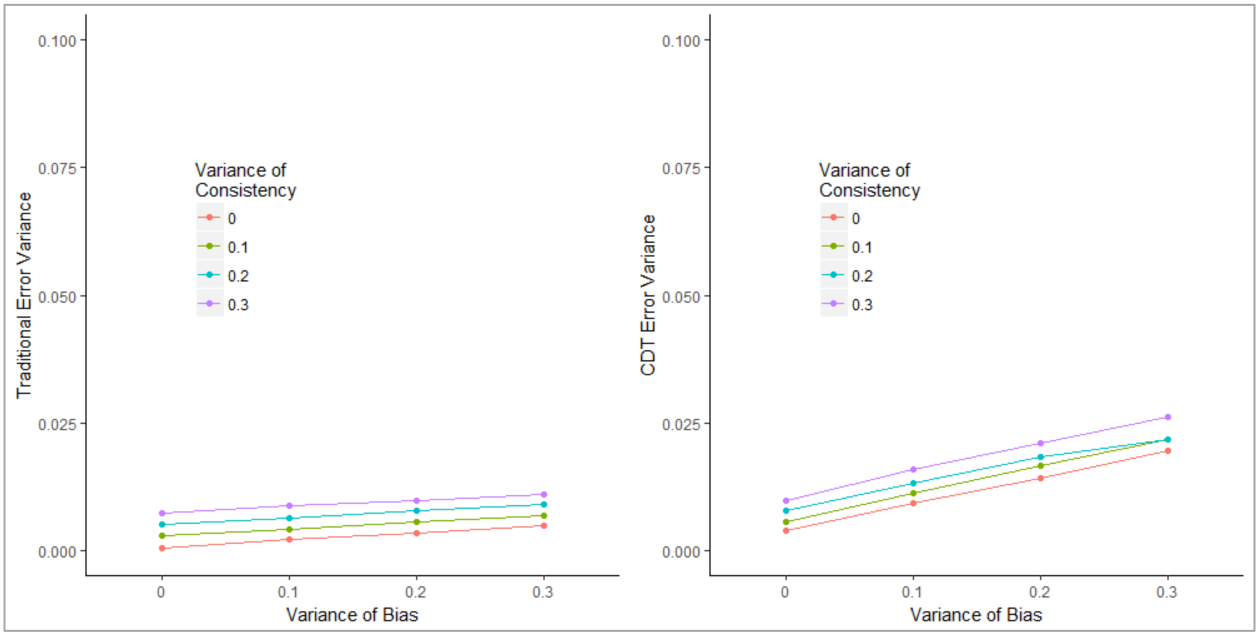

Figure 31. Cutscore 1 error variance estimates for each method, with 3 groups of 16 panelists.

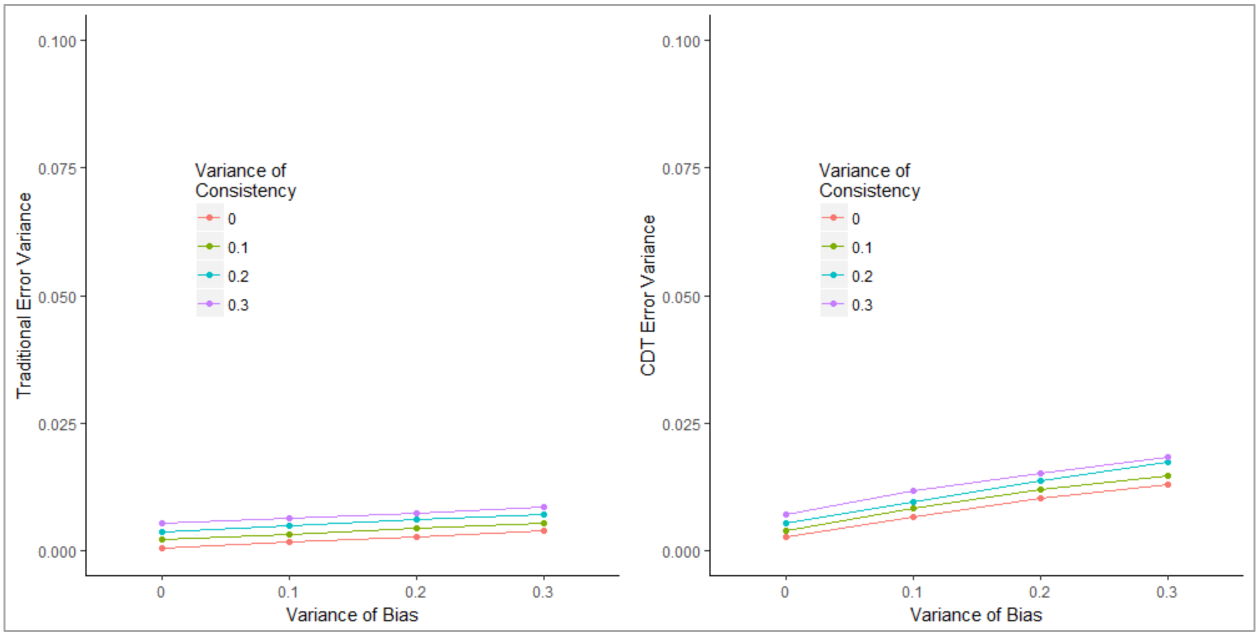

Figure 32. Cutscore 1 error variance estimates for each method, with 4 groups of 16 panelists.

Cutscore 2. Results for Cutscore 2 are presented in condition summary Tables 7 through 12, and results for each condition are shown in Figures 33 through 40.

Figures 33 and 34 are plots of the error variance estimates under each method with only one group of judges. The plots in these two figures show a roughly linear increase in error variance estimates under each method as the variance of consistency increases. As with Cutscore 1, the error variance estimates decreased substantially when the size of the panel was doubled.

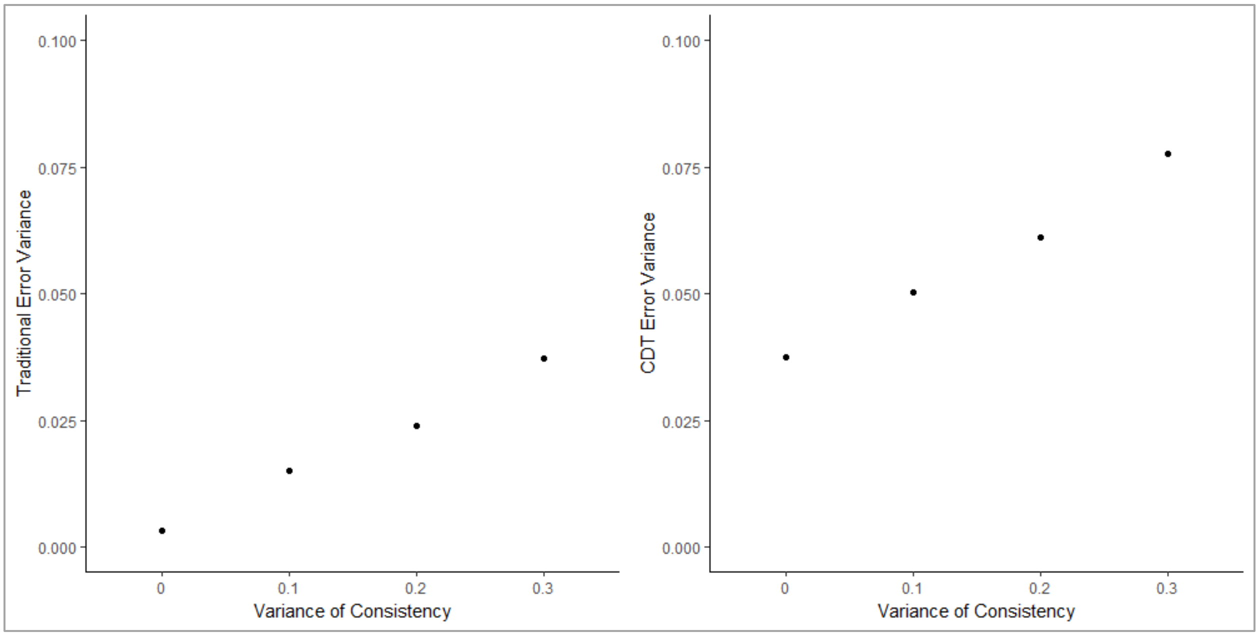

Figure 33. Cutscore 2 error variance estimates for each method, with 1 group of 8 panelists.

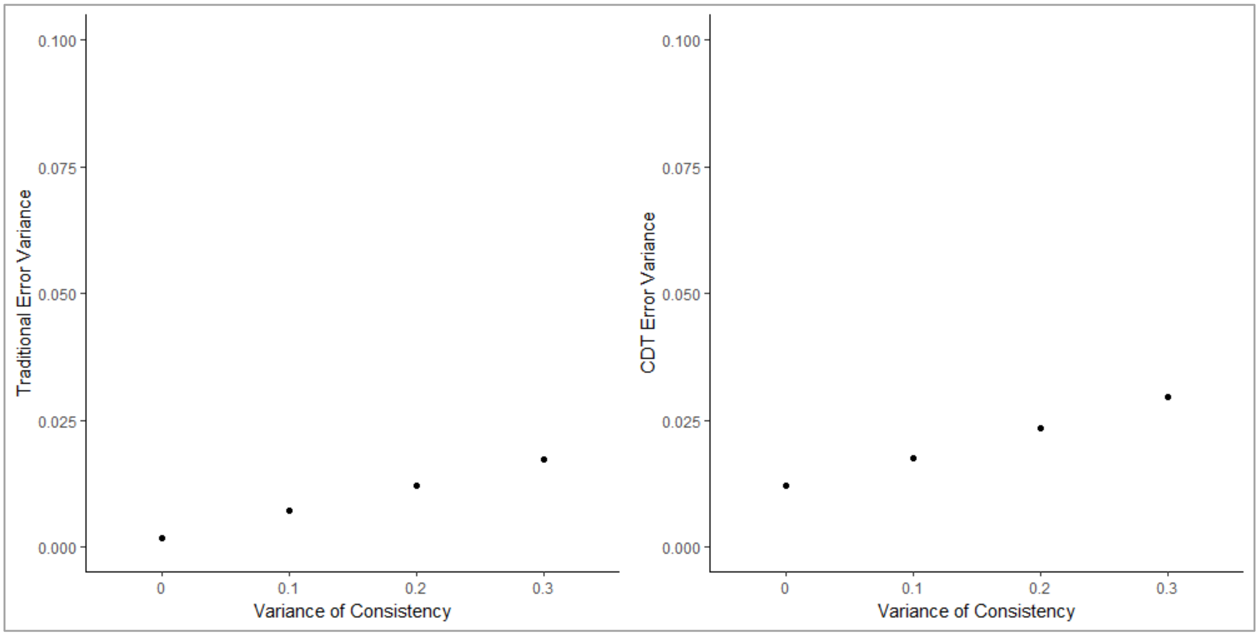

Figure 34. Cutscore 2 error variance estimates for each method, with 1 group of 16 panelists.

Figures 35 through 40 present the error variance estimates for Cutscore 2 in the multiple-group conditions. As with Cutscore 1, the most discrepant results between the two methods occurred in the conditions with two groups of panelists, where the CDT method was again more heavily influence by the group-level bias than the Traditional method. In particular, the two-group conditions were most influenced by bias when the within-group consistency value was 0.

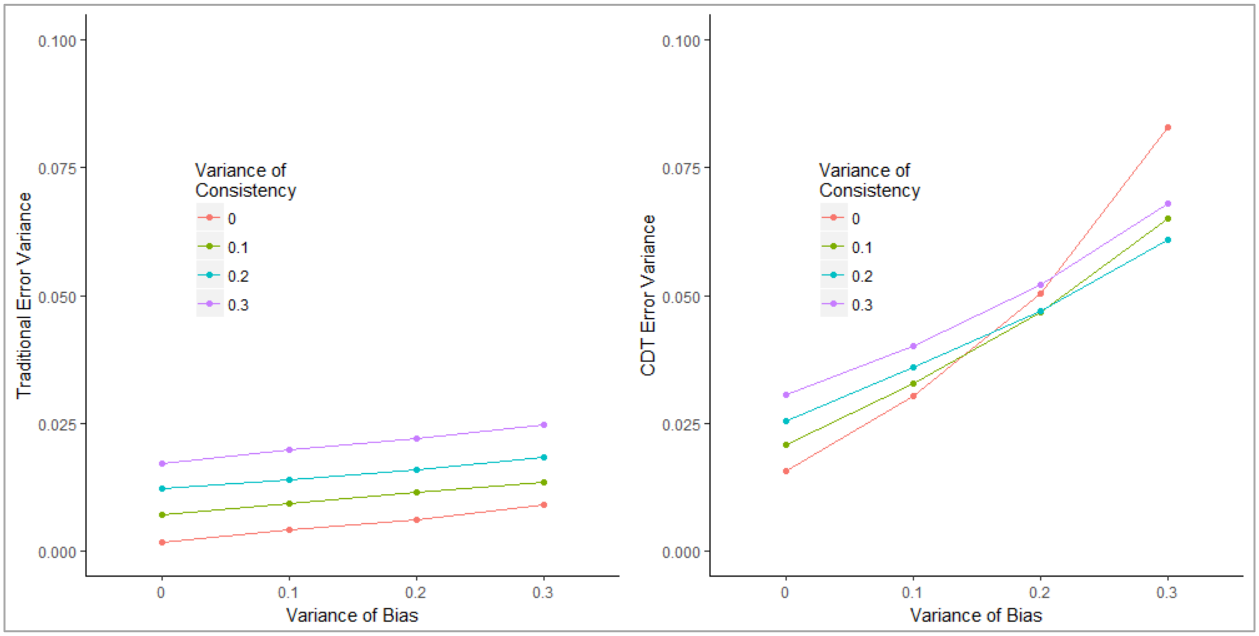

Figure 35. Cutscore 2 error variance estimates for each method, with 2 groups of 8 panelists.

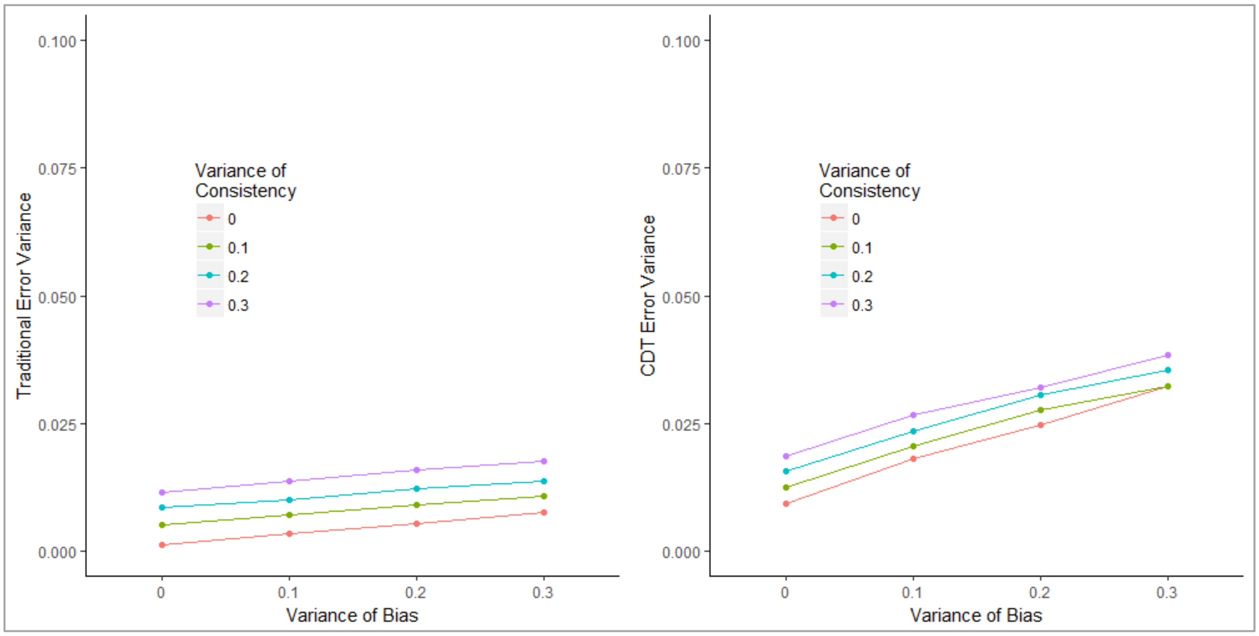

Figure 36. Cutscore 2 error variance estimates for each method, with 3 groups of 8 panelists.

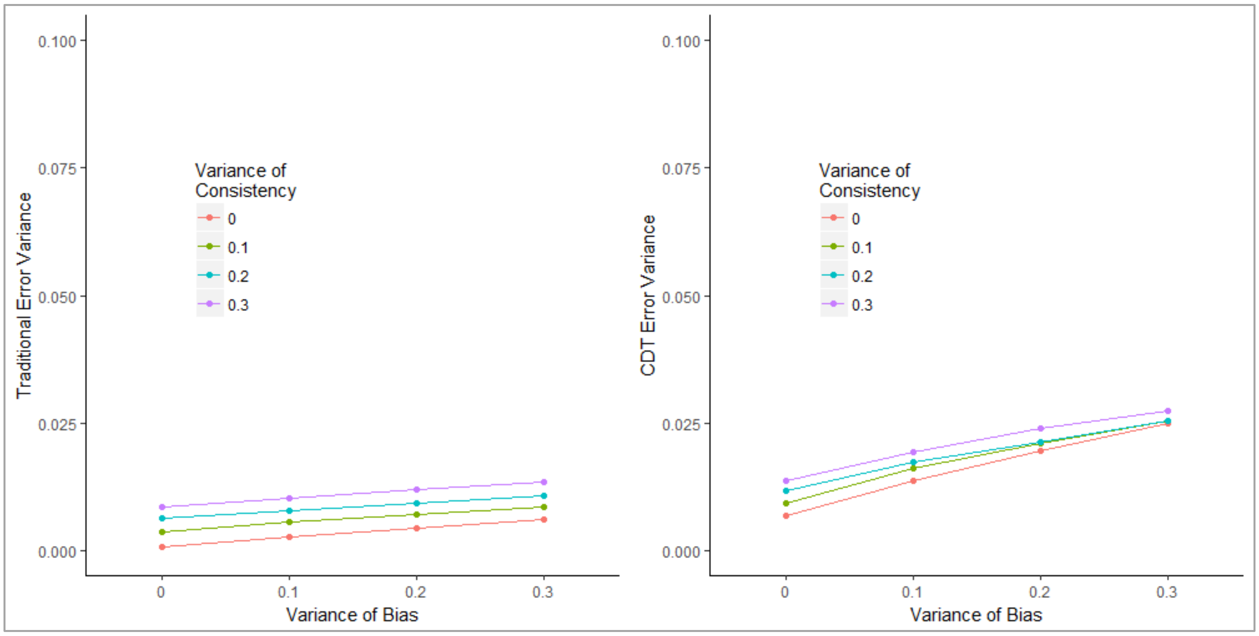

Figure 37. Cutscore 2 error variance estimates for each method, with 4 groups of 8 panelists.

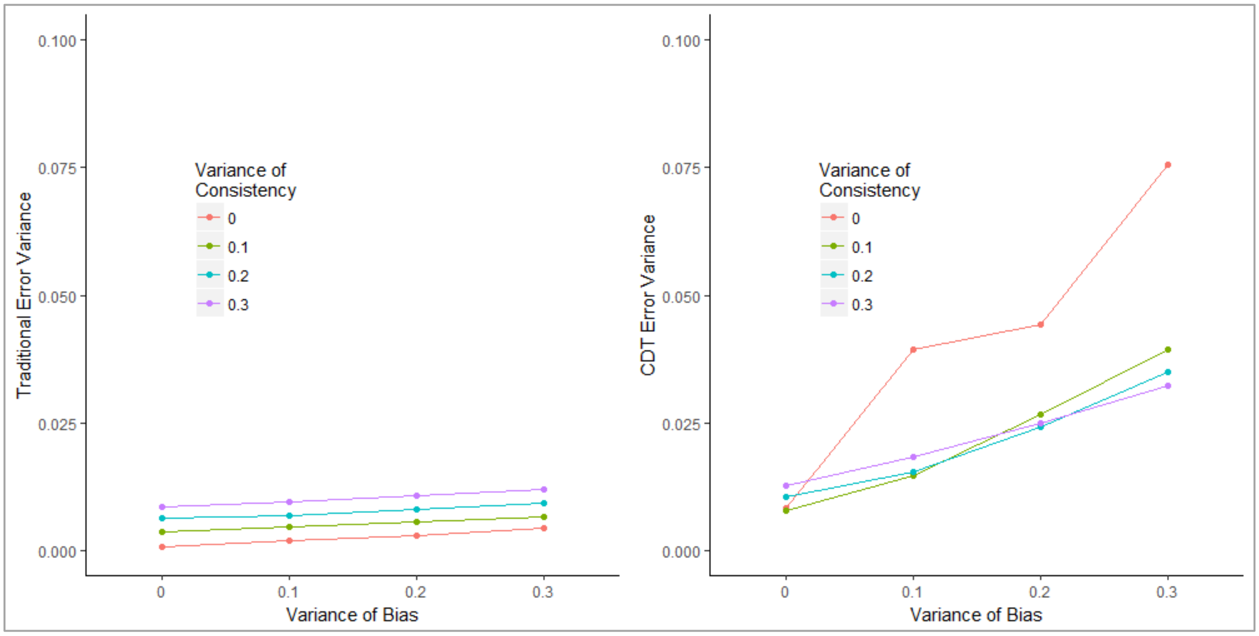

Figure 38. Cutscore 2 error variance estimates for each method, with 2 groups of 16 panelists.

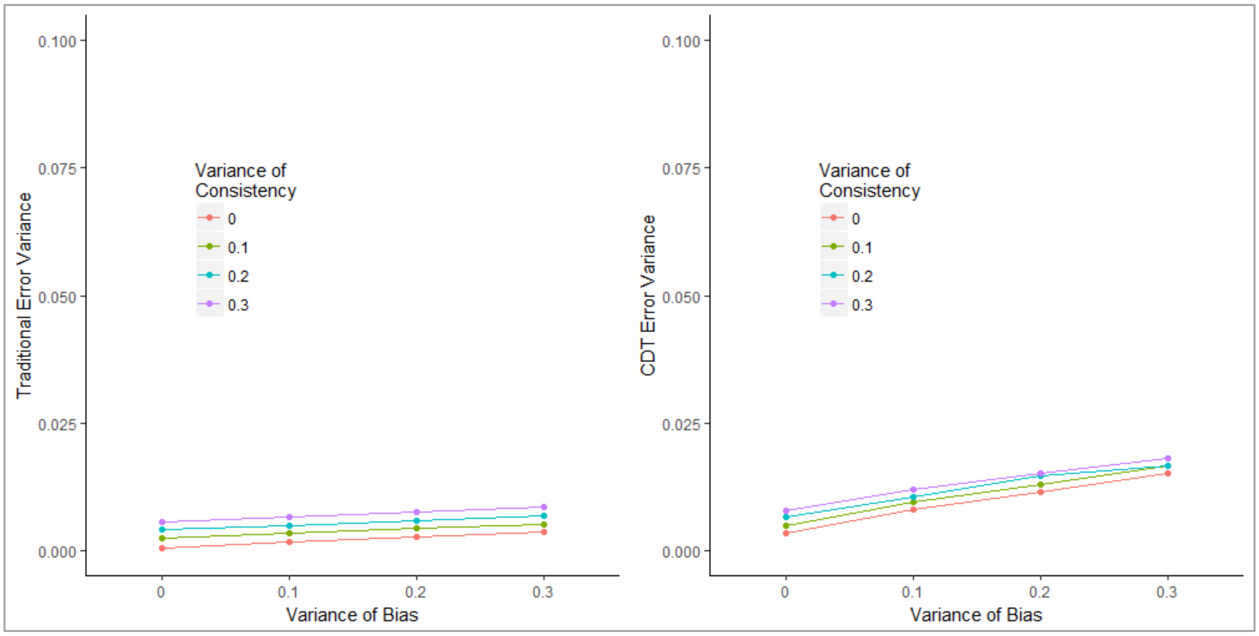

Figure 39. Cutscore 2 error variance estimates for each method, with 3 groups of 16 panelists.

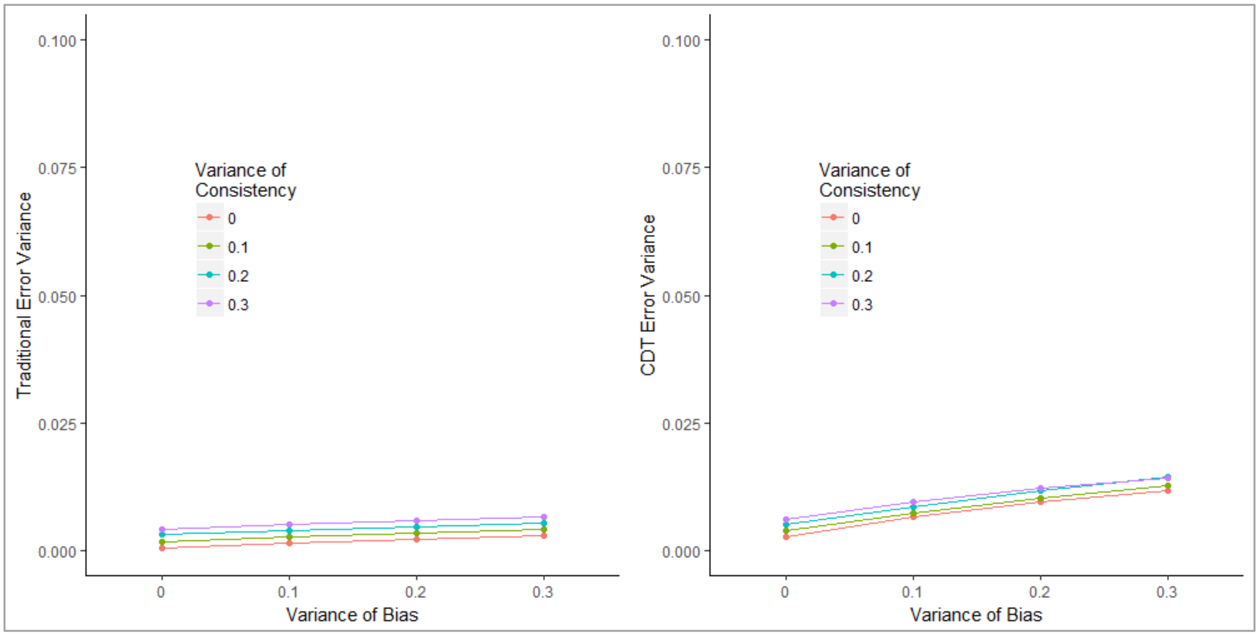

Figure 40. Cutscore 2 error variance estimates for each method, with 4 groups of 16 panelists.

Cutscore 3. Results for Cutscore 3 are presented in condition summary Tables 13 through 18, and results for each condition are shown in Figures 41 through 48.

Figures 41 and 42 are plots of the error variance estimates under each method with only one group of judges. The results for the single-group conditions for Cutscore 3 are very similar to the same conditions for Cutscores 1 and 2.

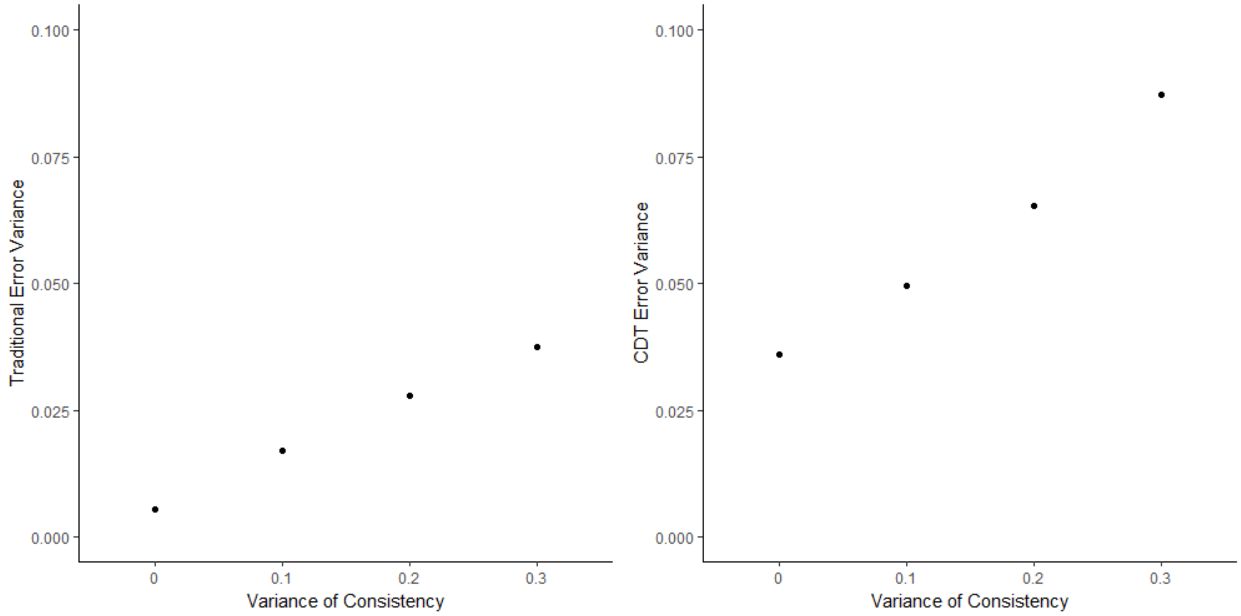

Figure 41. Cutscore 3 error variance estimates for each method, with 1 group of 8 panelists.

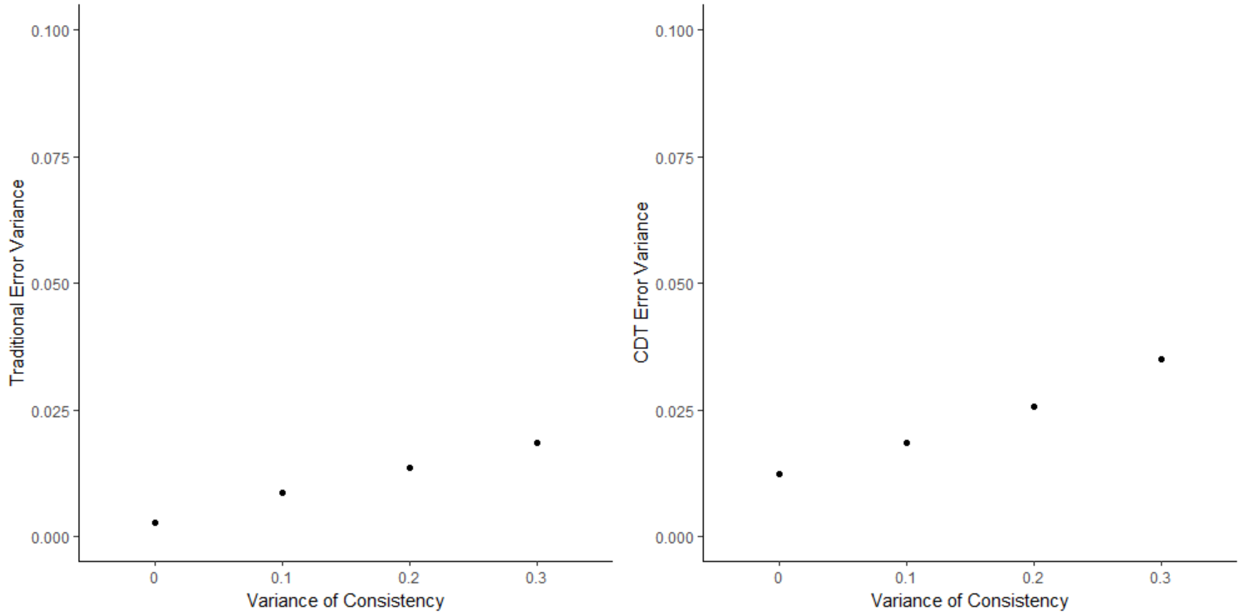

Figure 42. Cutscore 3 error variance estimates for each method, with 1 group of 16 panelists.

Figures 43 through 48 are plots of the Cutscore 3 error variance estimates for the multiple-group conditions. The results are similar to those of Cutscore 2, with the CDT estimates more influenced by the amount of bias, particular with only two groups and no within-group variance.

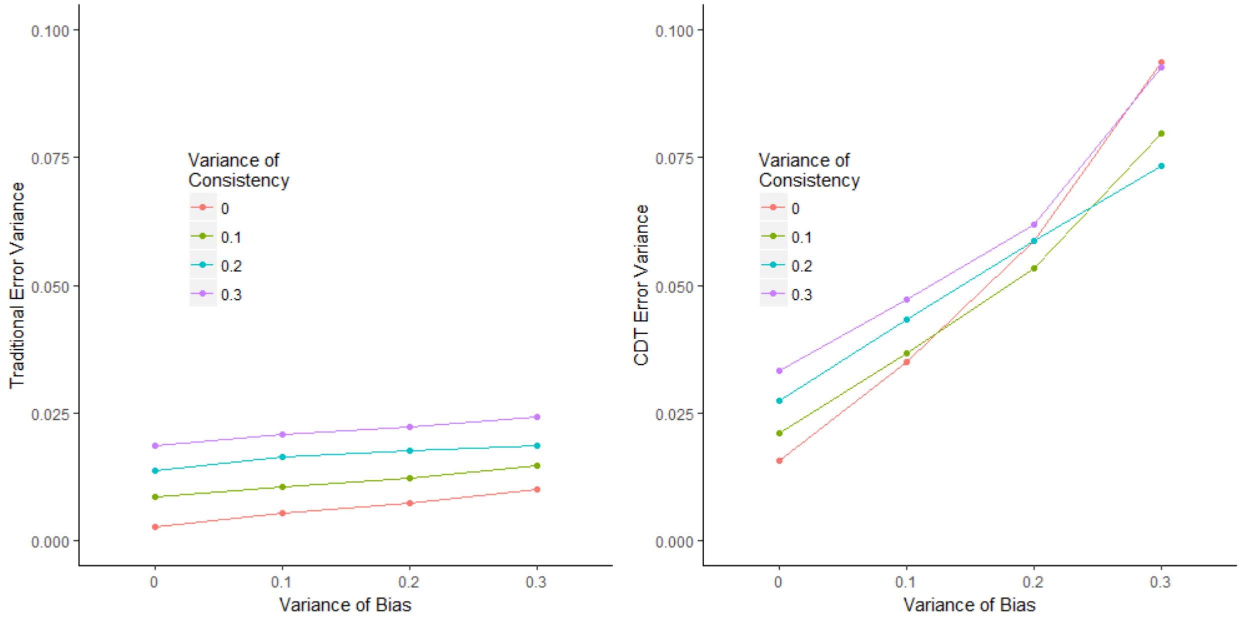

Figure 43. Cutscore 3 error variance estimates for each method, with 2 groups of 8 panelists.

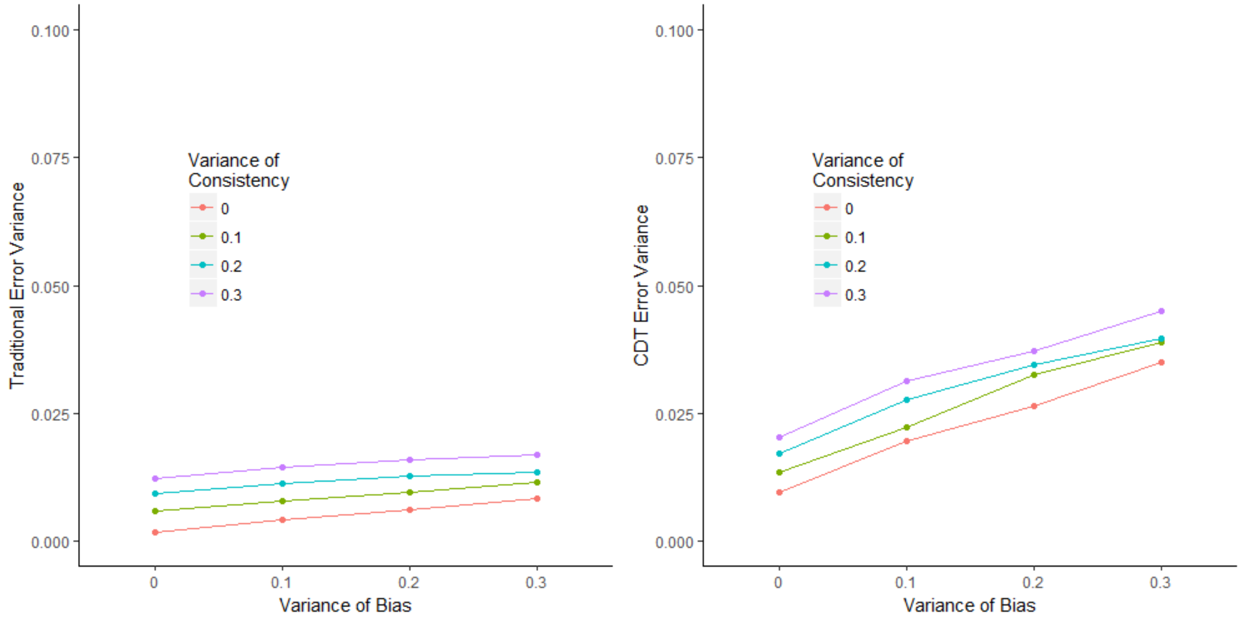

Figure 44. Cutscore 3 error variance estimates for each method, with 3 groups of 8 panelists.

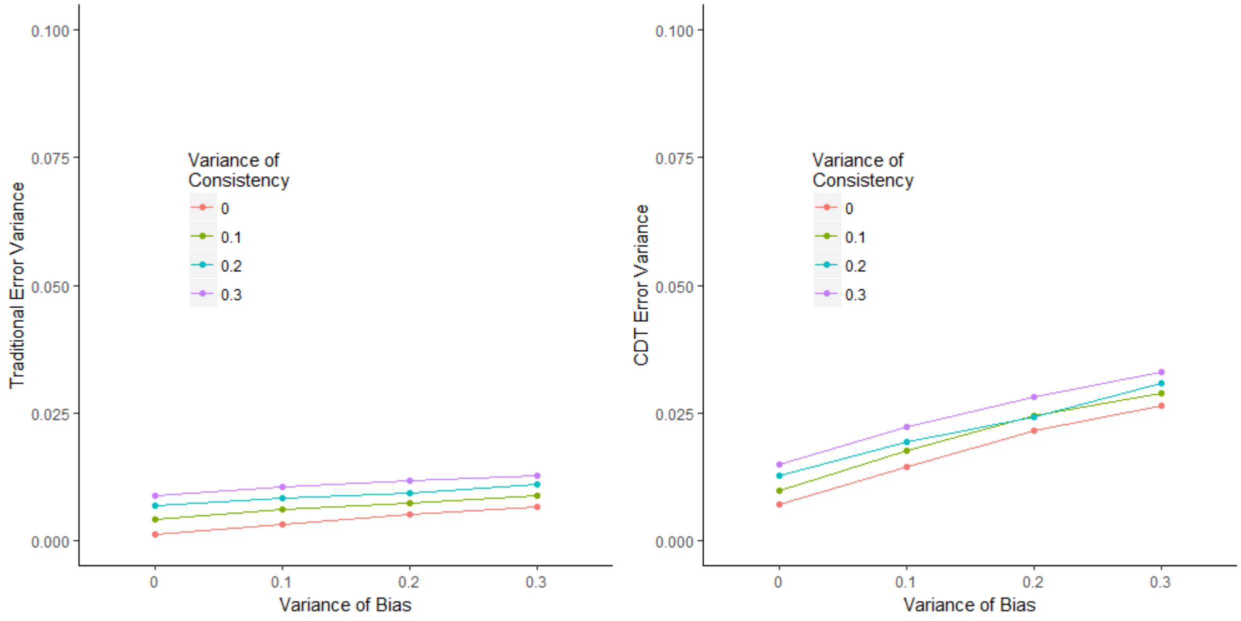

Figure 45. Cutscore 3 error variance estimates for each method, with 4 groups of 8 panelists.

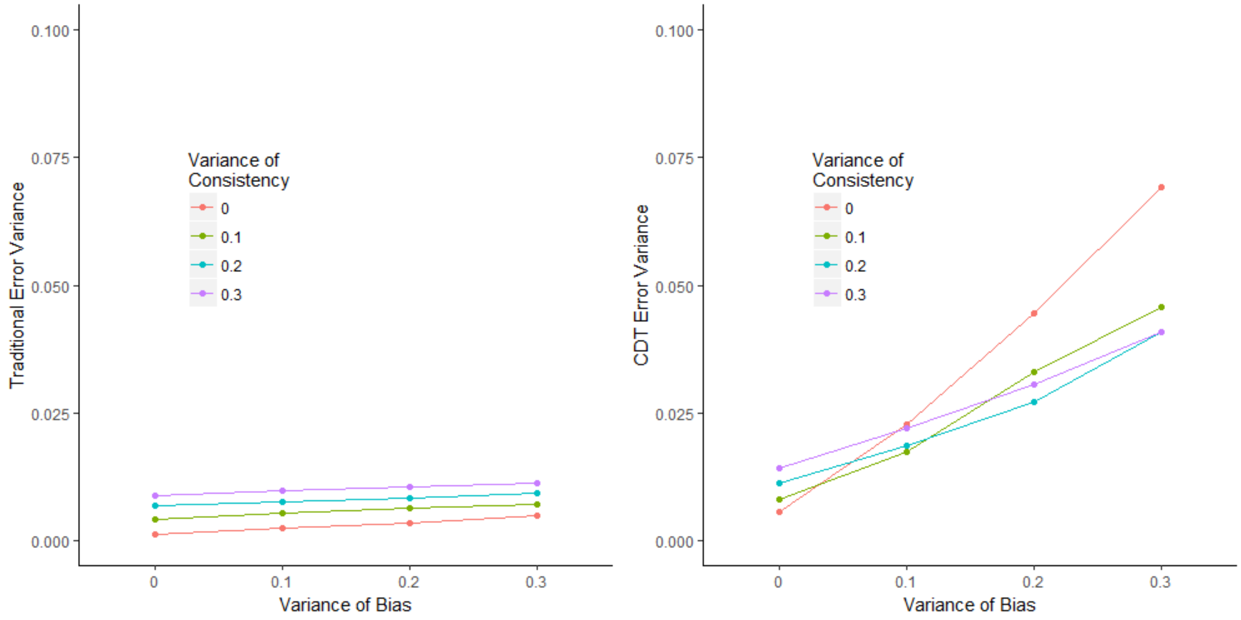

Figure 46. Cutscore 3 error variance estimates for each method, with 2 groups of 16 panelists.

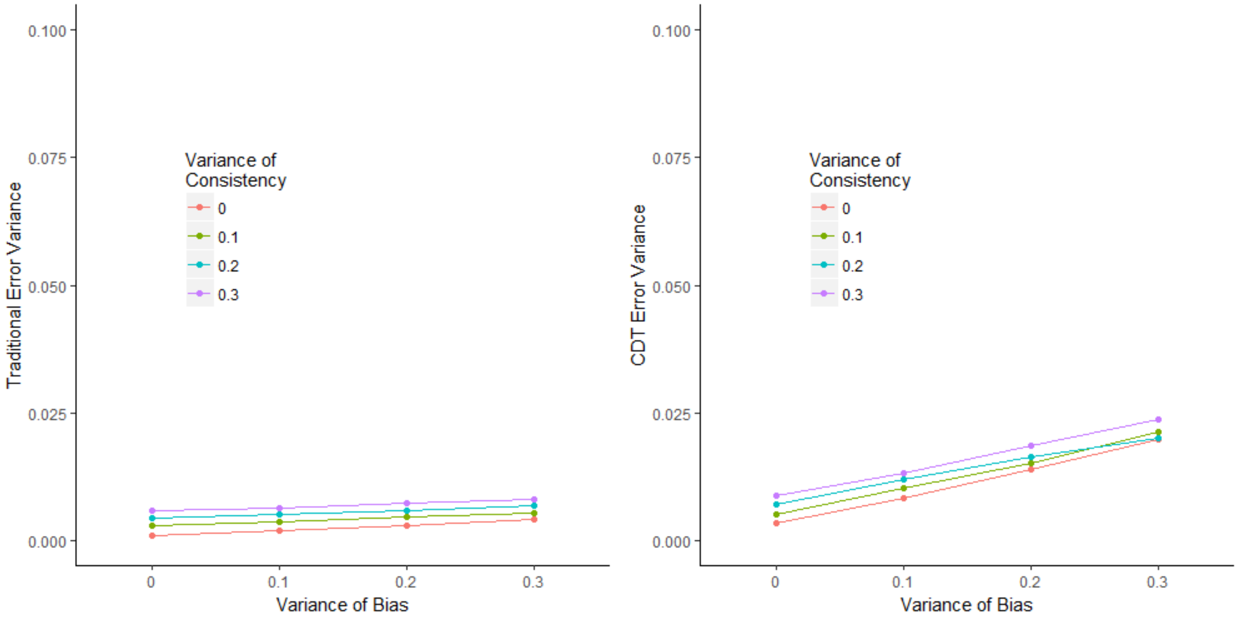

Figure 47. Cutscore 3 error variance estimates for each method, with 3 groups of 16 panelists.

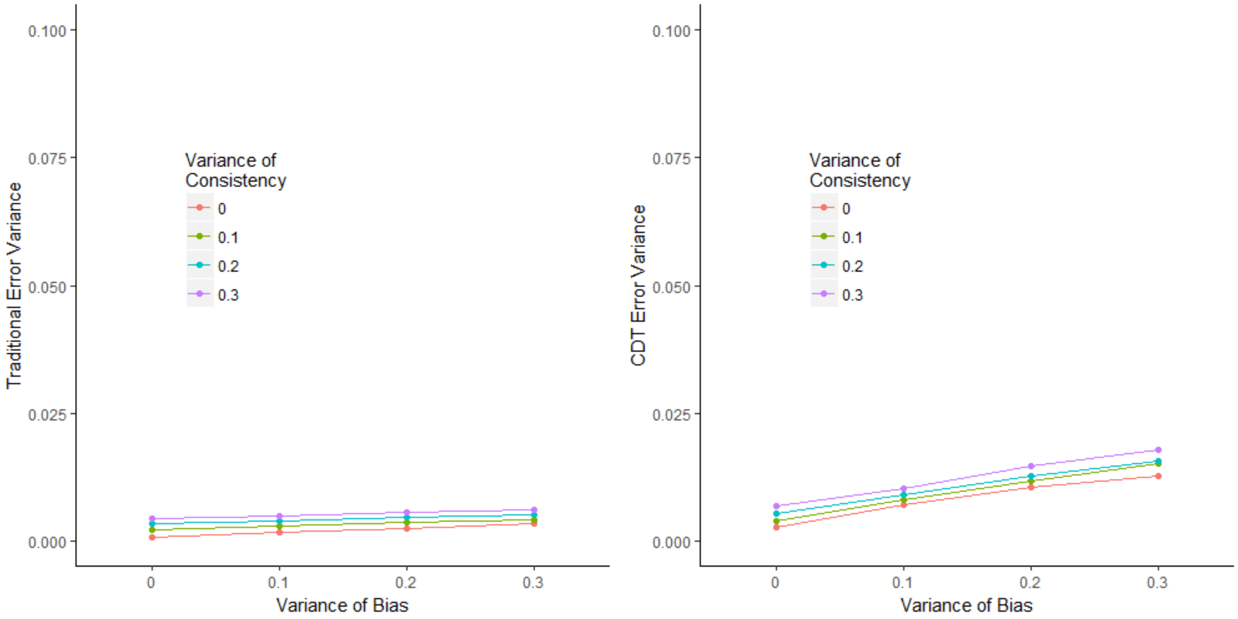

Figure 48. Cutscore 3 error variance estimates for each method, with 4 groups of 16 panelists.

Accuracy rates. As described in Chapter III, the accuracy rates presented in Tables 1 through 18 are simply the percentage of replications within each condition wherein the 95% confidence interval (for the Traditional method) or 95% credible interval (for the CDT method) contained the true cutscore. For the sake of simplicity, “C.I.” is used to describe either type of interval.

Cutscore 1. Accuracy rates for Cutscore 1 using the Traditional method ranged from 1% to 89%, with a mean of 54.3%. Using the CDT method, accuracy rates ranged from 18% to 100%, with a mean of 67.4%. The CDT method had higher accuracy rates in 92 of the 104 total conditions for Cutscore 1, while the Traditional method had a higher accuracy rate in 10 conditions. Of the 10 conditions in which the Traditional method had a higher accuracy rate, seven were conditions where the true variance of bias was 0.

Cutscore 2. Accuracy rates for Cutscore 2 using the Traditional method ranged from 0% to 86%, with a mean of 38.2%, while the CDT accuracy rates ranged from 38% to 100%, with a mean of 73.2%. The CDT method had higher accuracy rates in all 104 conditions for Cutscore 2.

Cutscore 3. Accuracy rates for Cutscore 3 using the Traditional method ranged from 4% to 86%, with a mean of 37.1%, while the CDT accuracy rates ranged from 13% to 100%, with a mean of 62.9%. As with Cutscore 2, the CDT method had higher accuracy rates in all 104 conditions for Cutscore 3.

Study 2

The second part of the studies reported here is an application of the Study 1 methodology to a real dataset. The data came from an operational standard-setting study for a large-scale state assessment program. As described in the methods section, data from two subjects (ELA and Mathematics) and several grade levels were available. However, the operational study employed a modified Bookmark procedure that varied from the traditional implementation in the way item difficulty values were calculated for ordering the OIBs. Due to these discrepancies, the panelists saw the items in a different order than what was suggested by the item *b*-parameters. Rather than relying on the operational results—which will be potentially confounded by the differences in methodology—the results reported here come from the single grade and subject in which the OIB order did match the *b*-parameter order. For the purpose of this research, the traditional cutscores calculated for this study may not match the ones calculated operationally, and were instead estimated using methods identical to those used in Study 1. The item *a*- and *b*-parameters and panelist bookmark placements were incorporated into methodology described in Chapter III. Despite the modifications to the scaling and estimation procedures employed in the operational study, the instructions to the panelists were typical of Bookmark standard-setting studies. As with the simulation study, the real panelists set three cutscores to identify four performance levels. The panelists began with Cutscore 2, as it was associated with the highest stakes.

Descriptive statistics by group are presented in Table 19. The within-group variance in Group 1 was consistent for all three cutscores, and lower than the Group 2 variance at each cutscore. The Group 2 within-group variance was large for Cutscores 1 and 3, and relatively small for Cutscore 2.

Table 20 presents the cutscores and error variance estimates for each cut under the traditional method and the CDT method. Note that the cutscores for the Traditional method in Table 20 match the “Overall” cutscores in Table 19. The CDT cutscore estimates were lower for Cutscore 1 and higher for the Cutscores 2 and 3 than those calculated with the traditional method, a finding that is fairly consistent with the simulation study results. The error variance estimates produced by each method were very similar for Cutscore 1, but larger with the CDT method for Cutscores 2 and 3.

The 95% confidence interval (for the traditional method) and 95% credible interval (for the CDT method) are presented in Table 21. As with the error variance estimates, these show that

CDT produces similar estimates for the first cutscore, but larger and less certain estimates for cutscores 2 and 3. Importantly, note that the Traditional cutscore estimate is always at the exact center of the 95% C.I., while the CDT estimate is not.

Table 1. Mean cutscores (and SD) for each group and overall.

	Group 1	Group 2	Total
N	8	7	15
Cut 1	0.06 (0.23)	-0.99 (0.93)	-0.43 (0.82)
Cut 2	1.03 (0.19)	0.67 (0.47)	0.86 (0.39)
Cut 3	1.68 (0.25)	2.76 (1.17)	2.18 (0.97)

Table 2. Cutscores and error variance estimates for each cutscore and method.

		Cut 1	Cut 2	Cut 3
Cutscores	Traditional	-0.43	0.86	2.18
	CDT	-0.51	1.29	2.71
$\hat{\sigma}_E^2$	Traditional	0.046	0.010	0.062
	CDT	0.049	0.039	0.085

Table 3. 95% C.I. for each method for each cutscore.

	Traditional		CDT	
	Lower	Upper	Lower	Upper
Cut 1	-0.85	-0.01	-1.00	-0.11
Cut 2	0.67	1.06	0.89	1.66
Cut 3	1.69	2.67	2.21	3.33

Chapter V: Discussion

The studies described here were conducted to try to develop a psychometrics of cutscores set using the Bookmark standard-setting method. Because the cutscores established during a standard-setting study are typically considered a recommendation rather than a final result, the amount of uncertainty around those cutscores carries great practical significance. Policy-making bodies that set the final cutscore often use the confidence intervals around a recommended cutscore as a justification for adjusting it slightly upward or downward (e.g., Cizek & Bunch, 2007).

Traditionally, the error variance around Bookmark cutscore recommendations has been estimated as a function of the consistency of the individual members of the standard-setting panel (i.e., the standard error of the mean recommendation). The current study attempted to improve that estimate through two main differences from the traditional methodology: 1) calculating both within- and between-group variance components, taking advantage of the fact that most standard-setting studies ask panelists to work in small-groups as part of the study design, and 2) estimating the variance components using Bayesian analysis. It was hypothesized that incorporating group-level variance in estimating cutscores would allow for more accurate estimates of the true amount of variance among the population of possible standard-setting panelists. By incorporating Bayesian estimation, the proposed methodology may be viewed as an extension of Cutscore Distribution Theory (CDT; Skorupski & Fitzpatrick, 2014; Skorupski, Zhao, Fitzpatrick, & Chen, 2015), which described a model-based method for estimating cutscores set by the modified Angoff method.

Study 1 was a simulation study conducted to compare the CDT method results with traditional results, varying the size of the standard-setting sample, the number of groups within which the panelists worked, and the size of the population-level variance within each group and between different groups. Study 2 applied the estimation methodology of Study 1 to the results of a real Bookmark standard-setting study.

Together, the two studies were designed to address three main research questions, which were presented in Chapter 1. Question 1 addresses the similarity of the 95% CIs between the traditional method and the proposed CDT method. Question 2 focuses on the cutscore estimates from each method. Question 3 asks how the estimates from each method are affected by the size of the panel and the number of groups. In this chapter, the results of each study are discussed separately, followed by some general conclusions and implications, as well as considerations for further research.

Study 1

The first study used simulated data to compare the CDT cutscore and error variance estimates to those calculated using traditional methods. Four independent variables were manipulated: the variance of bias (i.e., between-group variance) and consistency (i.e., within-group variance) of the simulated panelists' intended cutscores in the data-generating model; the number of simulated panelists per group; and the number of groups. The conditions were fully crossed, except for conditions with only one group (in which case the between-group variance could only be zero). The dependent variables of interest were the cutscore and error variance estimates under each method. Each condition was replicated 100 times, and the reported outcomes for each condition were averaged across replications.

Coverage of 95% CIs. The primary variables of interest in the simulation study were the 95% CIs around each cutscore estimate. Because the outcome of a typical Bookmark study is a recommendation rather than a final cutscore, these CIs are often presented to the policymaking body as a window within which to adjust the recommended cutscore up or down. As is evident from Figures A.1 through A.204 in the Appendix, the CDT method always produced larger CIs than the traditional method. As a result, those CIs also included the true cutscore more often than those of the traditional SEM method.

The difference is perhaps most striking in the conditions in which there was little to no variance in the panelists' intended cutscores. For example, Figures A.1 and A.2, along with Table A.1, show the results of all 100 replications for the condition with one group of eight panelists, where the generating variances of bias and consistency were 0. The only differences in bookmark placements across the sample were due to random error through the stochastic process by which the panelist "item responses" were generated. In this condition, the traditional CIs were expected to be small—which they were, compared with the CDT CIs—but they often missed the mark, and included the true values for Cutscores 1, 2, and 3 in just 47%, 66%, and 81% of the replications, respectively. The CDT method produced much larger estimates of error variance, resulting in 95% CIs that included the true cutscore in 100% of the replications for all three cutscores. Even though the simulated panelists in this condition were very consistent, the fact that the sample size was small and the study used only one group resulted in more uncertainty in the model-based approach than is accounted for in the traditional calculation.

Recovery of true cutscores. The traditional and CDT methods varied in their recovery of the true cutscores by the placement of the true cutscores along the ability continuum, and the method differences in cutscore estimates appeared to be fairly predictable when viewed as a whole. The traditional method overestimated all three true cutscores for every condition, while the CDT method estimates appeared to be biased outward. That is, the CDT method underestimated the lowest cutscore, overestimated the highest cutscore, and was very close to the most central cutscore (whether the estimate was high or low depended on the variances of bias and consistency).

Although these results require follow-up study to examine the CDT cutscore recovery across a broad range of true cutscores, the trend observed here suggests that the CDT model may need improvement when multiple cutscores are being set. The multivariate normal prior, in particular, may act to increase—or even simply equalize—the differences among the three cutscores. In some testing situations, it might be reasonable to expect roughly even spacing between cutscores along the ability continuum. In many cases, especially when many cutscores are being set on the same test, a wide dispersion of cutscores may even be preferable. For example, if multiple cutcores fall toward the high end of the scale (e.g., cutscores for "Proficient" and "Advanced" performance levels) it is possible that the difference between them is not much larger than one or two conditional standard errors of measurement for the test. In that case, though, it is probably preferable to address the dispersion of cutscores during panelist training activities or between standard-setting rounds, rather than asking the model to "fix" things after the fact. Ideally, the standard-setting model will let the panelists' bookmark placements speak for themselves. The largest sample size tested in Study 1 consisted of 64 panelists, which is much larger than a typical Bookmark standard-setting panel. Such a large sample should be more than enough to let the data outweigh the influence of the prior, but the same trends in cutscore estimates were observed in this condition as they were in conditions with smaller sample sizes. The CDT model presented here incorporates a very non-informative prior, and so specifying priors to ensure that results fall within a more "reasonable" range of values—say, -3 and +3—may improve the cutscore recovery results while still being theoretically defensible.

If the CDT method were to be used in an operational setting, it is plausible that the cutscore recommendations would still be calculated in the traditional way, with CDT used only for calculation of the error bands to be used by the policymaking body. This sort of hybrid approach to calculating standard-setting outcomes may be more amenable to stakeholders and psychometricians alike, especially if there is a desire to maintain consistency with previous standard-setting results.

Error variance estimates. The estimates of error variance were more consistent between methods than the estimates of cutscores, with the CDT estimates being larger than the traditional estimates for all three cutscores, and were often several times larger. For all conditions, it was clear that the variance of bias had a greater impact on the CDT error variance estimates than on the traditional estimates. The data-generating bias values—which reflect the true amount of between-group variance in panelists' intended cutscores—are modeled explicitly by CDT but are not directly accounted for in the traditional SEM formula. The CDT results were therefore expected to be more influenced by the level of bias, but it was also clear that the presence of both group-level bias and within-group inconsistency nearly off-set each other in many cases using the traditional method. The slopes of the lines in Figures 25 through 48 for the traditional method become close to zero as total panelist sample size increases, suggesting that a large value for group-level variance may be "absorbed" by collapsing the estimates across groups. By incorporating the multi-level design of a typical standard-setting study, the CDT method attempts to partition the variance into separate components, which is intended to produce a more accurate representation of the total amount of uncertainty about the cutscore. In practice, it is possible that some panelists—especially, perhaps, those with more extreme ICS values—will yield to the group consensus and adjust their bookmark placements up or down, ultimately resulting in panelists within a group looking more similar than panelists across groups. Although standard-setting studies try to mitigate this possibility by allowing for full sample discussion before the final round of standard-setting study, there may be more movement of individual bookmark placements between rounds 1 and 2 (i.e., from panelists working individually to working and discussing as part of their small groups) than there is between rounds 2 and 3.

The CDT error variance estimates are generally larger than those of the traditional method, but both methods underestimated the true error variance. The fact that neither method was able to recover the true amount of variability in the generating distributions was not altogether surprising. One main difference between the Bookmark method and other test-centered standard-setting methods is that it produces a much smaller amount of data. In an Angoff standard-setting study, for example, there is one data point per item, per panelist, per cutscore. Using the Study 1 design, that amounts to 300 data points per panelist, a number that multiplies again if multiple rounds are used. By contrast, the Bookmark panelists simulated here produced only 3 data points each (one per cutscore). Therefore, whatever uncertainty panelists may have about the difficulty of individual items is masked by requiring only one estimate. Although this is generally viewed as a strength of the Bookmark method because it can be less time-consuming for panelists (and for analysts who record the data), it is also true that we have far less information about panelists' level of certainty in their judgments. With multiple cutscores—which are currently mandated for the large-scale summative assessment systems in the U.S.—and the fact that those cutscores are forced both to be ordered and to fall within the boundaries of the ordered item booklet, there is simply less variability that can be recovered in the typical implementation of a Bookmark study. It is possible that panelists vary a great deal in their intended cutscores, but much of that variability will be lost if they need to translate those ICSs to a point within a range of, say, 20 or 30 items in an OIB.

Sample size. The sample sizes used in the simulation study were chosen to fit within the ranges of typical recommendations in standard-setting literature (e.g., Berk, 1996; Jaeger, 1991; Raymond & Reid, 2001). Those sources suggest anywhere from 5 to more than 80 panelists, with the mode seeming to be somewhere between 10 and 20. The smallest sample size in this study, one group of 8 panelists, was intended to be small but still defensible for a large-scale testing program. (The operational standard-setting study from which the Study 2 data were gathered stipulated that the study could not proceed if a grade had fewer than seven panelists.) Working upward from 8, groups of 16 were convenient as they allowed direct comparisons based solely on the groupings (e.g., two groups of 8 versus one group of 16). The largest sample size, four groups of 16 panelists, is likely to be larger than what would be practical for most operational standard-setting studies. Most standard-setting panels for large-scale assessments seem to comprise around 15 panelists who work in two or three small groups. Although multiple groups of 16 panelists are unlikely to be used in practice, fully crossing the numbers of groups and panelists per group allows for examining how much (if any) accuracy is gained by either method with very large samples, as well as making comparisons among different groupings of the same total number of panelists. In any case, the largest sample here (64 panelists) is still within the largest recommendation in the literature (87 panelists; Jaeger, 1991).

As expected, the width of the 95% CIs decreased for the for both methods as the number of panelists increased. As is evident from the figures in the Appendix, though, smaller standard errors do not necessarily mean better confidence intervals. For example, with three groups of 8 panelists and the true variance of bias and consistency both equal to 0.2 (Figures A.57 and A.58), the 95% CIs for Cutscores 2 and 3 included the true cutscore more often than the same condition but with four groups of 8 panelists (Figures A.89 and A.90). Increasing the sample size does not lead to more accurate CIs if the cutscore estimate itself is too high or too low. Again, this was true for both the CDT and traditional methods, and the better coverage of the CDT credible intervals (compared with the traditional CIs) in either condition is a product of their being wider to begin with.

However, a purported advantage of the CDT method is that it takes advantage of the small-group structure of a typical Bookmark panel. For the sake of comparison, see Figures A.5, A.6, A.101, and A.102. These figures show the same results for the Traditional and CDT methods, respectively, for zero variance conditions with two groups of 8 panelists (A.5. and A.6), and one group of 16 panelists (A.101 and A.102). Note that figures A.5 and A.101 are identical—because the data were generated using seeded random values, the bookmark placements across these two conditions were the same for the 16 total panelists. The traditional method ignores the groupings of panelists, and so there is no improvement (or worsening) in estimation by using grouped data. The CDT method, however, produced larger CIs when accounting for two small groups of panelists rather than one overall sample. Accounting for the nested design does not “improve” cutscore estimates in the sense that the mean values change much, but it does produce improved CIs in the sense that they are more likely to include the true cutscore. In other words, group differences may contribute to the overall uncertainty about a cutscore recommendation, and yet are ignored by the traditional method. When the panelists within a group are more similar than the total sample in their cutscore recommendations, it suggests a group effect is present. When this is the case, we should have less confidence that panelists are relying on their own judgment, which in turn gives us less confidence in the cutscore recommendation.

Study 2

The real data study was included in this research mainly as a “proof of concept” and to see how the CDT method would perform compared with the traditional method in a real data setting. Although results were calculated from only a single grade and subject, some important conclusions may be drawn.

First, the real data results reiterated that the CDT method produces wider CIs than the traditional method, as the CDT method produced larger cutscores for all three cutscores. However, while the simulated results tended to be fairly consistent across the three cutscores within a condition, the real data results were not. In fact, the CIs for Cutscores 1 and 3 were quite similar between the two methods (with those of the CDT method being slightly wider), but the Cutscore 2 CI from the CDT method was about twice as wide as that of the traditional method (0.77 versus 0.39). As a whole, these results are encouraging, as they reveal that the CDT method does not simply produce estimates that are much larger than the traditional methods, with the size of the CIs instead being driven by the data themselves.

It is also evident from the real data results that the CDT credible intervals are not symmetric about the cutscore estimate. This is an important point, and one that is not readily apparent from the simulation study results. None of the three CDT cutscore estimates in Study 2 is centered between the upper and lower bounds of the CIs, while the traditional estimates are—by definition—centered exactly. The fact that the CDT credible intervals are not required to be symmetric is a potentially strong advantage of the CDT method in an operational setting. For example, it makes little sense for a CI to include scores that are not possible on a test (e.g., a theta value above the one associated with a perfect score), but the traditional method CIs could include such scores. By using a model-based approach, we can calculate CIs that are more practically useful. The CDT credible intervals here tend to be fairly symmetric, perhaps due to the fact that a multivariate normal prior was used. Less restrictive priors may help this advantage to be more pronounced, and demonstrate the need for further study.

It is interesting that the cutscore associated with the largest CDT credible interval (Cutscore 2) had the smallest observed variance among the panelists. The cutscore estimates from each group for Cutscore 2 were closer together than those for the other two cutscores, and so it seems reasonable to expect the error variance to be smaller for this cutscore. Indeed, Groups 1 and 2 were about a scale unit apart for both Cutscores 1 and 3, but the difference was less than half that for Cutscore 2. The group-level variance was fairly consistent for Group 1 across all three cutscores. However, the variance in Group 2 was about twice as large for Cutscores 1 and 3 than for Cutscore 2. It may be that the wide dispersion of bookmark placements for the lowest and highest cutscores effectually “forced” the middle cutscore into a tighter window. Ultimately, more research is needed before claiming the CDT method to be more accurate in this case. But it may be reasonable to expect a panel to be “consistently uncertain” about a series of cutscores, rather than much more certain about one cutscore than the others. Such results could also indicate that the performance levels above and below Cutscore 2 were more clearly defined than the ones separated by the other two cutscores—a plausible notion since this was the “high-stakes” cutscore. Although the CDT method treats multiple cutscores as fitting the same model, in practice panelists may not view multiple cutscores to be so connected, particularly when those cutscores carry different stakes.

Limitations

In addition to those already discussed, some of the limitations of this research are listed and briefly discussed in this section. Although the CDT method overall was promising, the results here suggest that the Bayesian model proposed here—particularly the multivariate normal prior assigned to panelists' intended cutscores—may need refinement before the CDT methodology is ready to be applied in an operational setting.

Aside from the Bayesian estimation itself, the main difference between CDT and the traditional method is that CDT attempts to model group-level variance as well as within-group variance. The traditional method considers only the consistency of the full sample of judges, with no regard paid to group-level variance in the calculation of final cutscore recommendations. In theory, however, the two models should be equivalent in the conditions where $N_G = 1$, which was not the case in these results. This suggests that the Bayesian estimation method itself may need to be improved. The multivariate normal prior placed on panelists' estimates may be overly influential with three cutscores being calculated, and a study which simplifies the data here to a single cutscore would be particularly helpful.

Another factor that made the error variance estimates for both the CDT and traditional methods appear to underestimate the true error variance was that the amount of variance in the sample was generally smaller than what the data-generating model would imply. For every condition within the simulation study, the true amount of cutscore error variance across replications was expected to be the sum of the generating variances of consistency (i.e., within-group variance) and bias (i.e., between-group variance). For example, in the conditions where σ_c^2 and σ_b^2 were both equal to 0.2, the expected variability across all 100 replications was expected to be 0.4. However, the expected variance of the panelists' ICSs was rarely reflective of the variance of their simulated bookmark placements, particularly for the Cutscores 2 and 3. The ICS values for these cutscores were generated using truncated normal distributions, which effectively reduced the total variance across replications. Likewise, the process of rejecting "illegal" bias values (i.e., bias values that would not have allowed ordered cutscores) also reduced the total true variance in the samples. Nevertheless, the "actual" true variance in each condition was still much larger than the CDT estimates (except when the true variances of bias and consistency were 0). Refinements to the CDT method should focus on recovering the error variance components more completely.

Conclusions & Implications

The simulated and real data studies described here provide some support for a model-based method of describing the uncertainty involved in the Bookmark standard-setting process. Building upon Bayesian methods of estimating the error variance of Angoff cutscores (Skorupski et al., 2015), Cutscore Distribution Theory (CDT) for the Bookmark method proposes a psychometrics of Bookmark cutscores that attempts to model both within-group and between-group variance in the bookmark process in order to estimate more accurate confidence intervals around cutscore estimates. Although much of the extant standard-setting literature focuses on calculation of the cutscores themselves, this line of research is more focused on how the uncertainty around them is calculated. The error bands around the cutscore recommendations carry a great deal of practical importance, as policymakers often adjust a panel's recommendation upward or downward (e.g., after reviewing impact data, or to help reconcile discrepancies from year to year or even across grade levels). The current research, therefore, is not intended to reduce the error variance around cutscores, but simply to estimate it more accurately.

The results of these studies also suggest some broad recommendations for implementing a Bookmark standard-setting study, and provide some groundwork for additional research. The CDT results showed that dividing panelists into two groups produces larger estimates of error variance than with one, three, or four groups. If two groups are very far apart in their recommended cutscore, it is not readily apparent which group is closer to the "true" cutscore. Dividing panelists into three or four groups may allow for a better estimation of the group-level bias, and therefore a better estimate of the true cutscore. (Using more groups with fewer panelists per group may also mitigate the possibility of a few dominant panelists overly influencing a panel.) Further research into panelist structure—particularly with real data—appears warranted. Additional studies should also examine the effect of varying the placement of the true cutscore on the CDT cutscore estimates. Such research might simplify the design used here by looking only at a single cutscore. Restricting the study to one cutscore would eliminate the need for truncation in the data-generation process, as well as the use of a multivariate normal prior.

The most promising aspect of the CDT methodology was its ability to generate 95% credible intervals that included the true cutscore more often, and often much more often, than the traditional method. Although part of that is certainly a byproduct of simply producing larger CIs, the real data results indicate that the CDT method is capable of producing CIs similar to those of the traditional method. Ultimately, the goal of the proposed methodology is to provide a psychometrics of cutscores in order to model standard-setting results. For the Bookmark method, model-based cutscores should be preferable to averaging the distance between item difficulty values, as is typically done under the traditional method. The results described here suggest that modeling cutscores in this way produces more accurate cutscore recommendations, in the sense that they are more likely to reflect the opinions of a population of potential standard-setting panelists. Again, the goal of this research is not so much to improve cutscore estimation as it is to give policymakers a more accurate representation of the uncertainty involved in the standard-setting process.

References

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

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Beretvas, S. N. (2004). Comparison of Bookmark difficulty locations under different item response models. *Applied Psychological Measurement, 28*(1), 25–47.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45,* 4–9.

Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education*, *9*(3), 215-225.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison–Wesley.

Box, G. E., & Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. New York: Wiley Classics. (Original work published 1973)

Brandon, P. R., (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17* (1), 59–88.

Brennan, R. L. (1995). Standard setting from the perspective of generalizability theory. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessment Governing Board (NAGB) and the National Center for*

Education Statistics (NCES) (pp. 269–287). Washington, DC: U.S. Government Printing Office.

Brennan, R. L. & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using Generalizability Theory. *Applied Psychological Measurement, 4*(2), 219–240.

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement*, *39*(3), 253-263.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. Journal of Educational Measurement, 27(2), 145-163.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, *30*(2), 93–106.

Cizek, G. J. (1996). Standard-Setting Guidelines. *Educational Measurement: Issues and Practice*, *15*(1), 13–21.

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). New York, NY: Routledge.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.

Clauser, B. E., Swanson, D. B., & Harik, P. (2002). Multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an

Angoff-style standard-setting procedure. *Journal of Educational Measurement, 39*(4), 269–290.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Egan, K. L, Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York, NY: Routledge.

Elementary and Secondary Education Act of 1965, U.S.C. §§ 1111 (2016).

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor and Department of Justice. (1978, August 25). Uniform guidelines on employee selection procedures. *Federal Register, 43,* 38290–38315.

Engelhard, G. & Anderson, D. W. (1998). A binomial trials model for examining the ratings of standard-setting judges. *Applied Measurement in Education, 11*(3), 209–230.

Engelhard, G. (2009). Evaluating the judgments of standard-setting panelists using Rasch measurement theory. In E. V. Smith Jr. & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 312-346). Maple Grove, MN: JAM Press.

Engelhard, G. (2011). Evaluating the Bookmark judgments of standard-setting panelists. *Educational and Psychological Measurement*, 71(6), 909–924.

Fehrmann, M. L., Woehr, D. J., & Arthur, W., Jr. (1991). The Angoff cutoff score method: The impact of frame-of-reference rater training. Educational and Psychological Measurement, 51, 857–872.

Ferdous, A. A. & Plake, B. S. (2005). Understanding the factors that influence decisions of panelists in a standard-setting study. *Applied Measurement in Education, 18*(3), 257–267.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd. Retrieved September 28, 2017 from http://krishikosh.egranth.ac.in/bitstream/1/ 2048218/1/0039_2689A.pdf.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004).

Giraud, G. & Impara, J. C. (2005). Making the cut: The cut score setting process in a public school district. *Applied Measurement in Education, 18*(3), 289–312.

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education*, *18*(3), 223–232.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, *15*(4), 237-261.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237-288.

Glaser, R. & Nitko, A. J. (1971). Measurement in learning and instruction. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 625–670). Washington, DC: American Council on Education.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, *15*(4), 277–290.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, (pp. 89–115). Mahwah, NJ: Erlbaum.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*(4), 355–366.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*(1), 1–47.

Hein, S. F. & Skaggs, G. E. (2009). A qualitative investigation of panelists' experiences of standard setting using two variations of the Bookmark method. *Applied Measurement in Education, 22*(3), 207–228.

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19–20.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35,* 69–81.

Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judges. *Applied Measurement in Education, 1*(1), 17–31.

Jaeger, R. M. (1989). Certification of student competence. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). Washington, DC: American Council on Education.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice*, *10*(2), 3-14.

Kane, M. T. (1987). On the use of IRT models with judgmental standard setting procedures. *Journal of Educational Measurement*, *24*(4), 333–345.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment*, *5*(3), 129-145.

Kane, M. & Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. *Applied Psychological Measurement*, *8*(1), 107–115.

Karantonis, A. & Sireci, S. G. (2006). The Bookmark standard-settting method: A literature review. *Educational Measurement: Issues and Practice*, *25*(1), 4–12.

Lee, G. & Lewis, D. M. (2008). A Generalizability Theory approach to standard error estimates for Bookmark standard settings. *Educational and Psychological Measurement*, *68*(4), 603–620.

Lewis, D. M., Mitzel, H. C., Mercado, R.L., & Schulz, E.M. (2012). The Bookmark standard setting procedure. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225–253). New York, NY: Routledge.

Linn, R. L. (2003, September 1). Performance standards: Utility for different uses of assessments. Education Policy Analysis Archives, 11(31). Retrieved January 15, 2016, from http://epaa.asu.edu/epaa/v11n31/.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Loomis, S. C. (2012). Selecting and training standard setting participants. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 107–134). New York, NY: Routledge.

Lunn, D., Spiegelhalter, D., Thomas, A., Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*, 28, 3049-3067.

McGinty, D. (2005). Illuminating the "black box" of standard setting: An exploratory qualitative study. *Applied Measurement in Education, 18*(3), 269–287.

Mehrens, W. A. & Cizek, G. J. (2012). Standard setting for decision making: Classifications, consequences, and the common good. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 33–46). New York, NY: Routledge.

Mehrens, W. A. & Popham, W. J. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education, 5*(3), 265–283.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research, 46*(1), 133–158.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.

National Research Council. (2005). *Measuring literacy: Performance levels for adults*. Washington, DC: The National Academies Press. doi:10.17226/11267

Phillips, S. E. (2012). Legal issues for standard setting in K–12 educational contexts. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 535–569). New York, NY: Routledge.

Plake, B. S. (2008). Standard setters: Stand up and take a stand! *Educational Measurement: Issues and Practice*, *27,* 3–9.

Plake, B. S. & Impara, J. C. (2001). Ability of panelists to estimate item performance for a target group of candidates: An issue in judgmental standard setting. *Educational Assessment, 7*(2), 87–97.

Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement*, *15*(4), 297–300.

Poggio, J. P. (1984, April). Practical considerations when setting test standards: A look at the process used in Kansas. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3–19.

Norcini, J. J., Shea, J. A., & Kanya, D. T., (1988). The effects of various factors on standard setting. *Journal of Educational Measurement, 25,* 57–65.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, (pp. 119–157). Mahwah, NJ: Erlbaum.

Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice, 25*(2), 4-18.

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice*, *10*(2), 11-14.

Schulz, E. M. (2006). Commentary: A response to Reckase's conceptual framework and examples for evaluating standard setting methods. *Educational Measurement: Issues and Practice*, *25*(3), 4–13.

Skorupski, W. P. (2012). Understanding the cognitive processes of standard setting panelists. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 135–147). New York, NY: Routledge.

Skorupski, W. P. & Fitzpatrick, J. (2014, April). *Sampling distributions of cutscores based on panelist accuracy and consistency*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.

Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education*, *18*(3), 233-256.

Skorupski, W. P., Zhao, Y., Fitzpatrick, J., & Chen, F. (2015, April). *Cutscore Distribution Theory (CDT): A comparison with G-Theory*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, 12(3), 1-16.

Van de Watering, G. & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review*, *1*(2), 133–147.

Van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19* (4), 295–308.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement*, *40*(3), 231–253.

Williams, N. J., & Schulz, E. M. (2005, April). An investigation of response probability (RP) values used in standard setting. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Yin, P. & Sconing, J. (2008). Estimating standard errors of cut scores for item rating and mapmark procedures: A Generalizability Theory approach. *Educational and Psychological Measurement*, *68*(1), 25–41.

Zieky, M. J. & Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment Tests.* Princeton, NJ: Educational Testing Service.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Appendix

OpenBUGS script for Cutscore Distribution Theory Estimates

```
model {
#Specify likelihood and model
for (i in 1:(N*N.cut*NG*n) ) {
   datamat1[i] ~ dbern(p[R[i], K[i], G[i], J[i]])
   p[R[i], K[i], G[i], J[i]] <-
      (c[i]) + ((1-c[i])/(1+exp(-a[i]*(theta[G[i], R[i], K[i]]-b.RP[i]))))
}
#Prior for panelists is multivariate normal.
for (g in 1:NG) {
   for (i in 1:N) {
      theta[g, i, 1:N.cut] ~ dmnorm(mu.cut[ , g], T[ g, , ])
   }
}

### Hyperparameters for panelist theta values.
### Expected value is from data (mean across panelists).
for (g in 1:NG) {
   for (k in 1:N.cut) {
      mu.cut[k, g] <- CUT[k]
      hyper.mu.g[k,g] <- mean(ind.theta.cut.r[ , k, g])
   }
}

for (k in 1:N.cut) {
     hyper.mu[k] <- mean(hyper.mu.g[k, ])
     CUT[k] ~ dnorm(hyper.mu[k], 0.001)
}

### sig is the estimated V/C matrix.
for (g in 1:NG) {
   for (k in 1:N.cut) {
      for (m in 1:N.cut) {
            sig[g, k, m] <- 2*INV[ g, k, m]
      }
   }
}

### Variance/covariance matrix is inverse-Wishart.
for (g in 1:NG) {
   INV[g, 1:N.cut, 1:N.cut] <- inverse(T[ g, , ])
   T[g, 1:N.cut, 1:N.cut] ~ dwish(rho[ , ], N.cut)
}
}
```

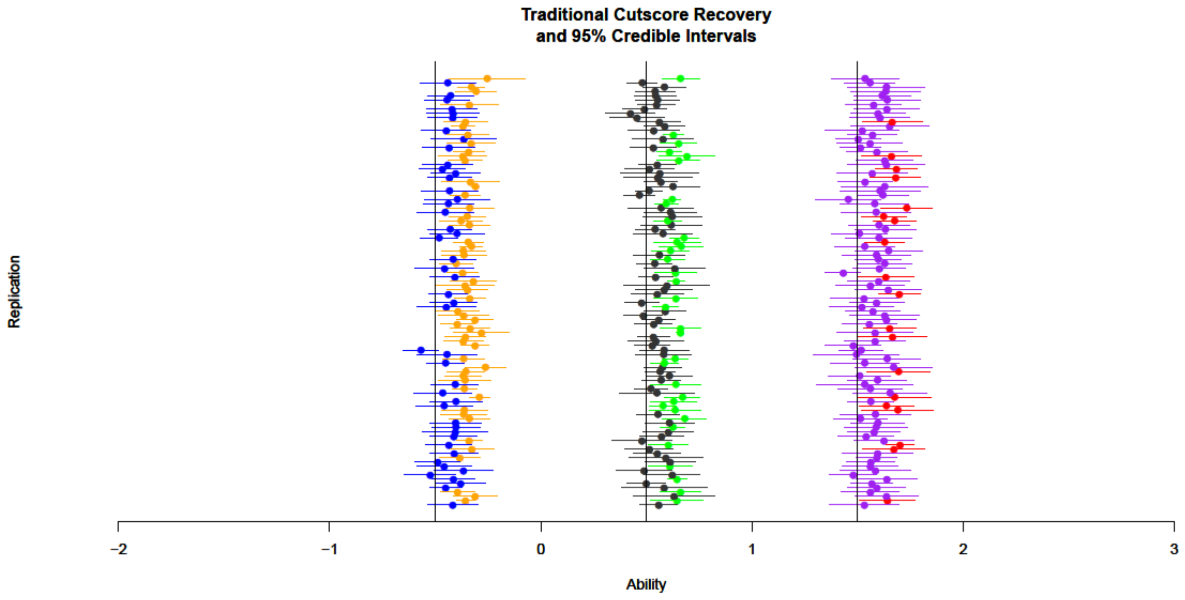

Figure A.1. $N_I = 8$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

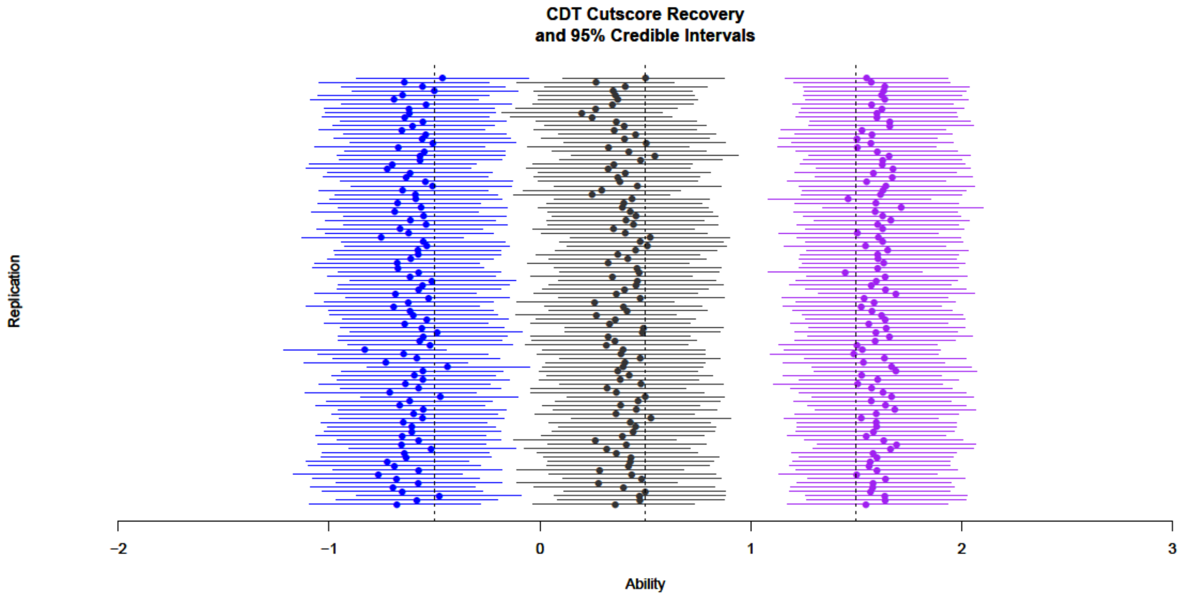

Figure A.2. $N_I = 8$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

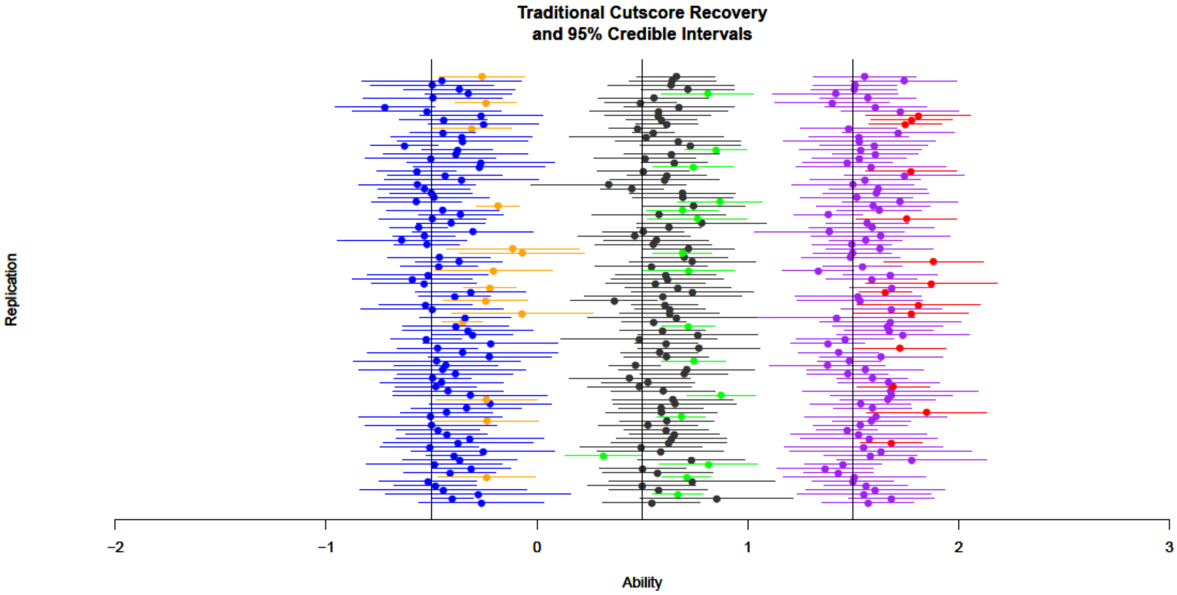

Figure A.3. $N_I = 8$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

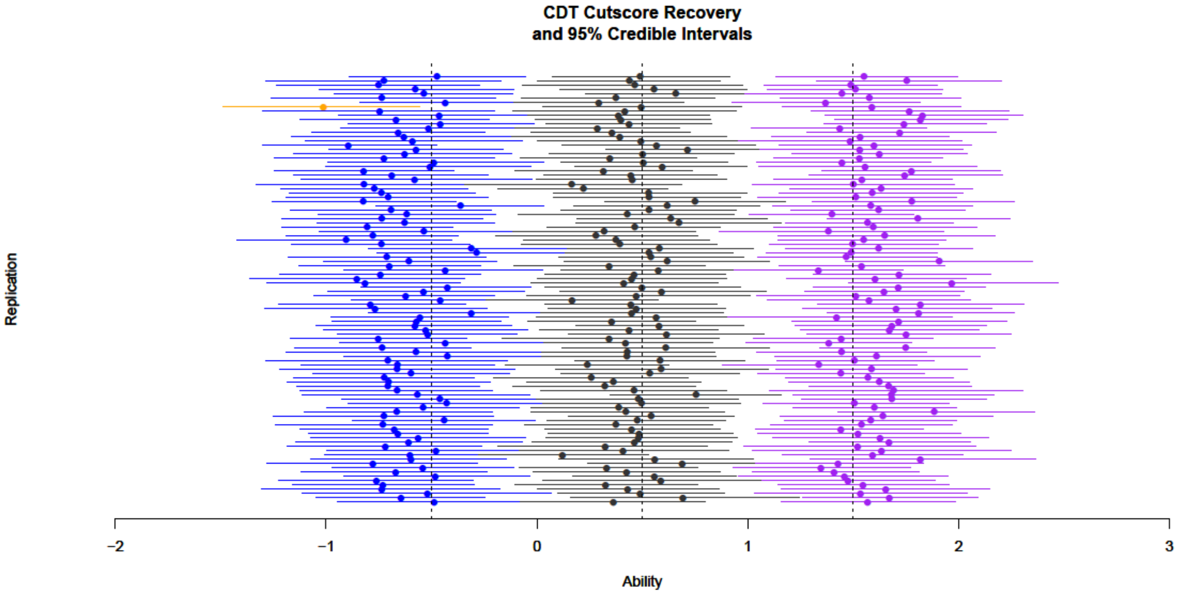

Figure A.4. $N_I = 8$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

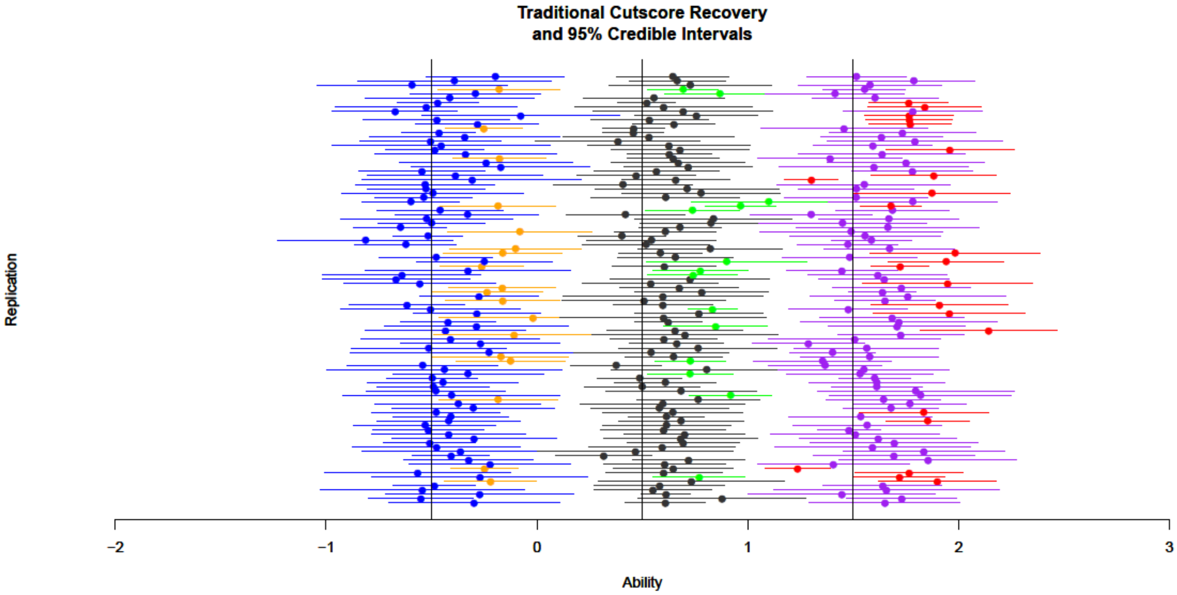

Figure A.1. $N_I = 8$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

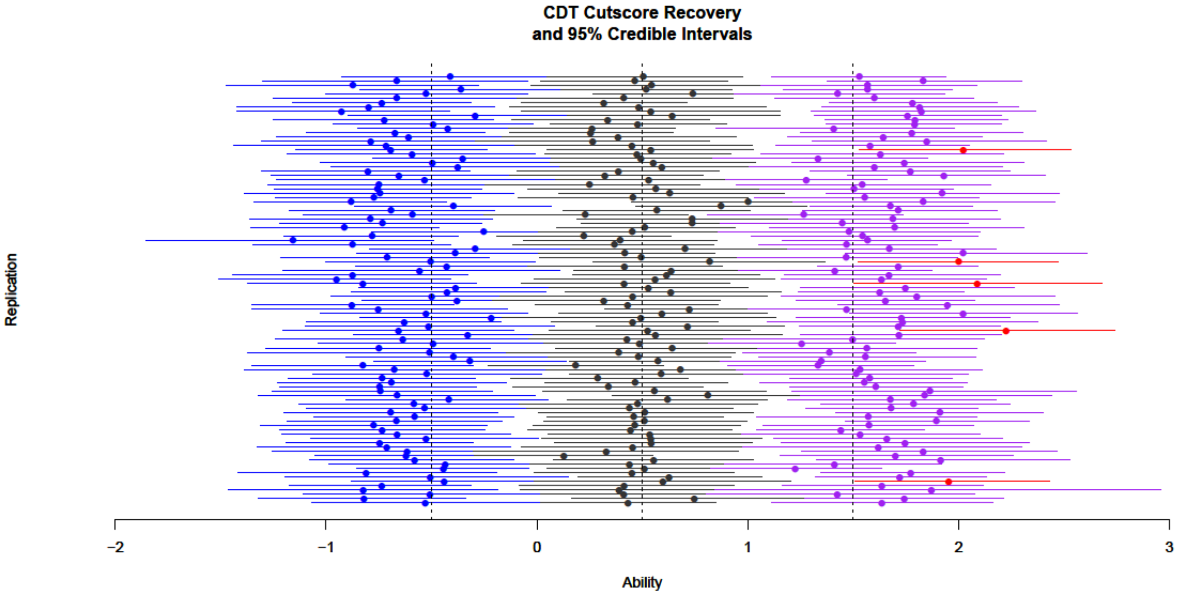

Figure A.2. $N_I = 8$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

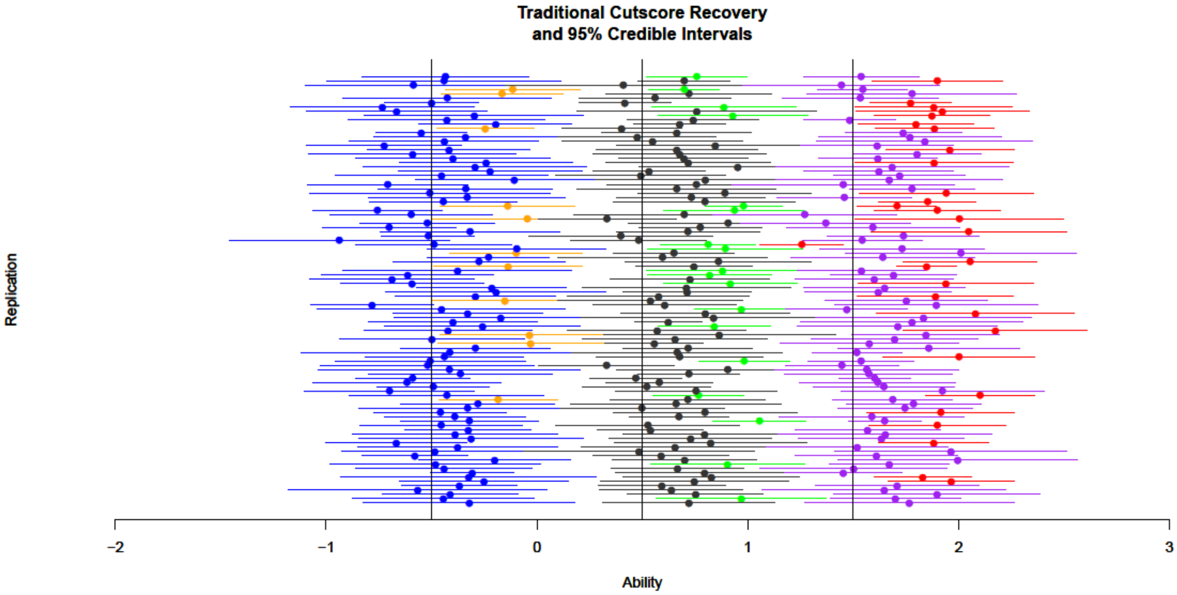

Figure A.3. $N_I = 8$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

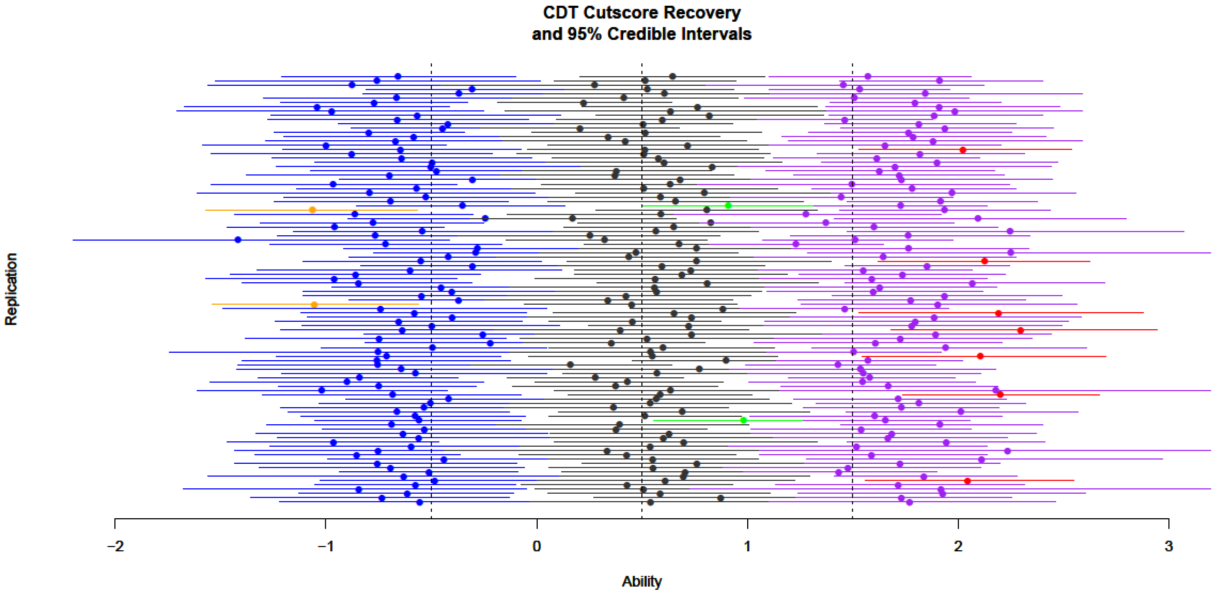

Figure A.4. $N_I = 8$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

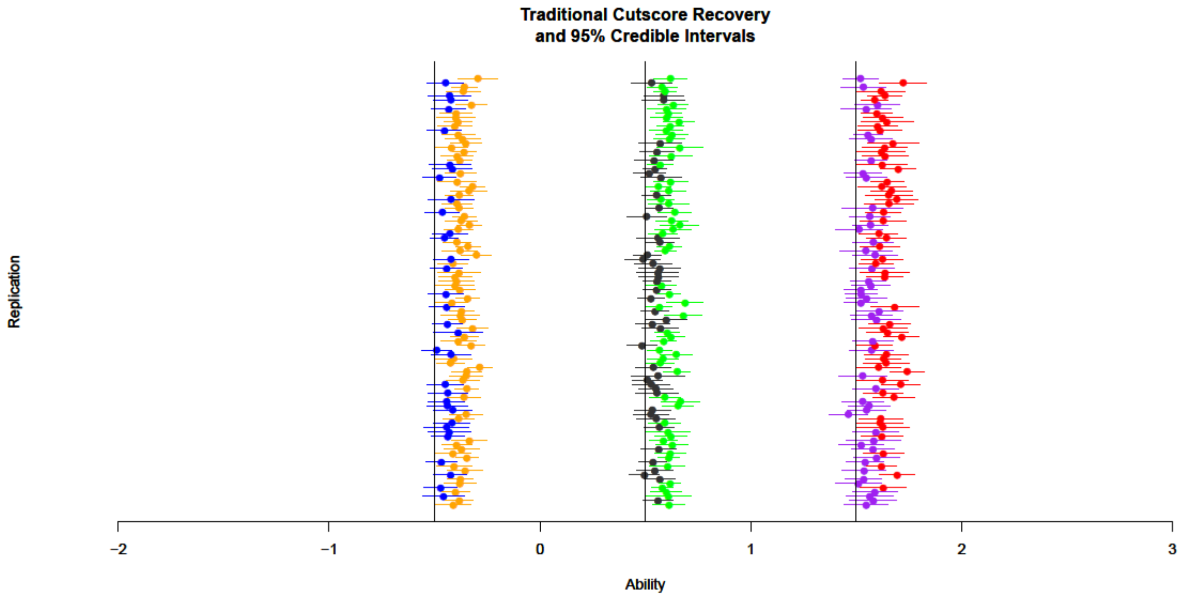

Figure A.5. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

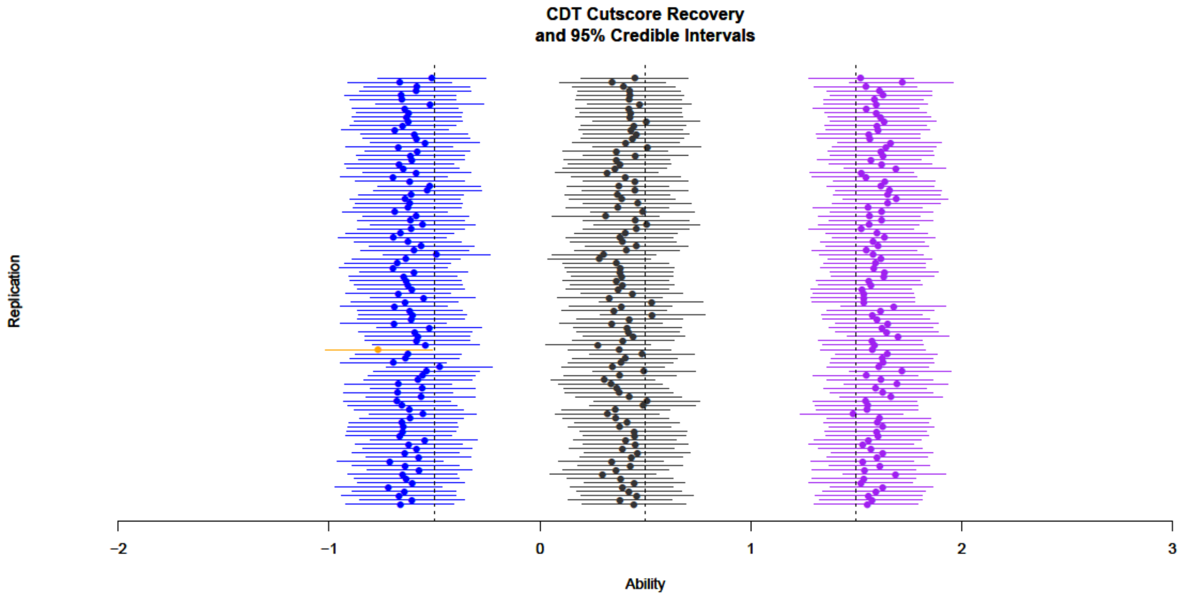

Figure A.6. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

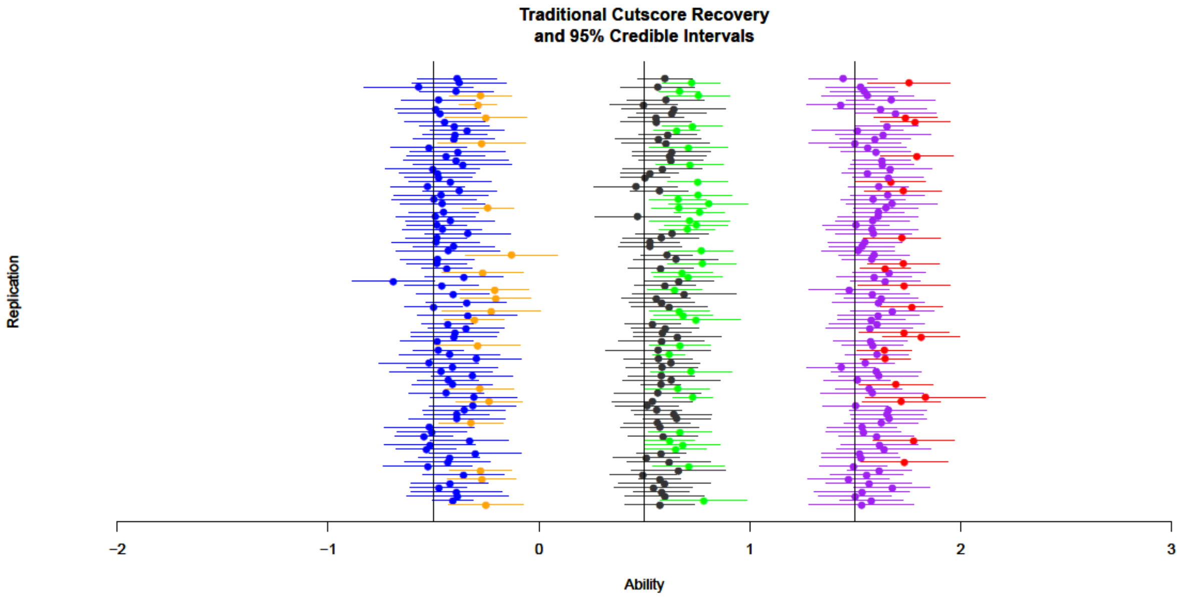

Figure A.7. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

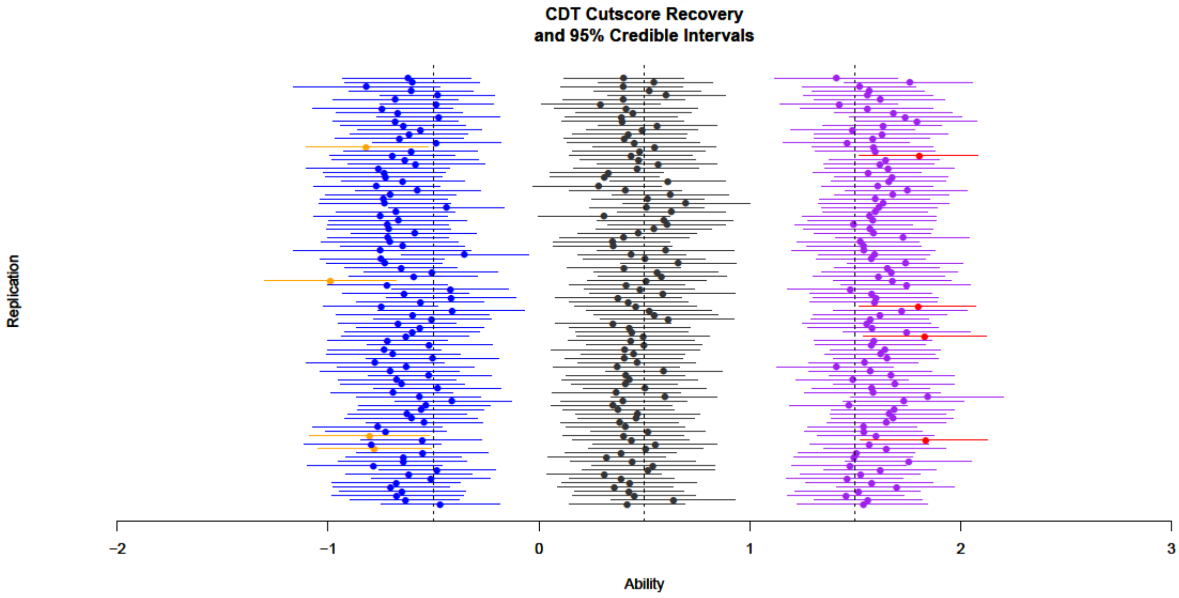

Figure A.8. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

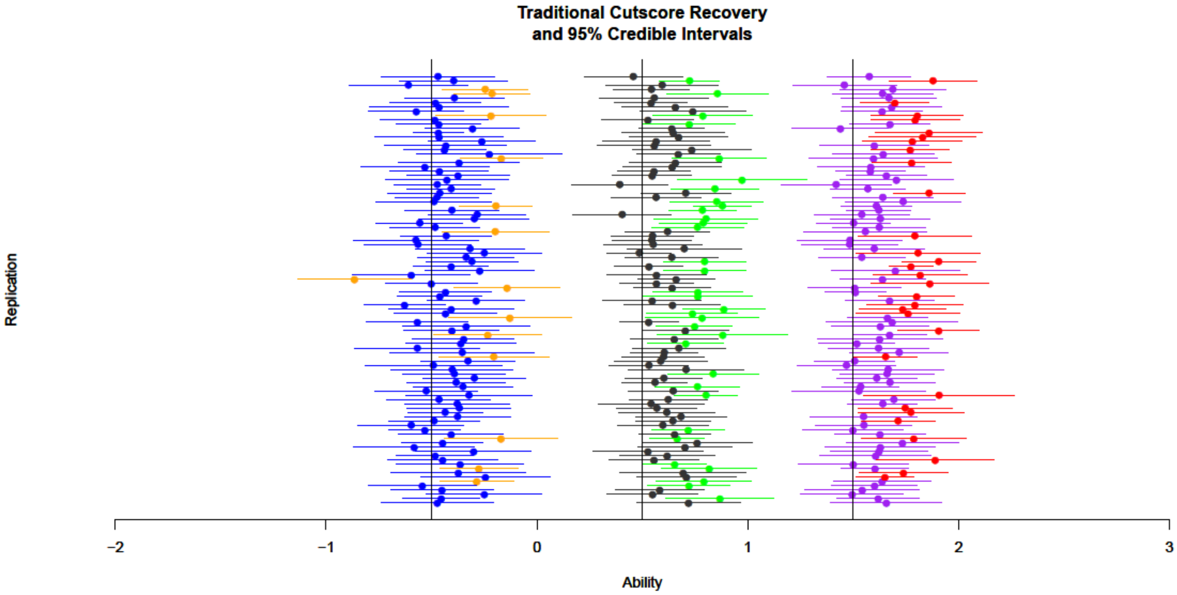

Figure A.9. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

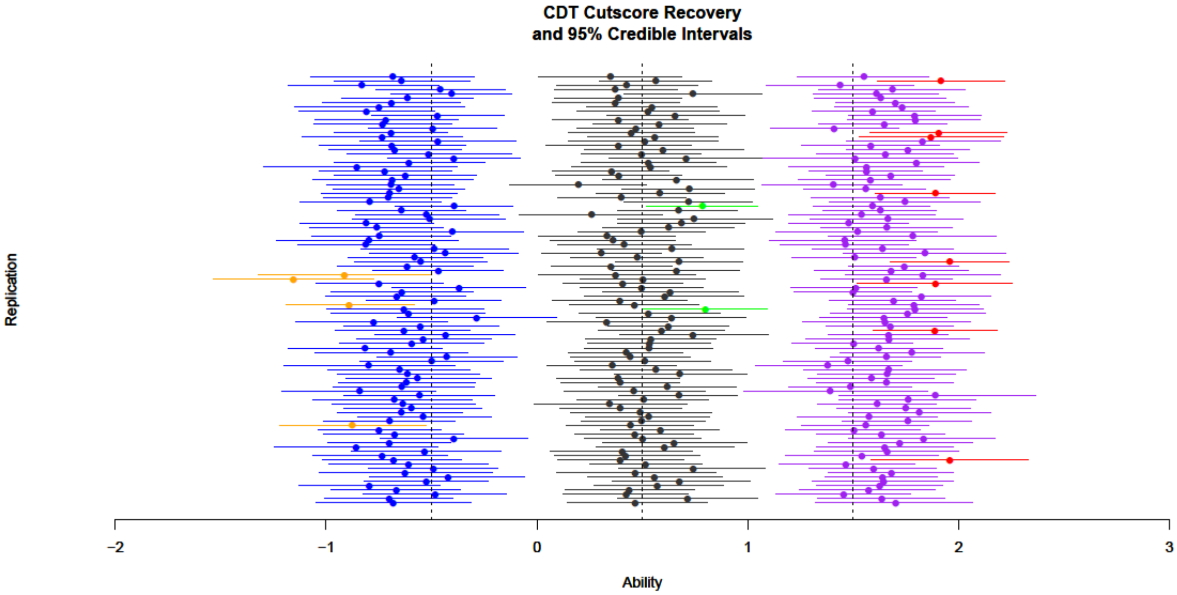

Figure A.10. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

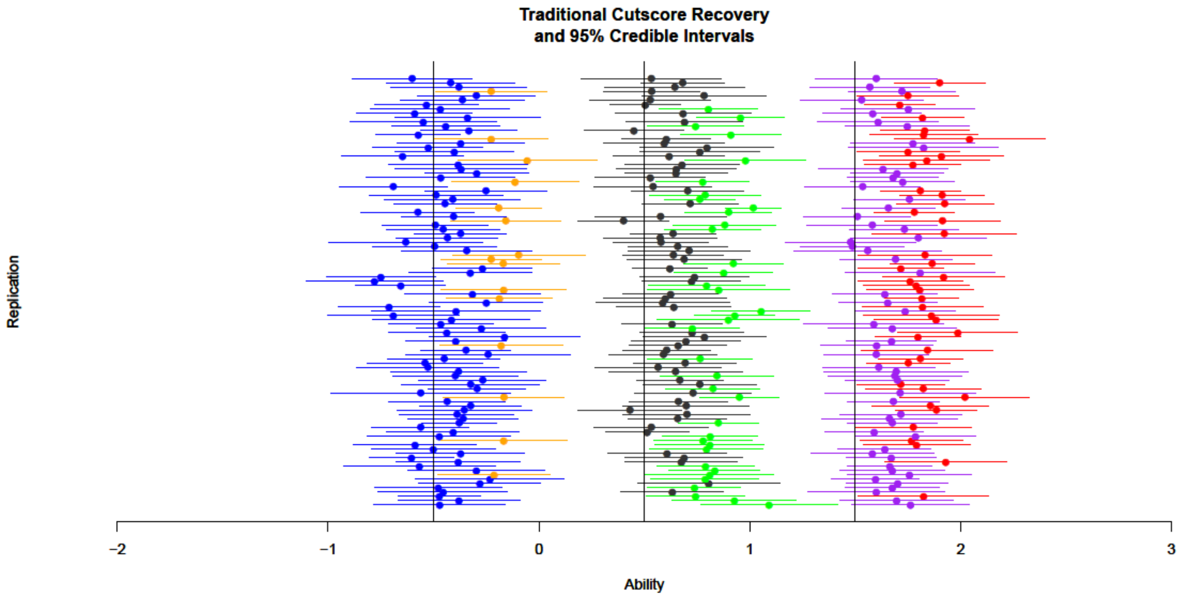

Figure A.11. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

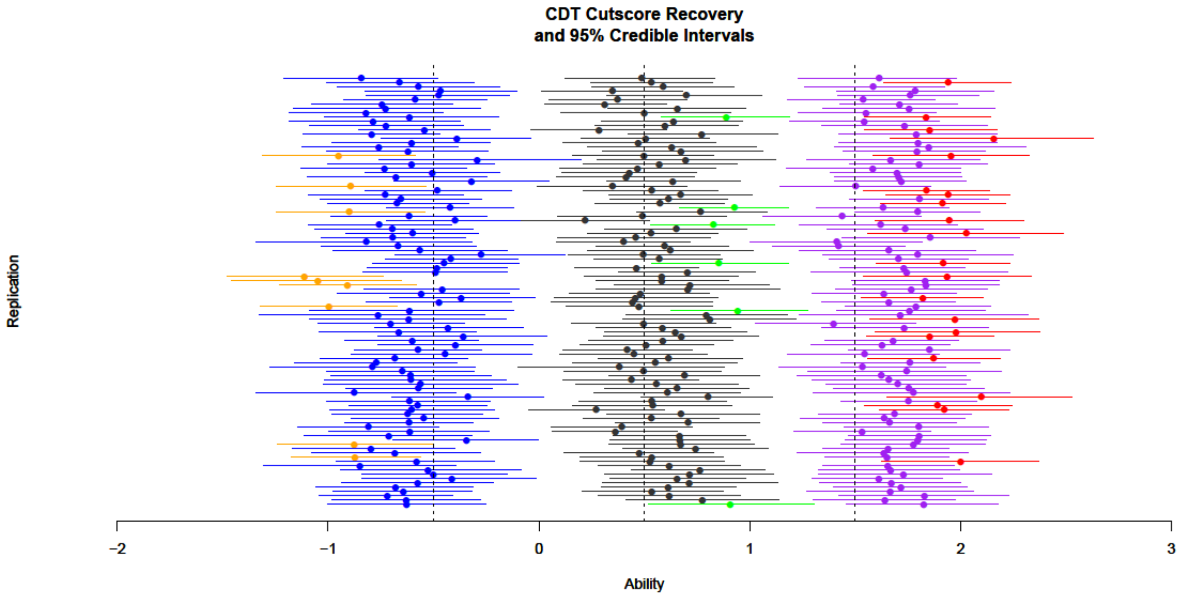

Figure A.12. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

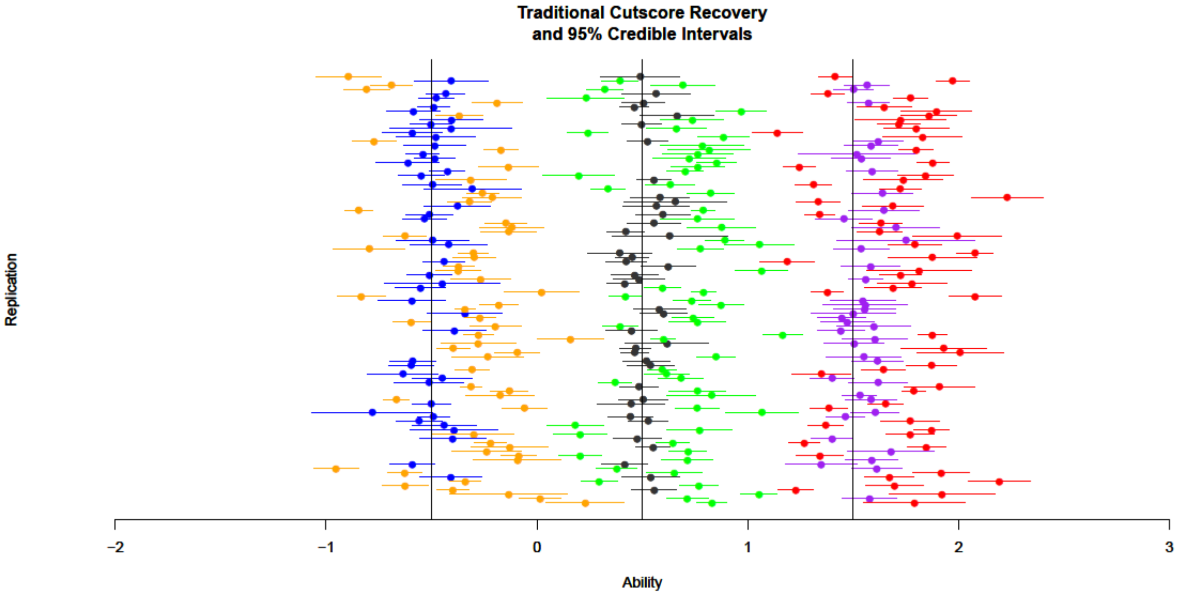

Figure A.13. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

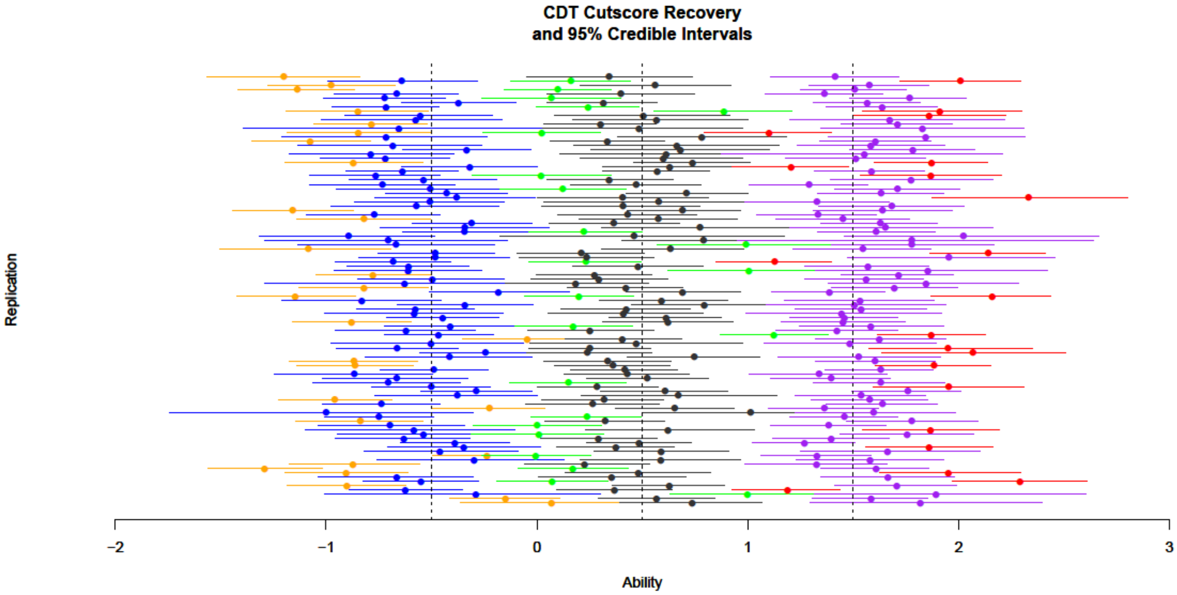

Figure A.14. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

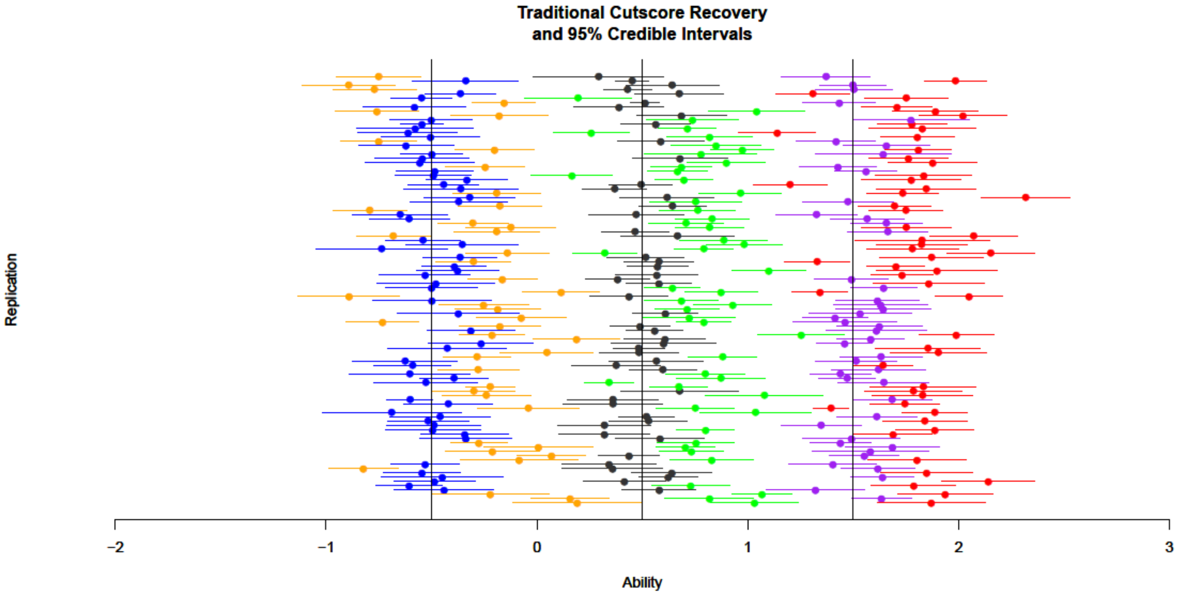

Figure A.15. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

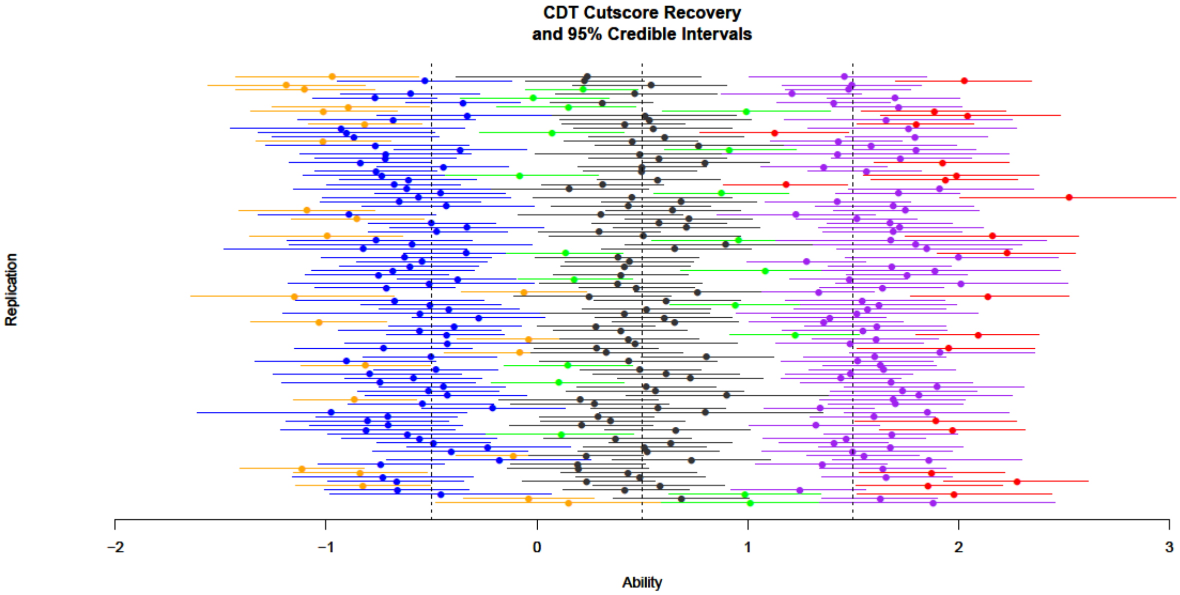

Figure A.16. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

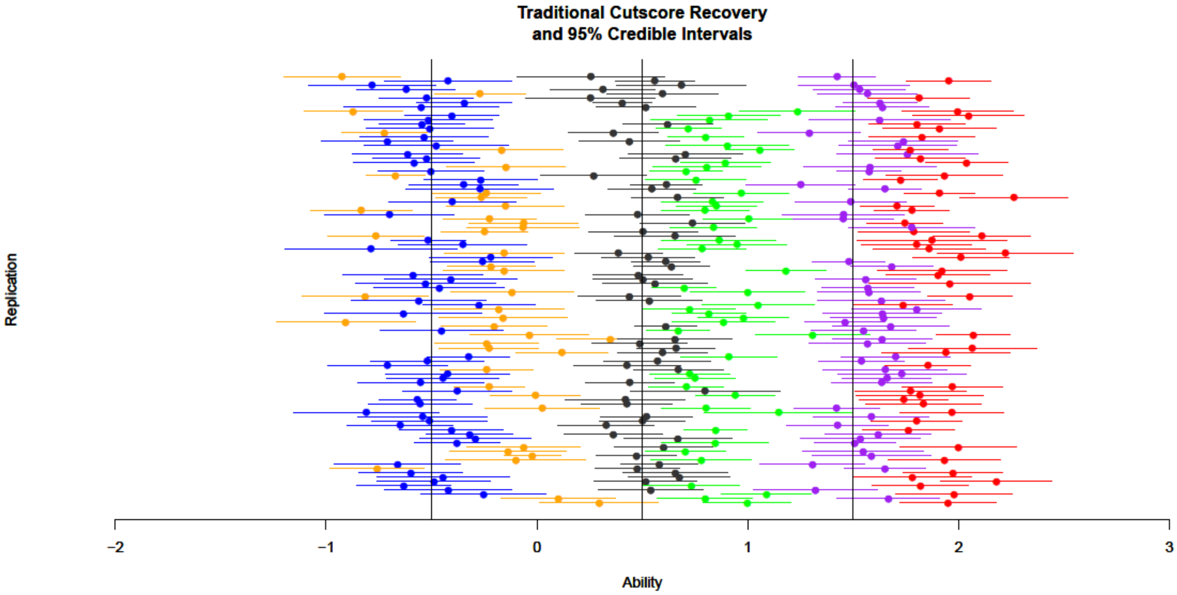

Figure A.17. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

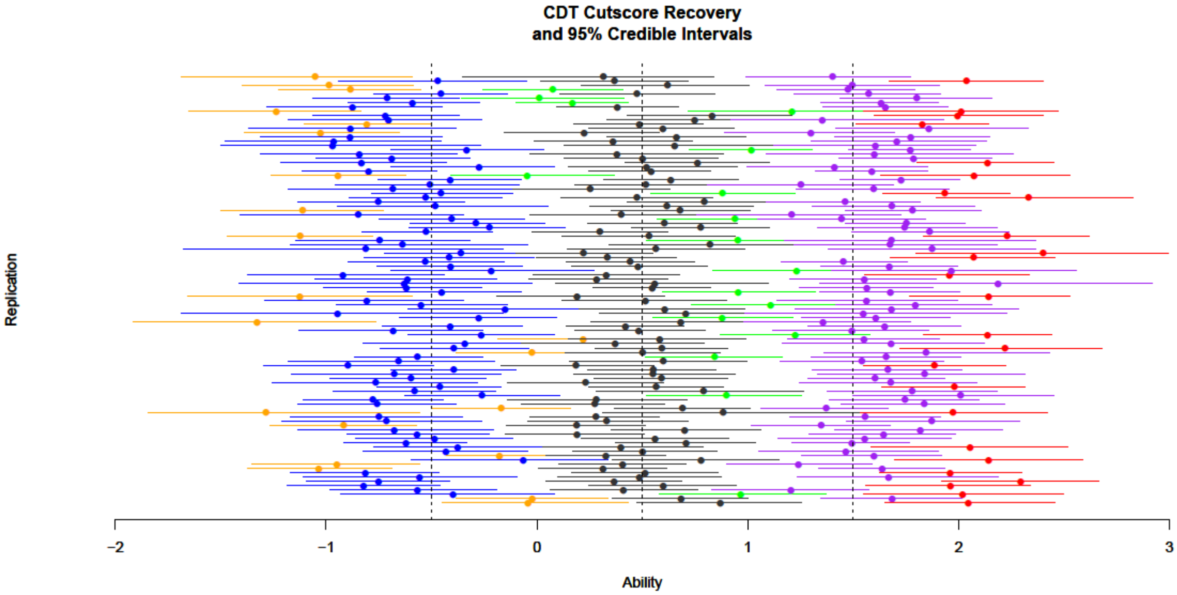

Figure A.18. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

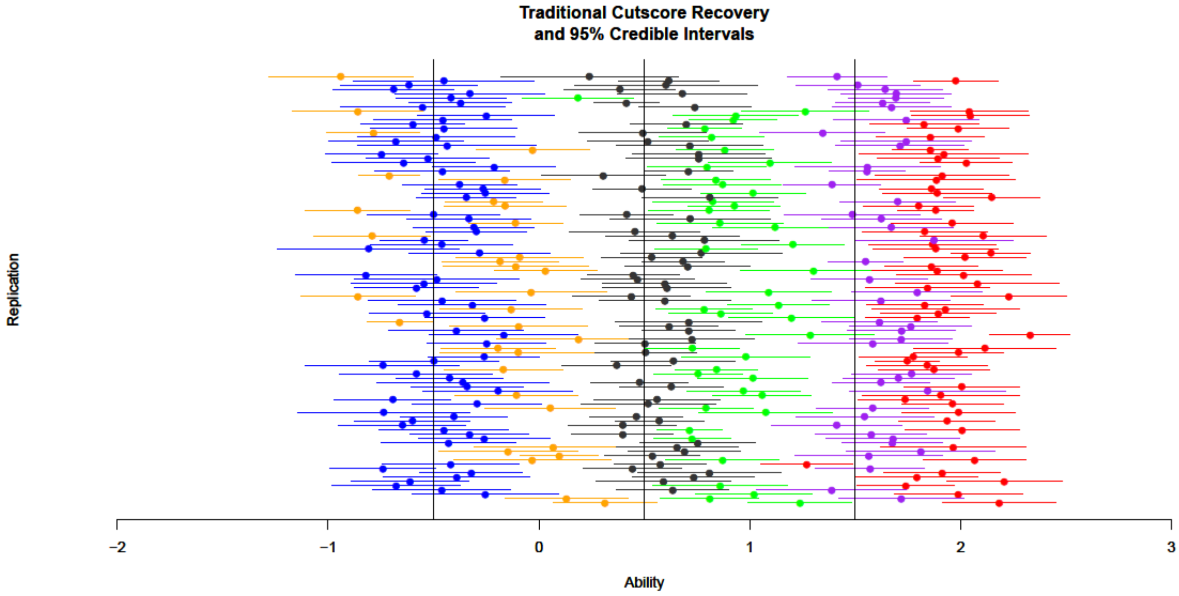

Figure A.19. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

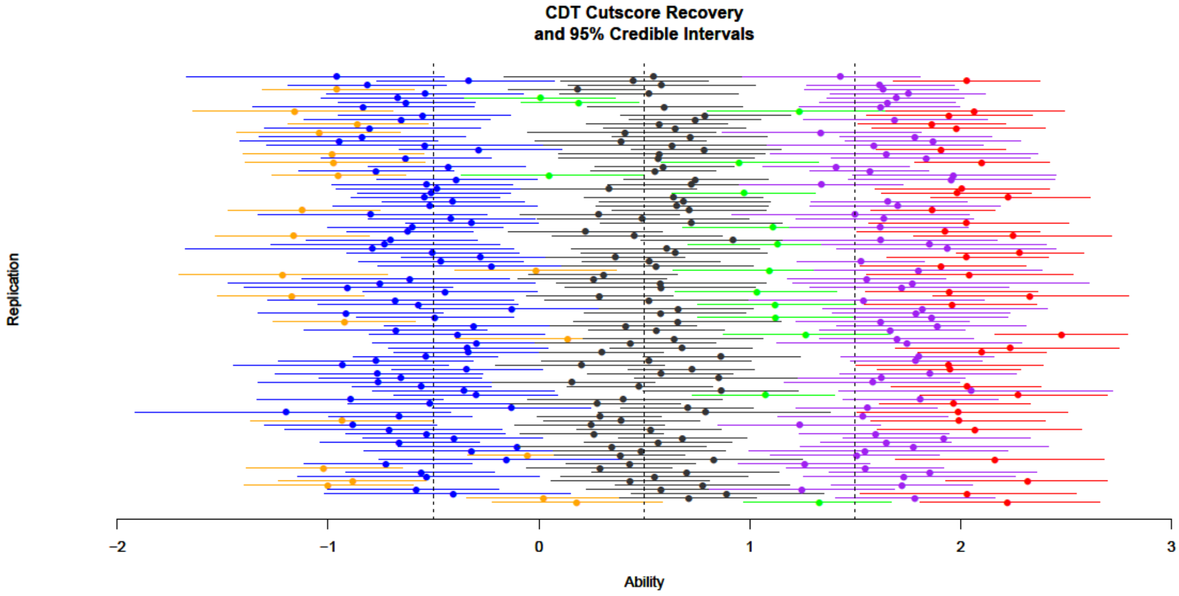

Figure A.20. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

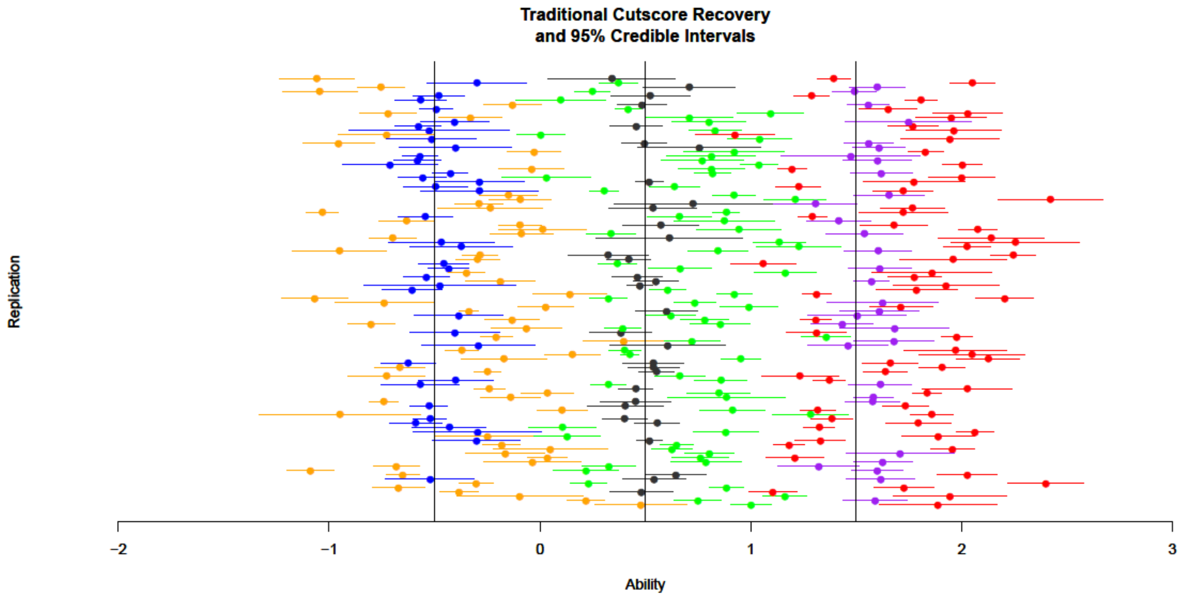

Figure A.21. N_I = 8, N_G = 2, σ^2_B = 0.2, σ^2_C = 0.0; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

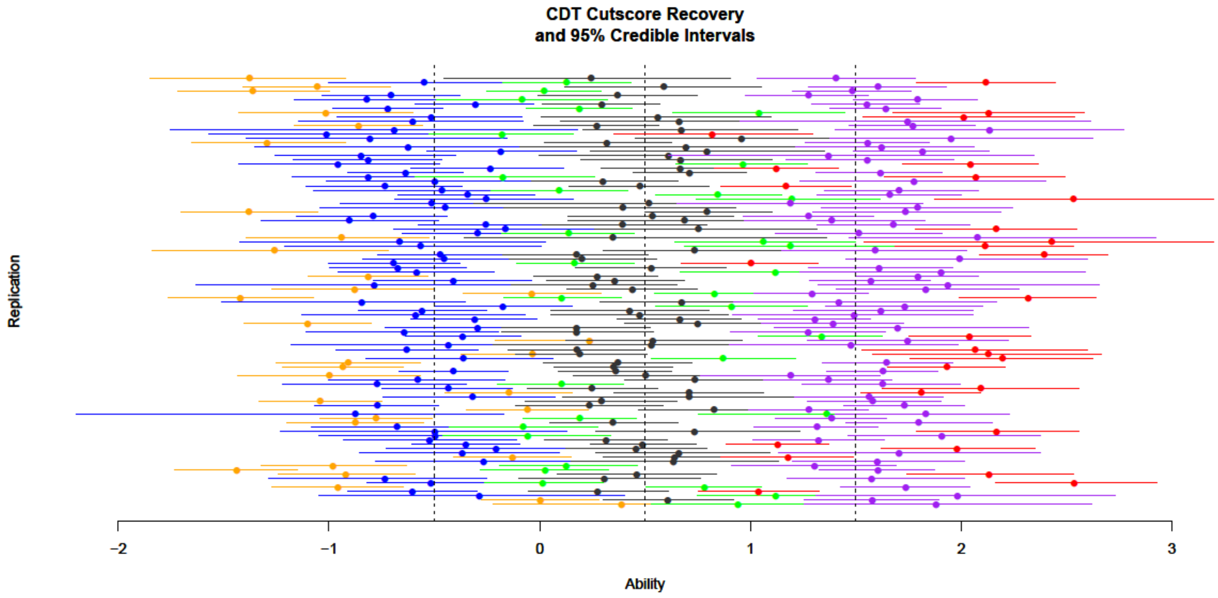

Figure A.22. N_I = 8, N_G = 2, σ^2_B = 0.2, σ^2_C = 0.0; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

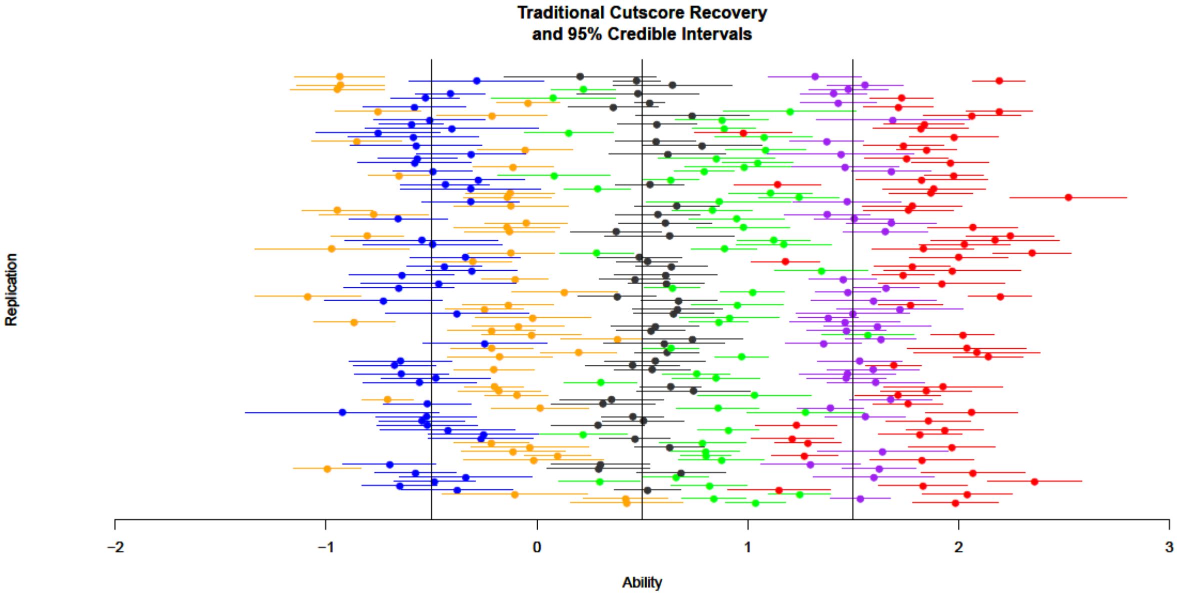

Figure A.23. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

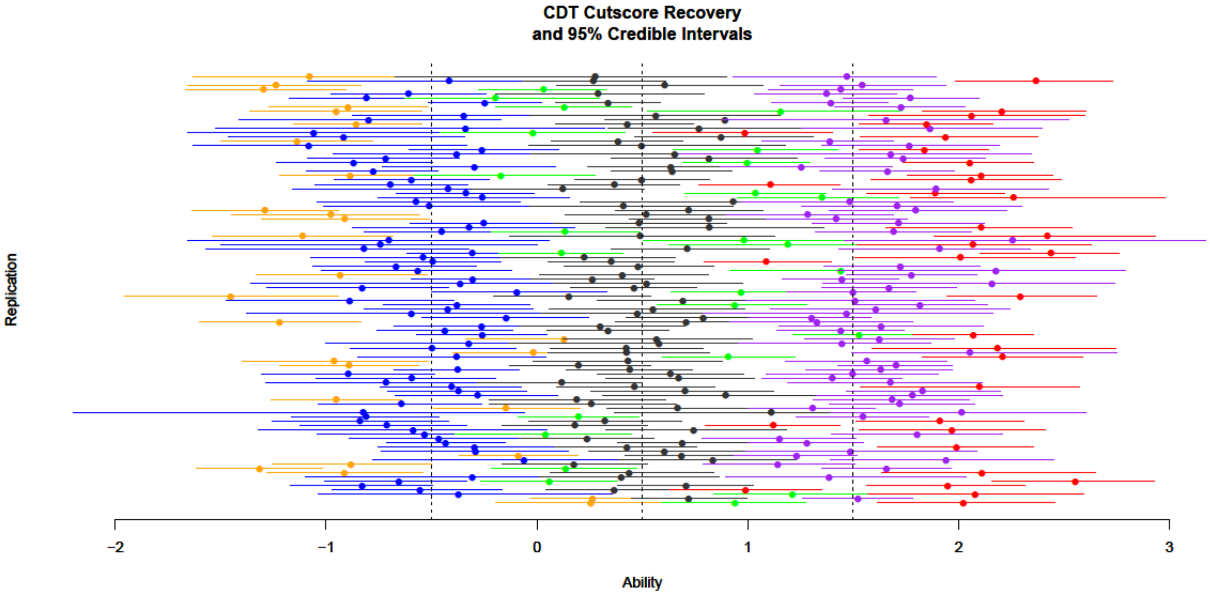

Figure A.24. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

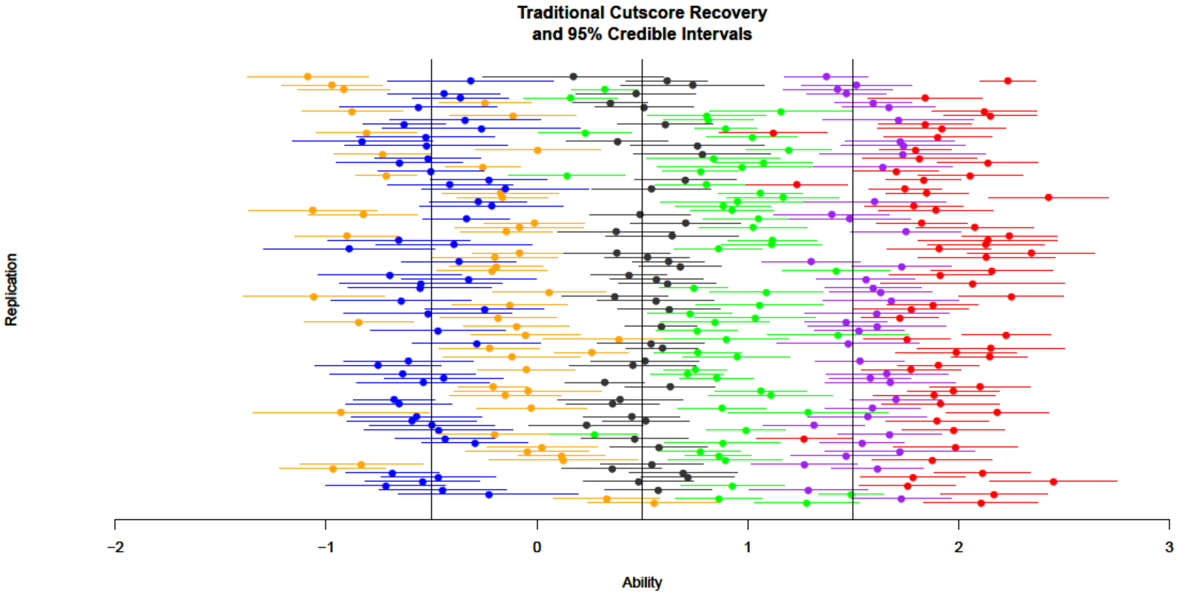

Figure A.25. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

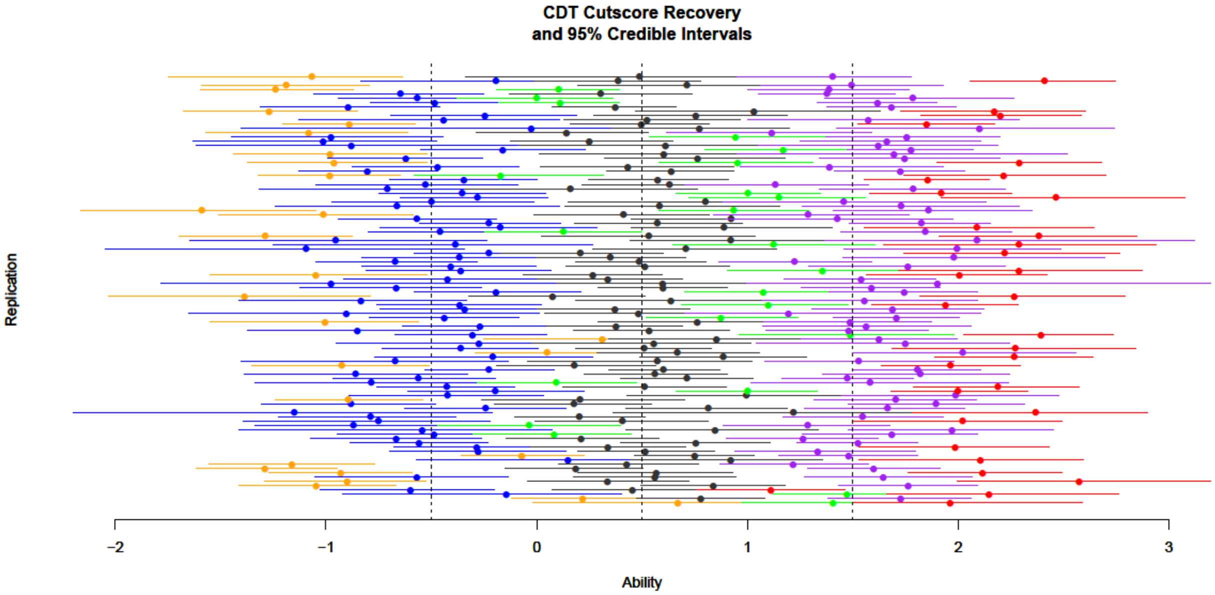

Figure A.26. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

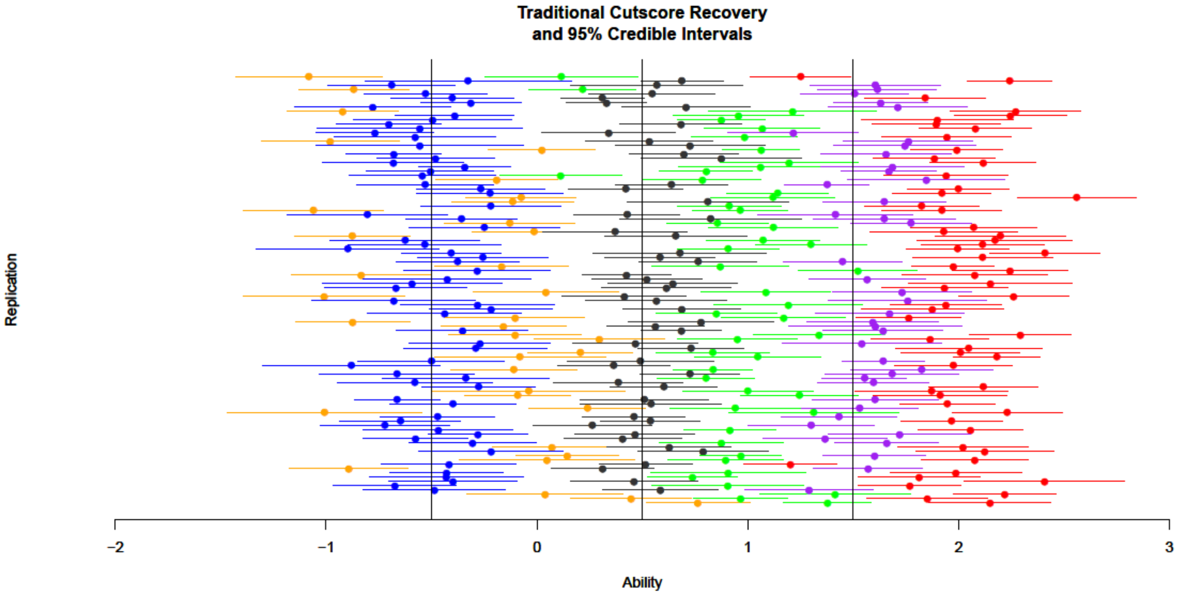

Figure A.27. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

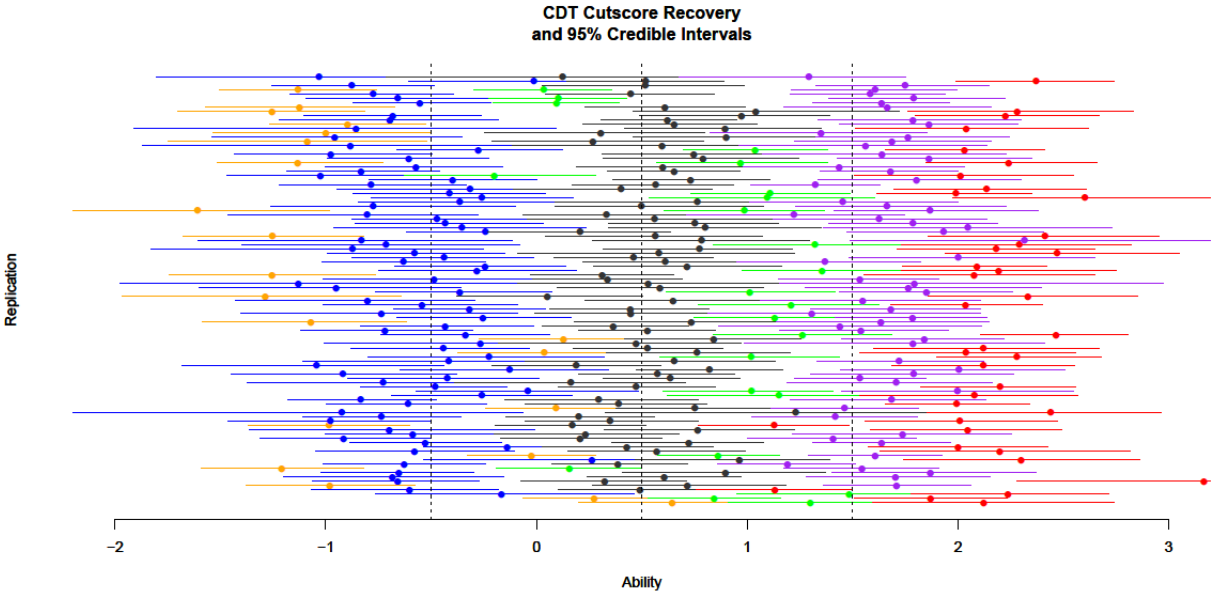

Figure A.28. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

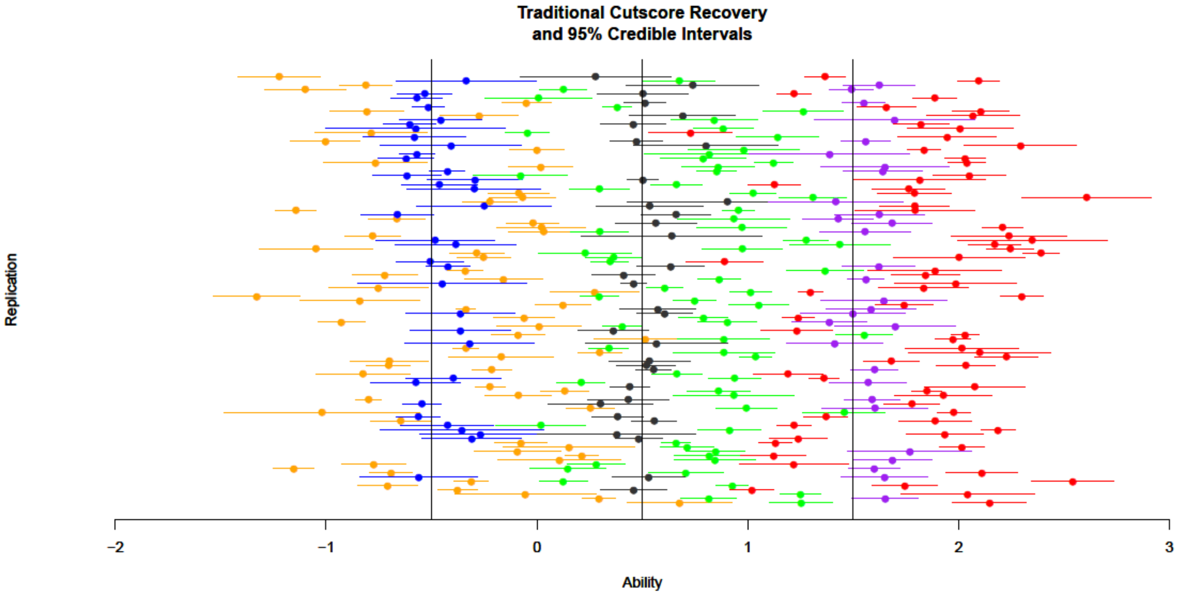

Figure A.29. N_I = 8, N_G = 2, σ^2_B = 0.3, σ^2_C = 0.0; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

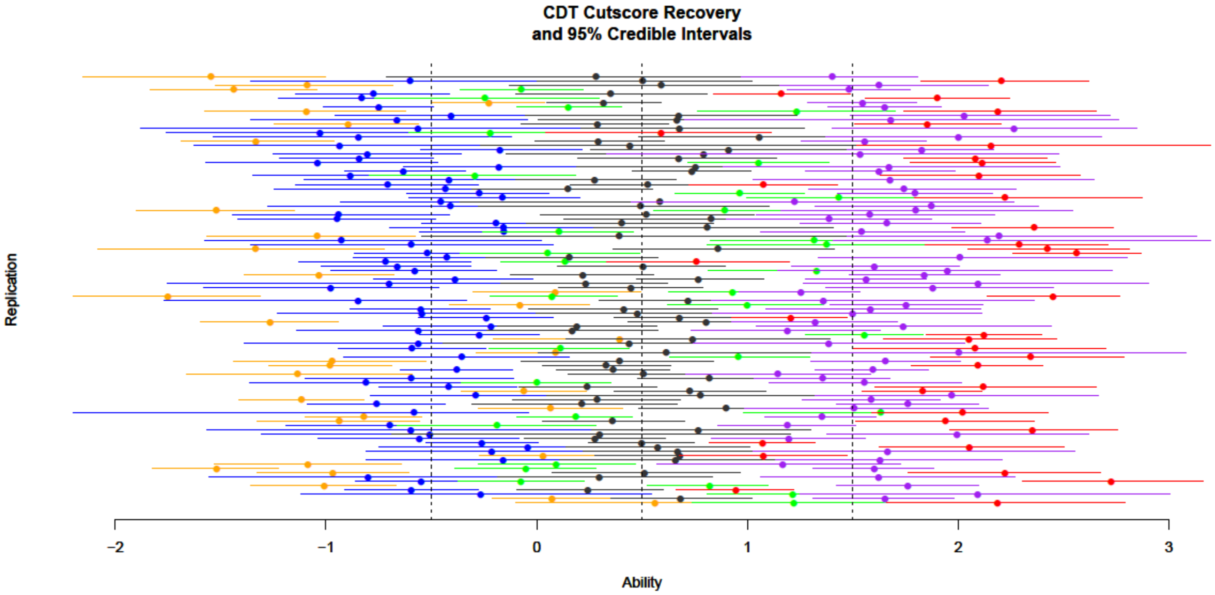

Figure A.30. N_I = 8, N_G = 2, σ^2_B = 0.3, σ^2_C = 0.0; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

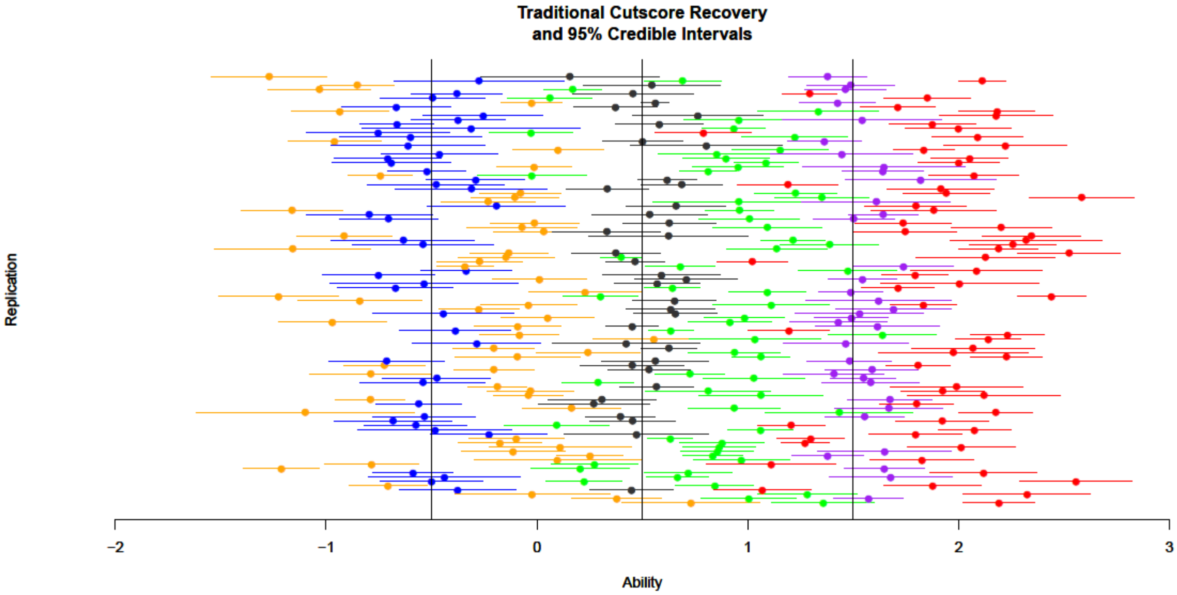

Figure A.31. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

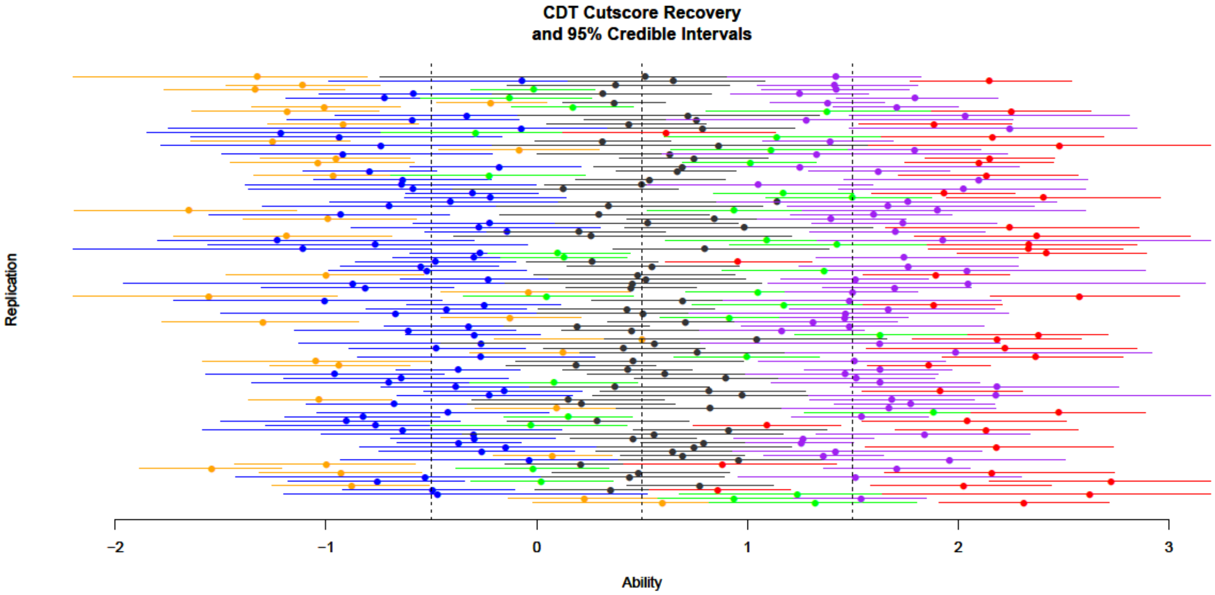

Figure A.32. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

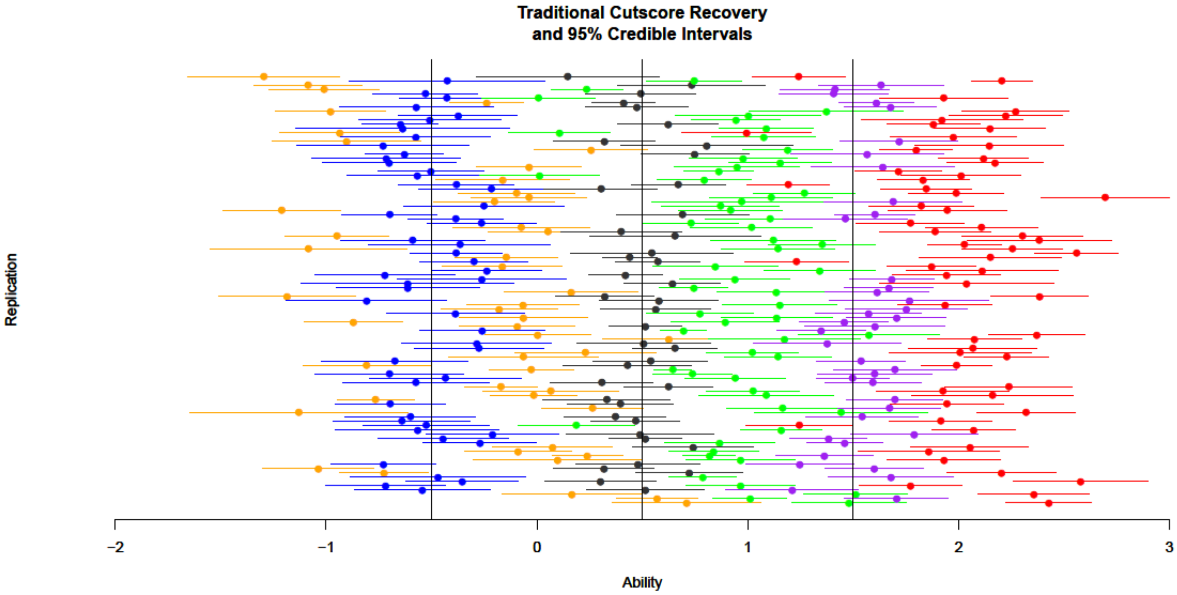

Figure A.33. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the include the true cutscore.

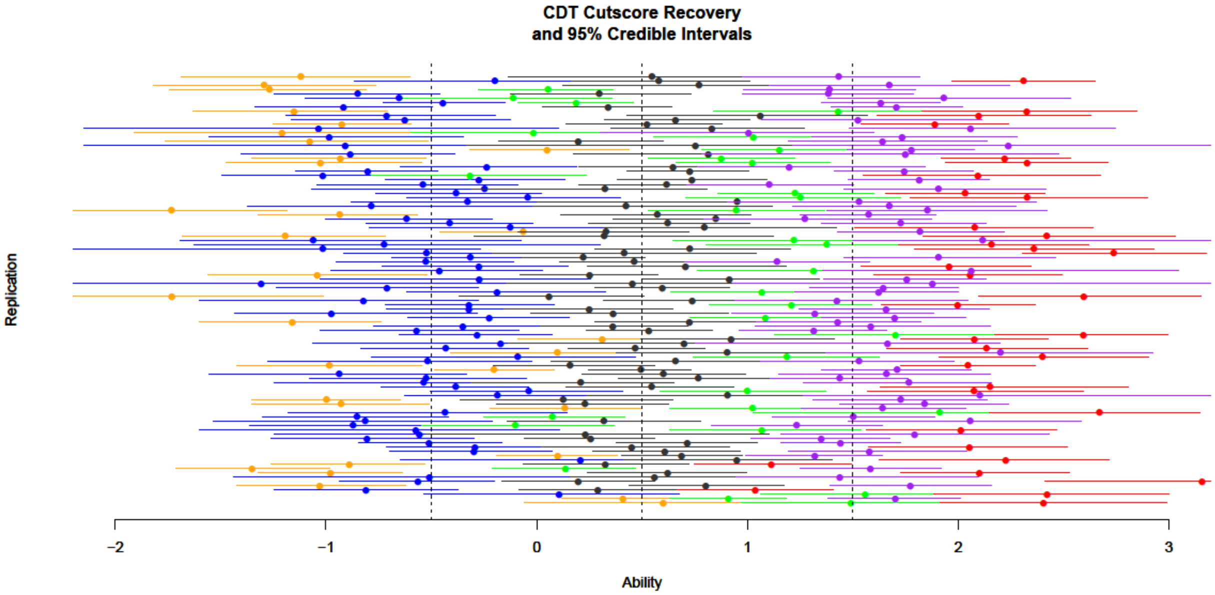

Figure A.34. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

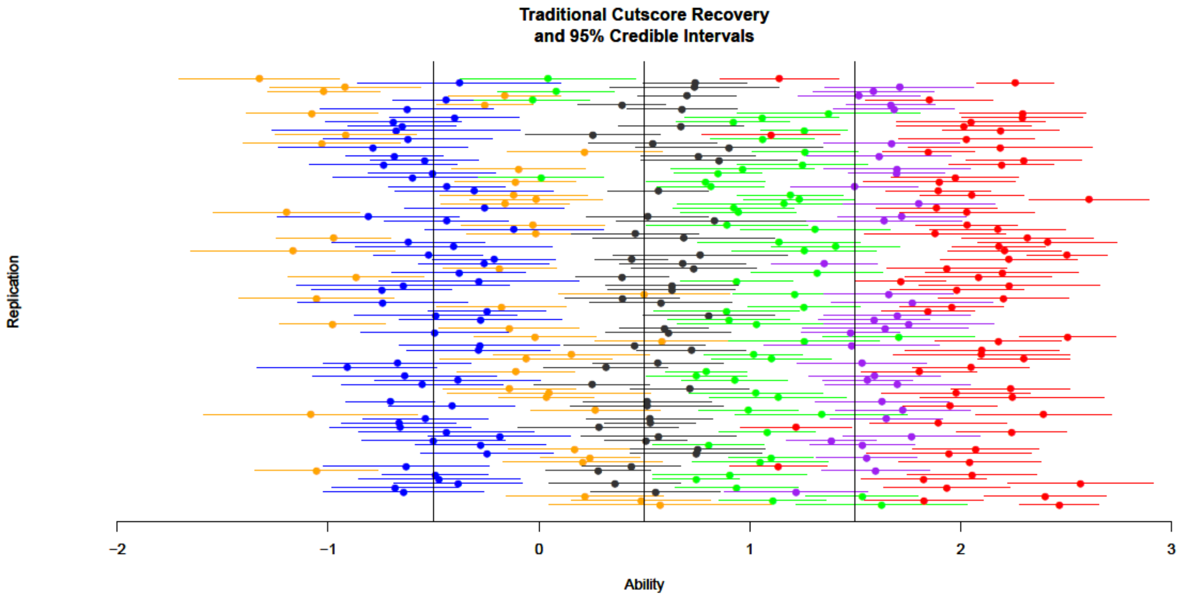

Figure A.35. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

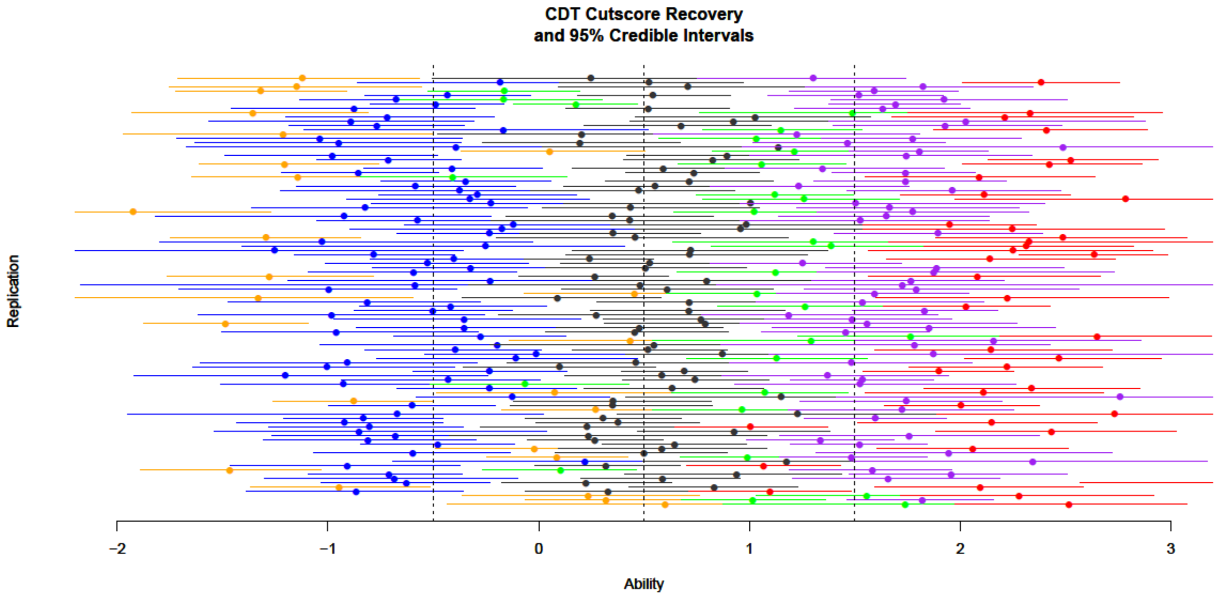

Figure A.36. $N_I = 8$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

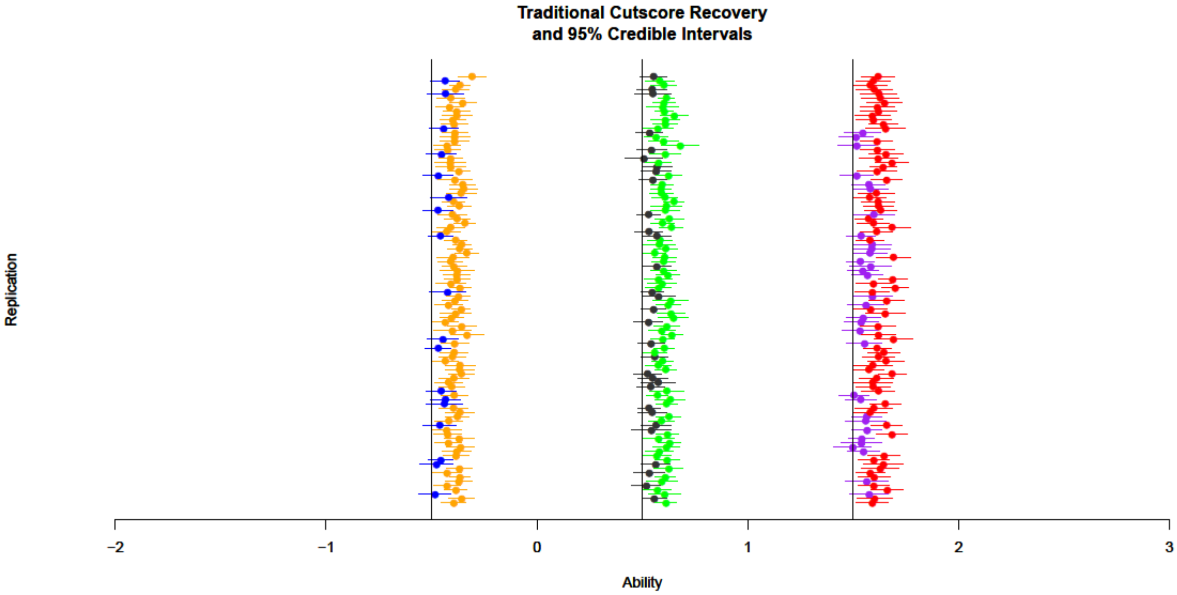

Figure A.37. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

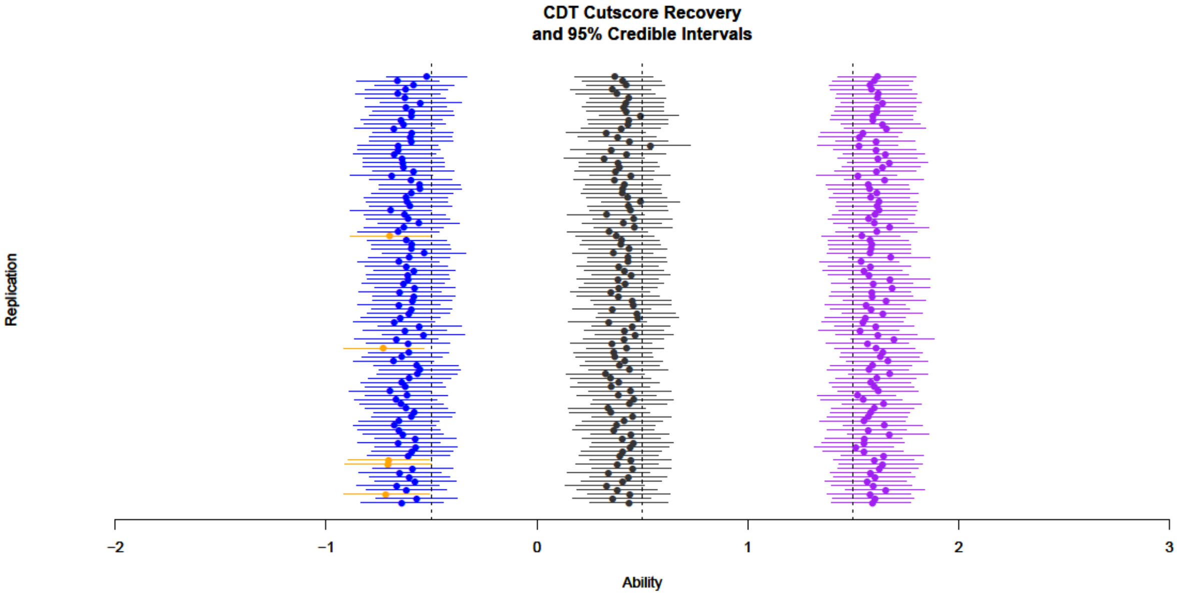

Figure A.38. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

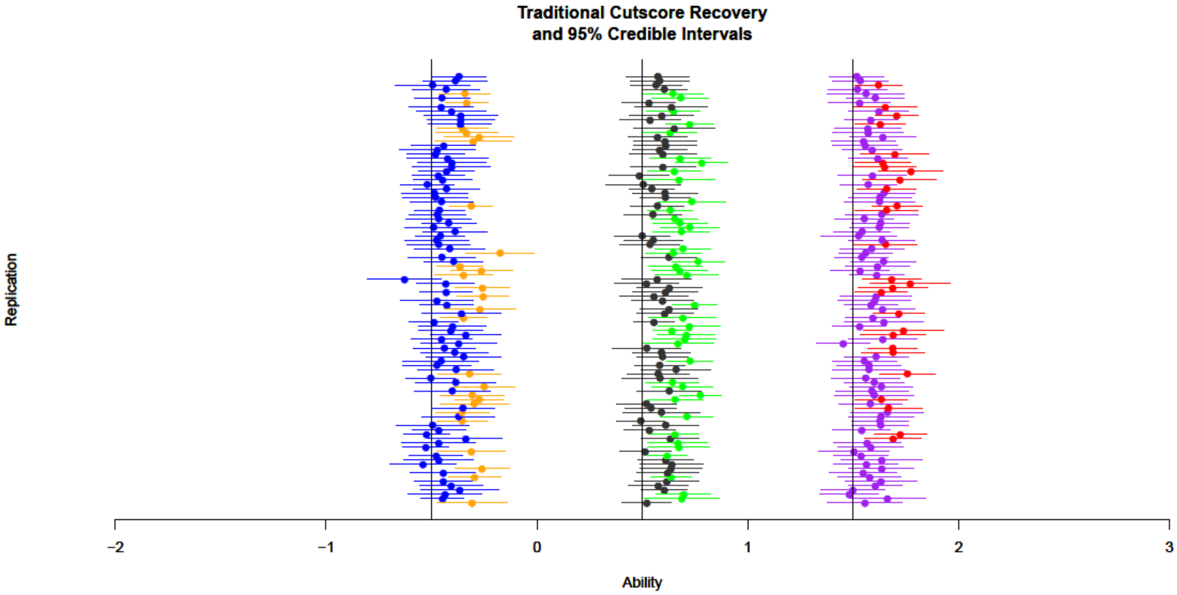

Figure A.39. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

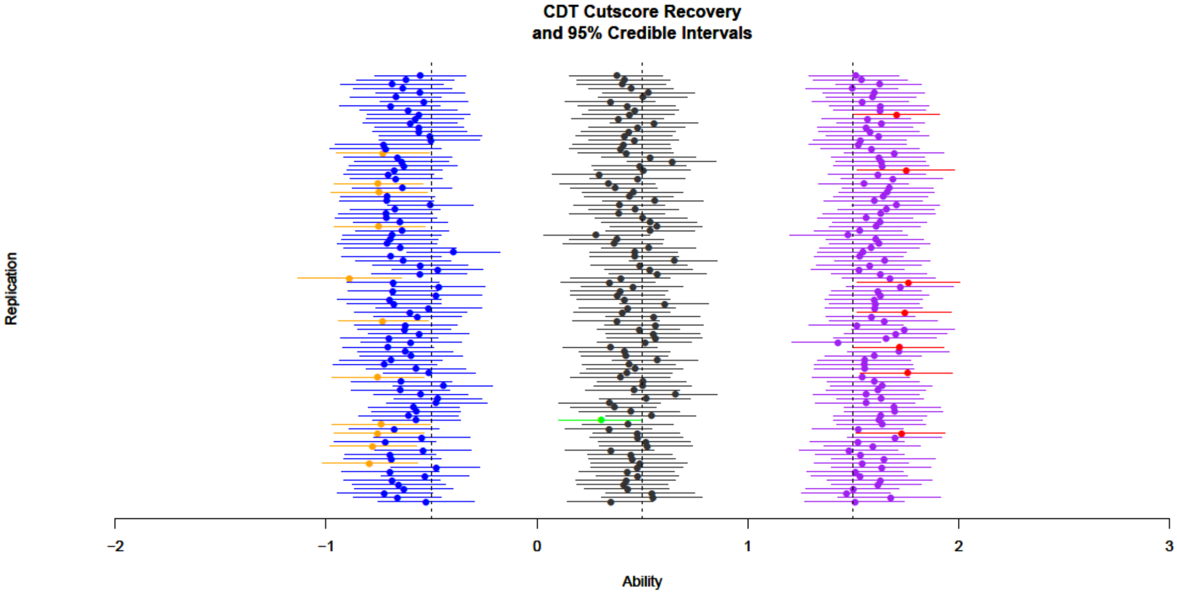

Figure A.40. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

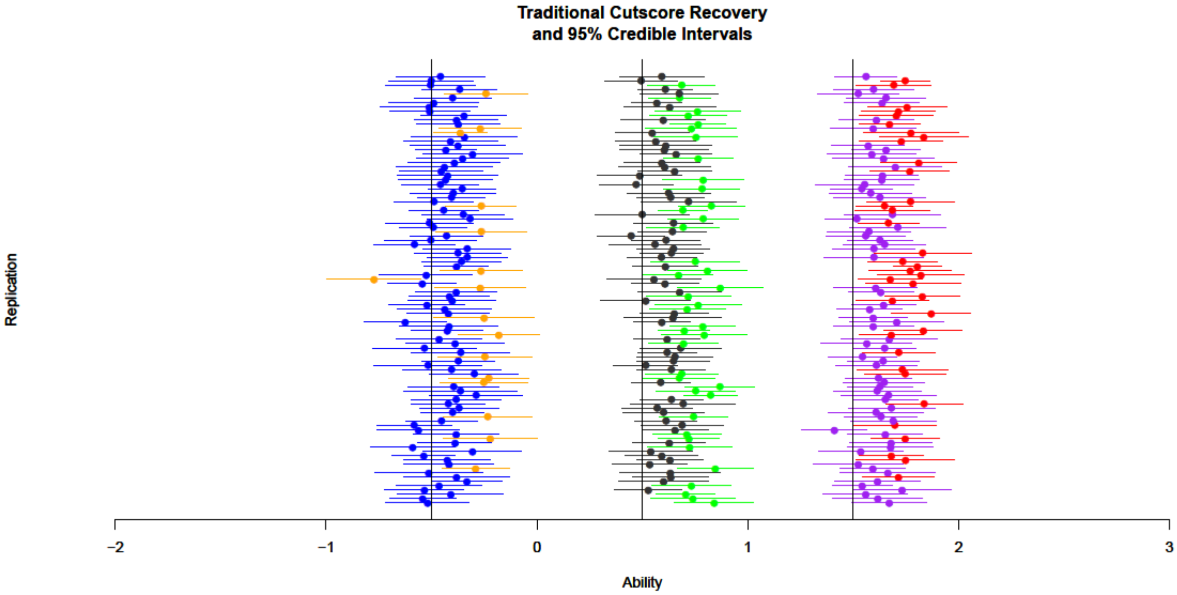

Figure A.41. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

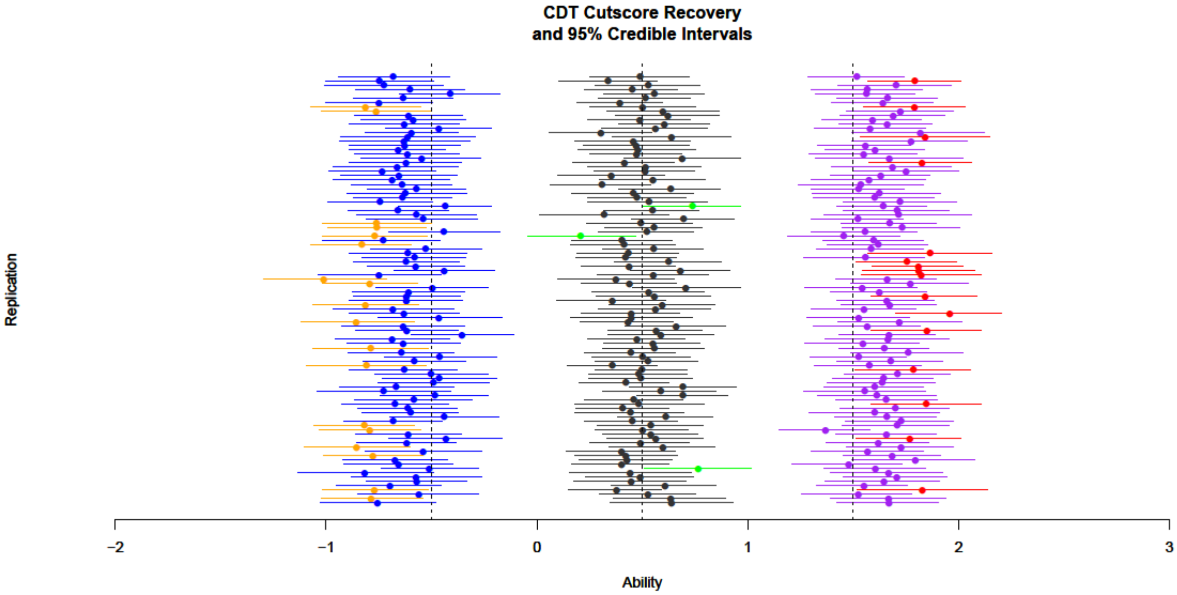

Figure A.42. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

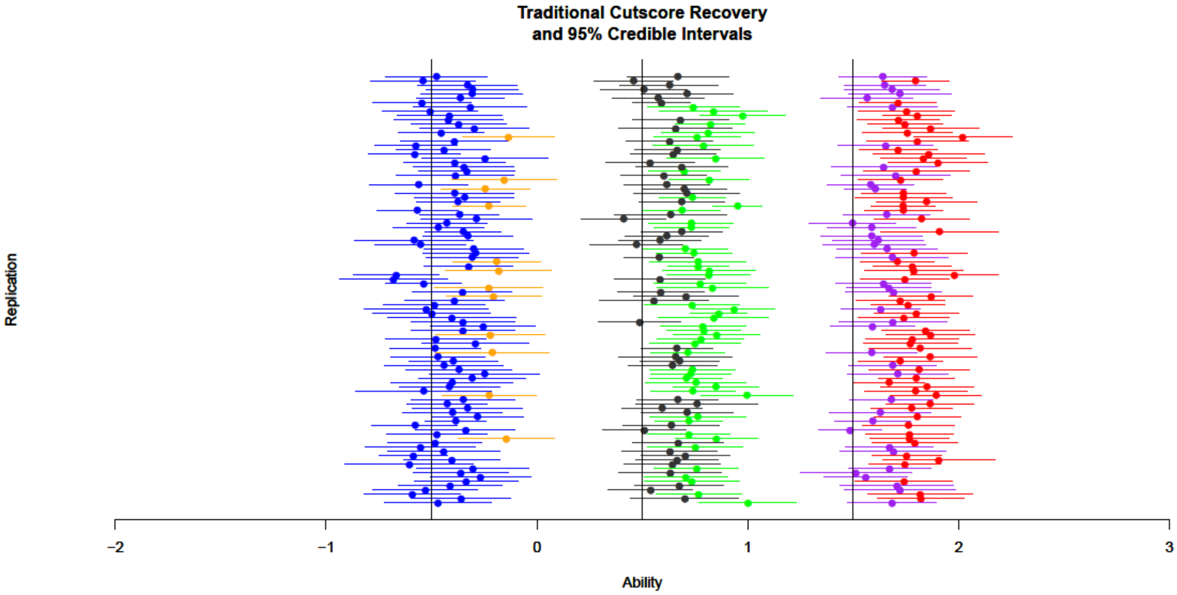

Figure A.43. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

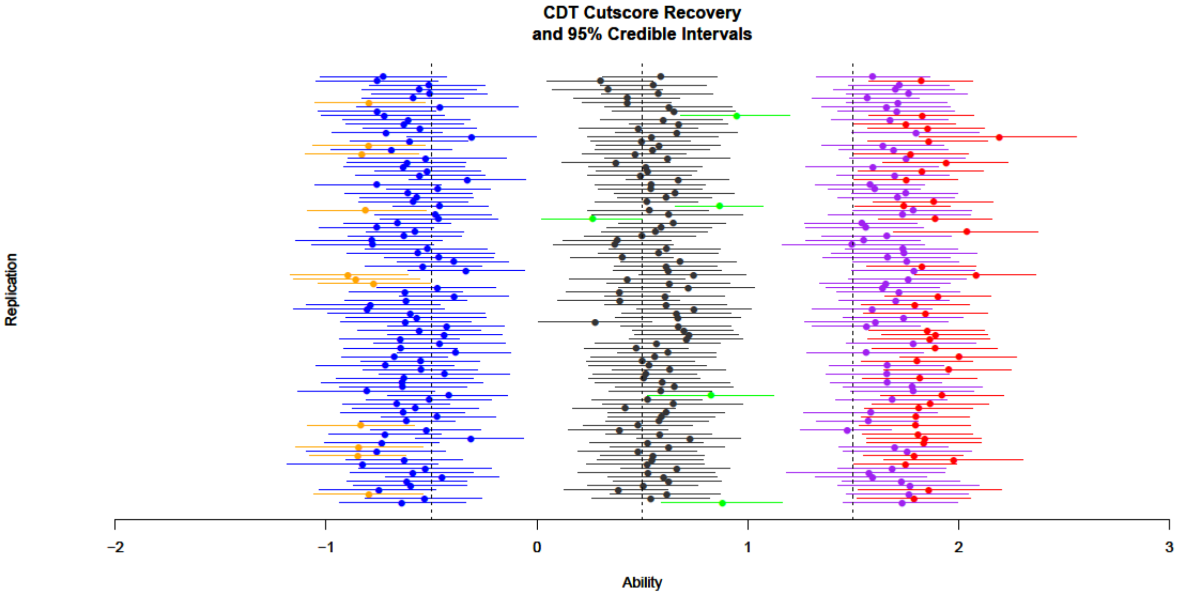

Figure A.44. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

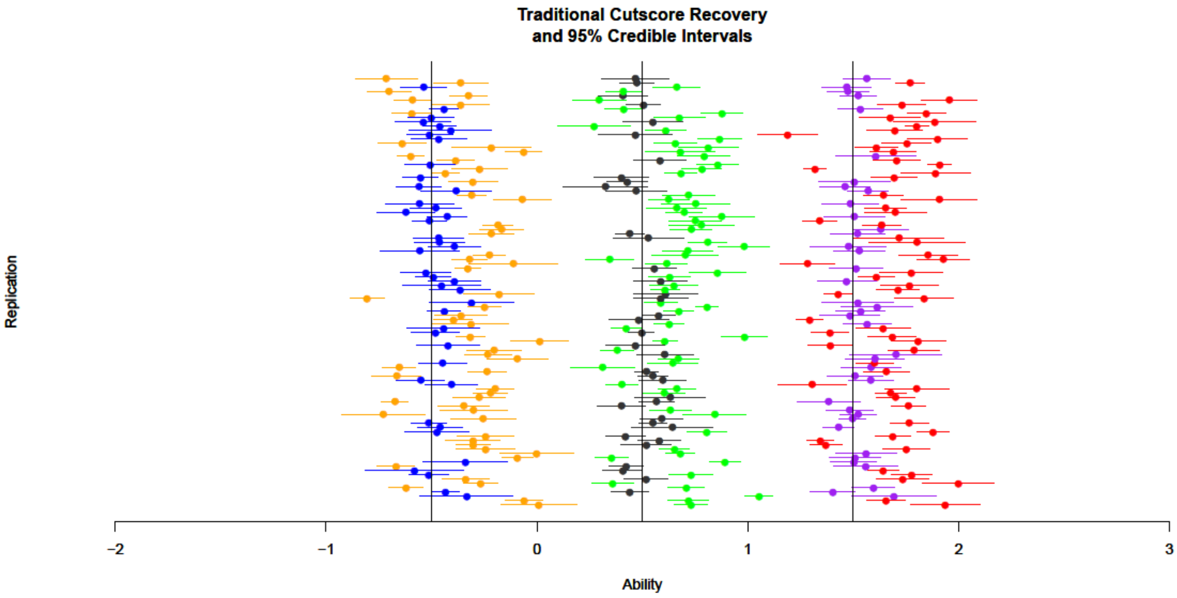

Figure A.45. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

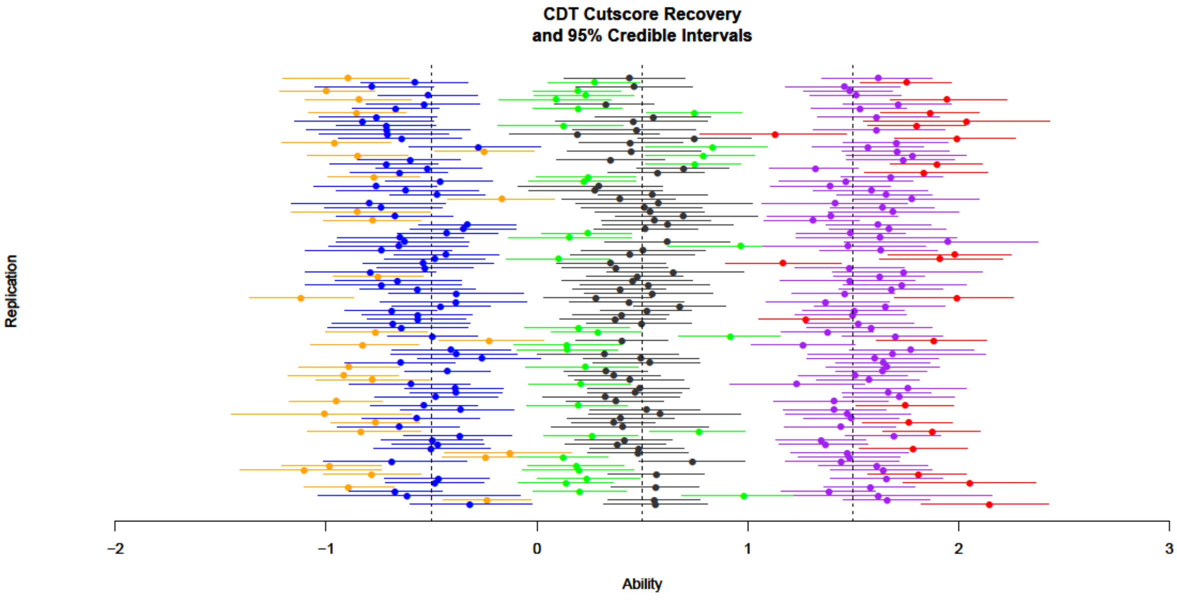

Figure A.46. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

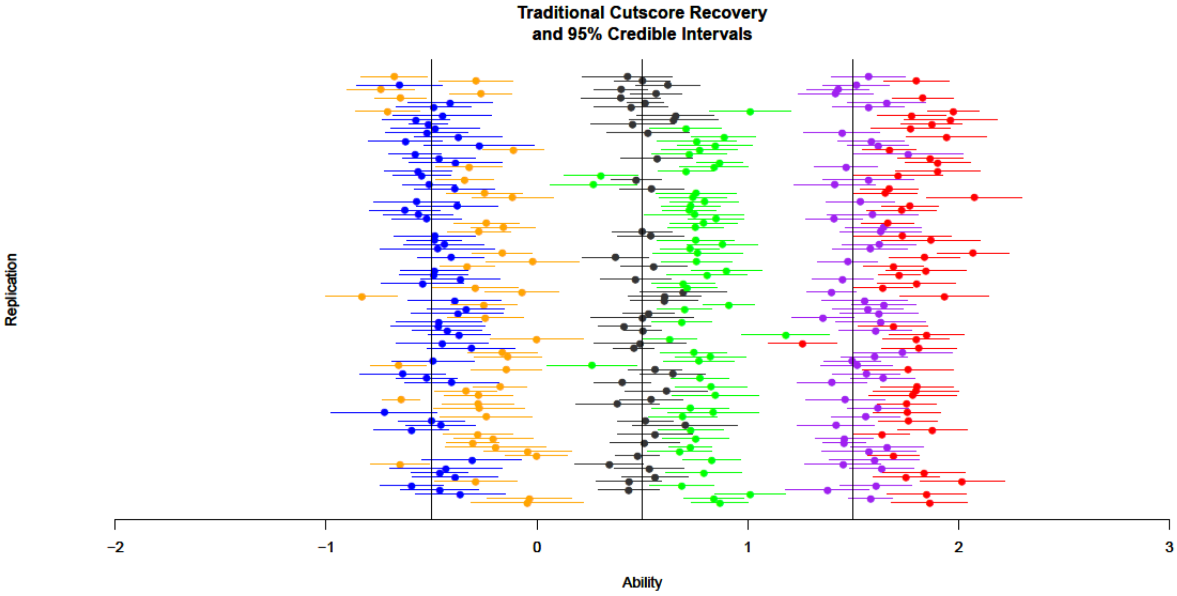

Figure A.47. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

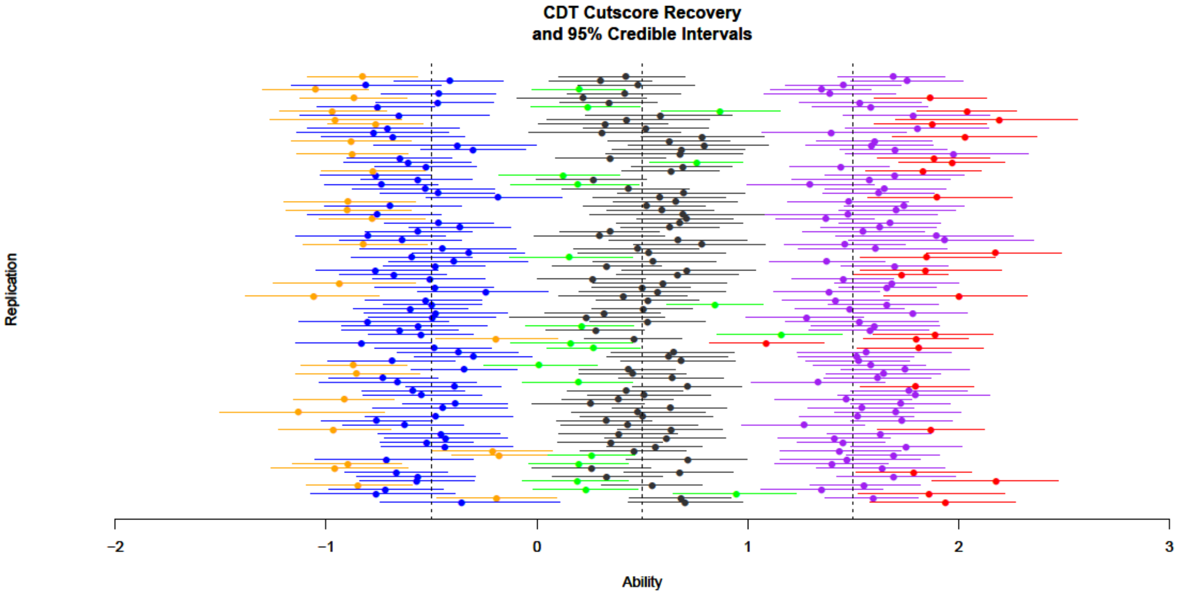

Figure A.48. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

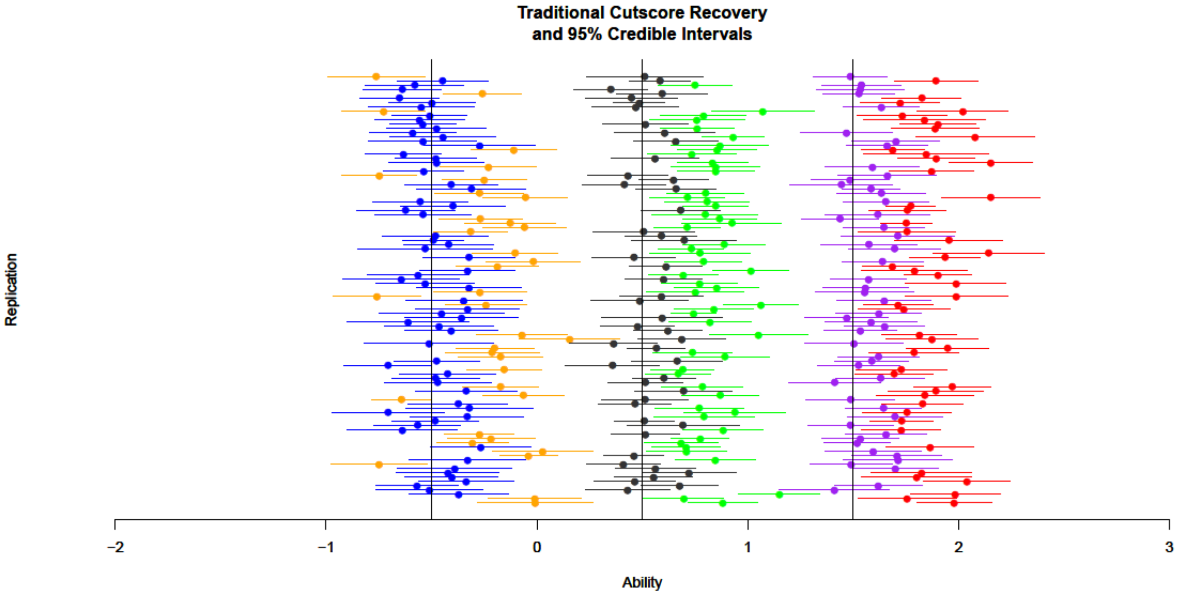

Figure A.49. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

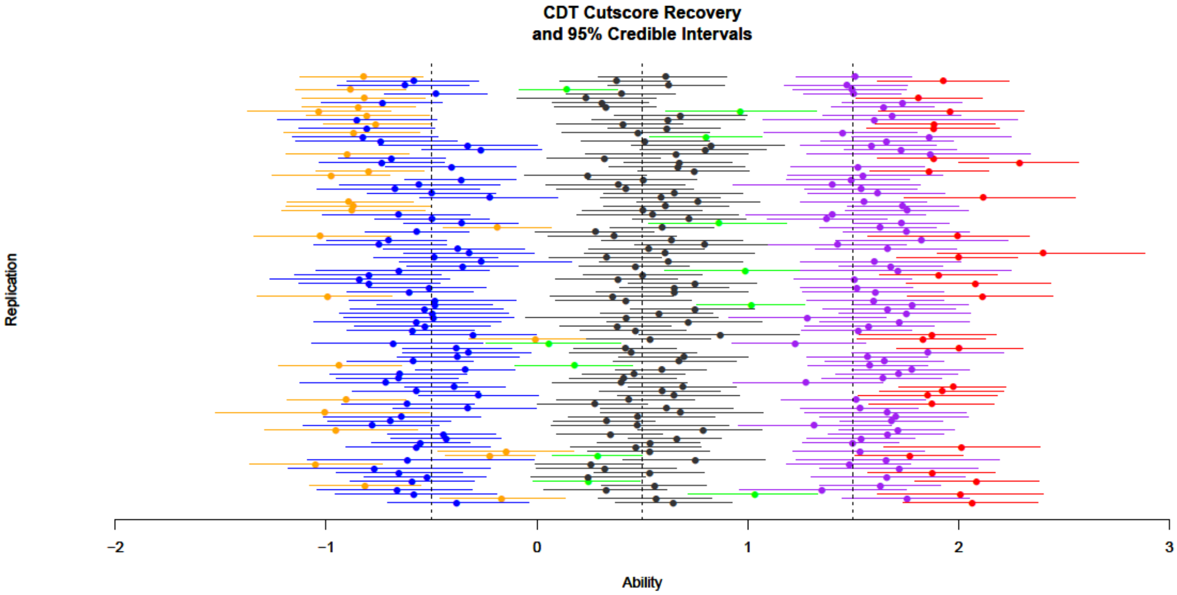

Figure A.50. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

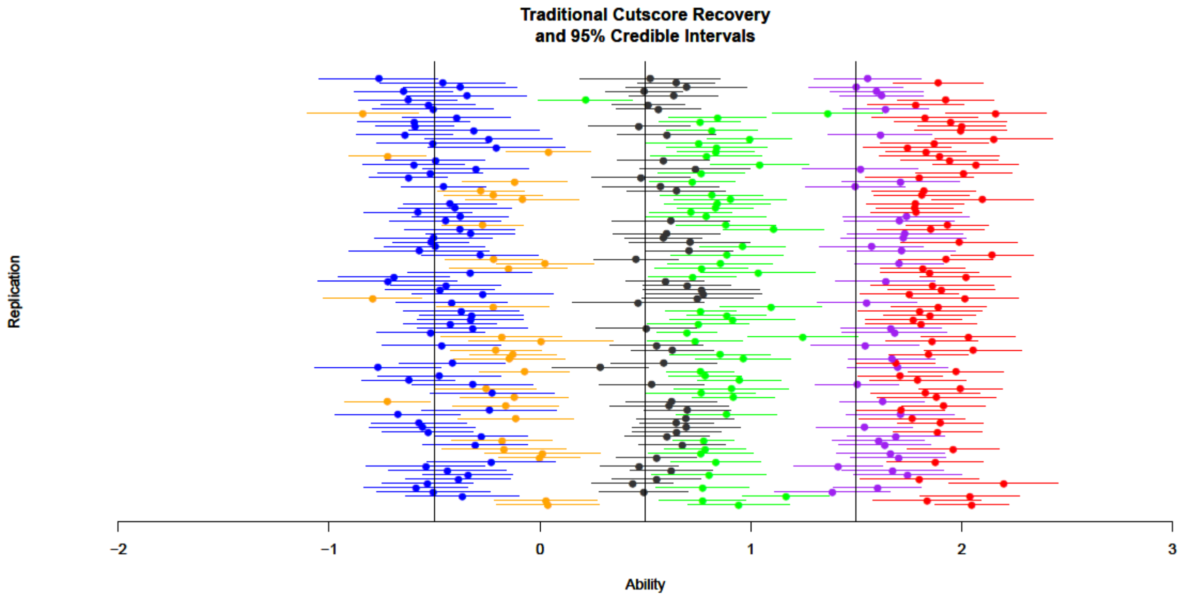

Figure A.51. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

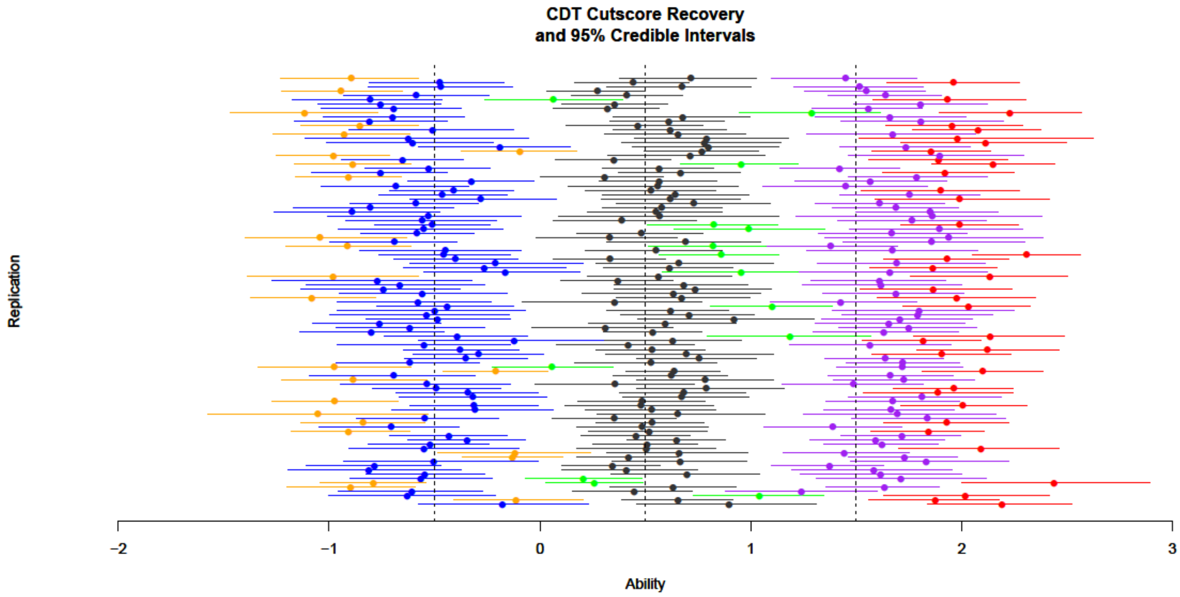

Figure A.52. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

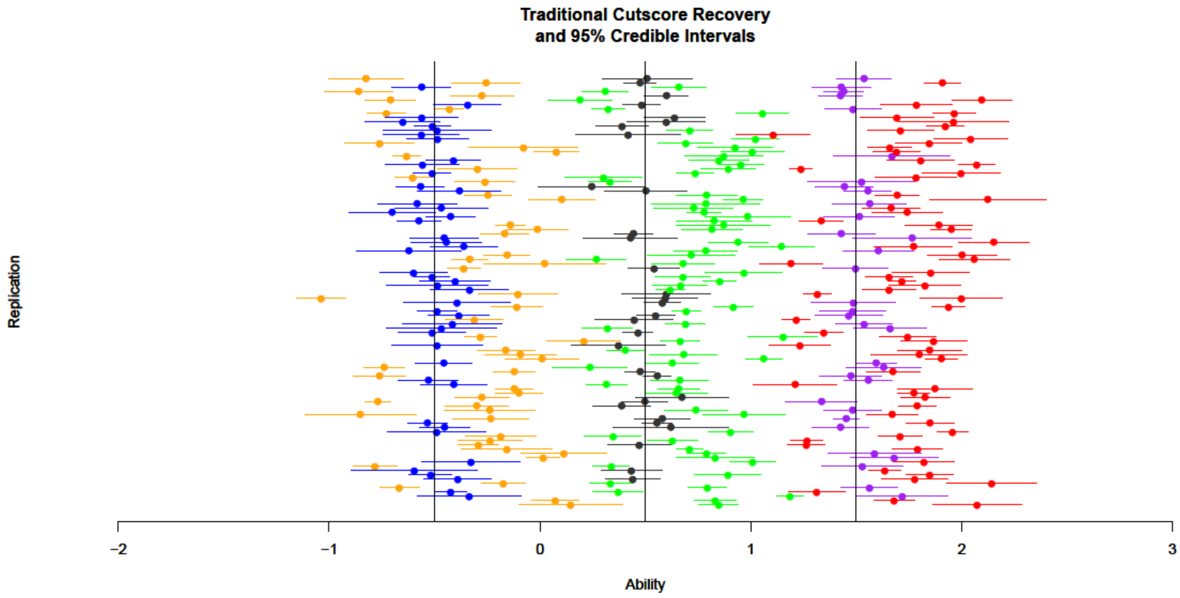

Figure A.53. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the include the true cutscore.

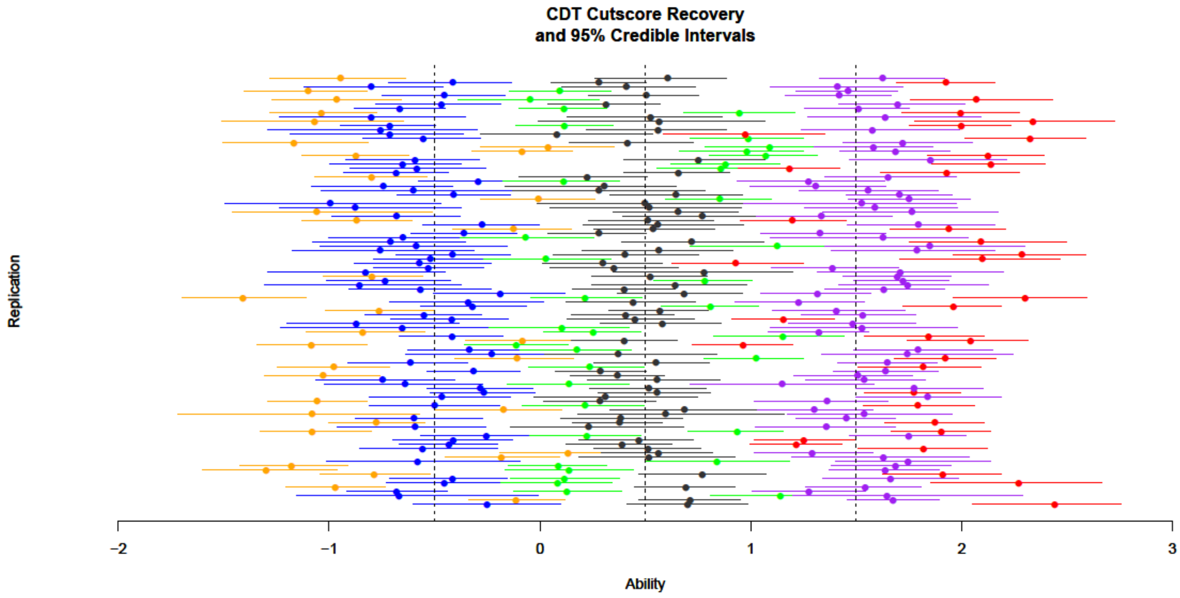

Figure A.54. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

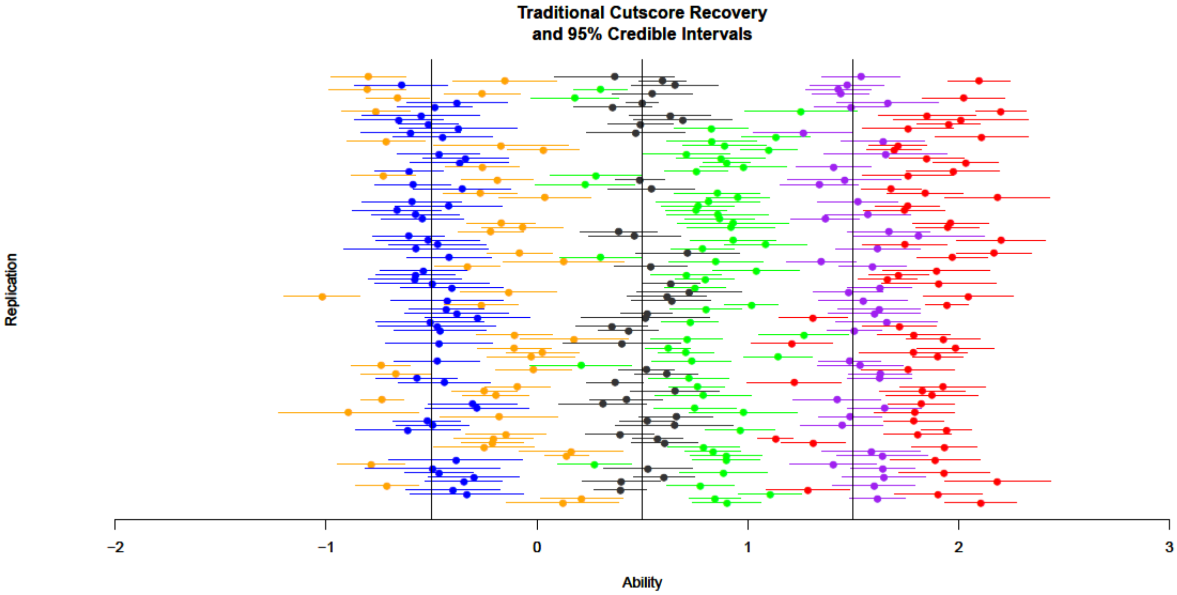

Figure A.55. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

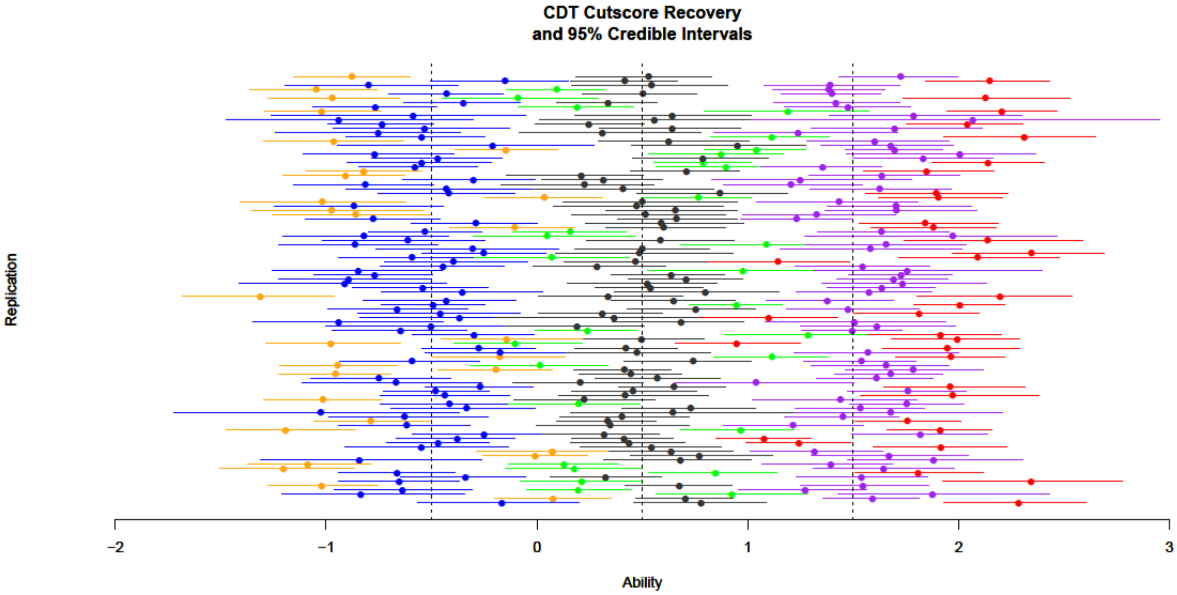

Figure A.56. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

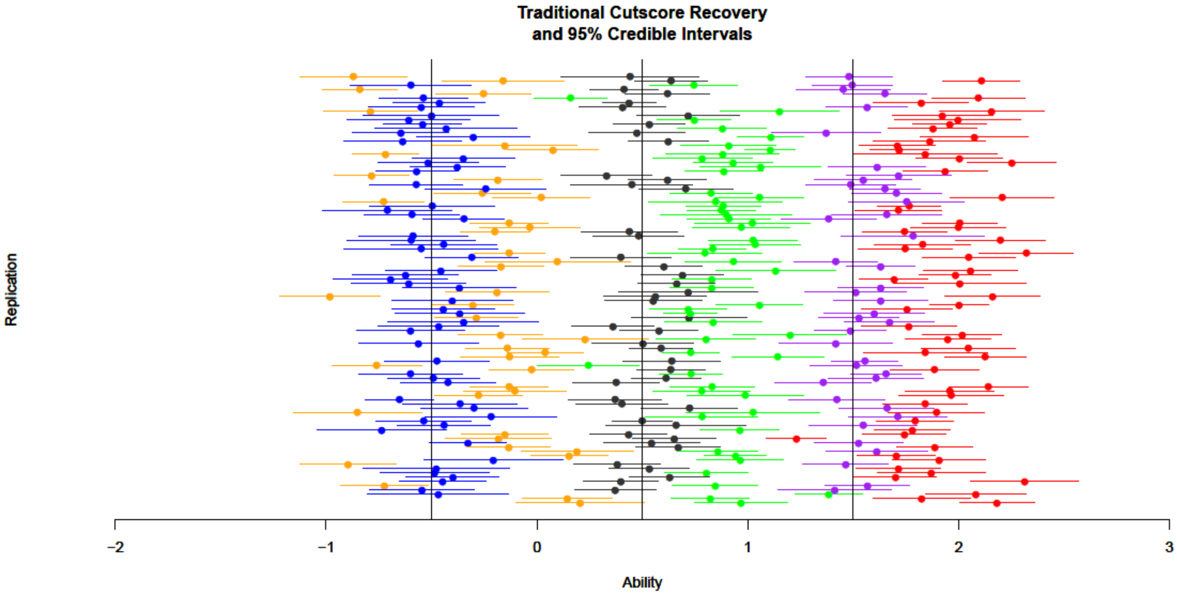

Figure A.57. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

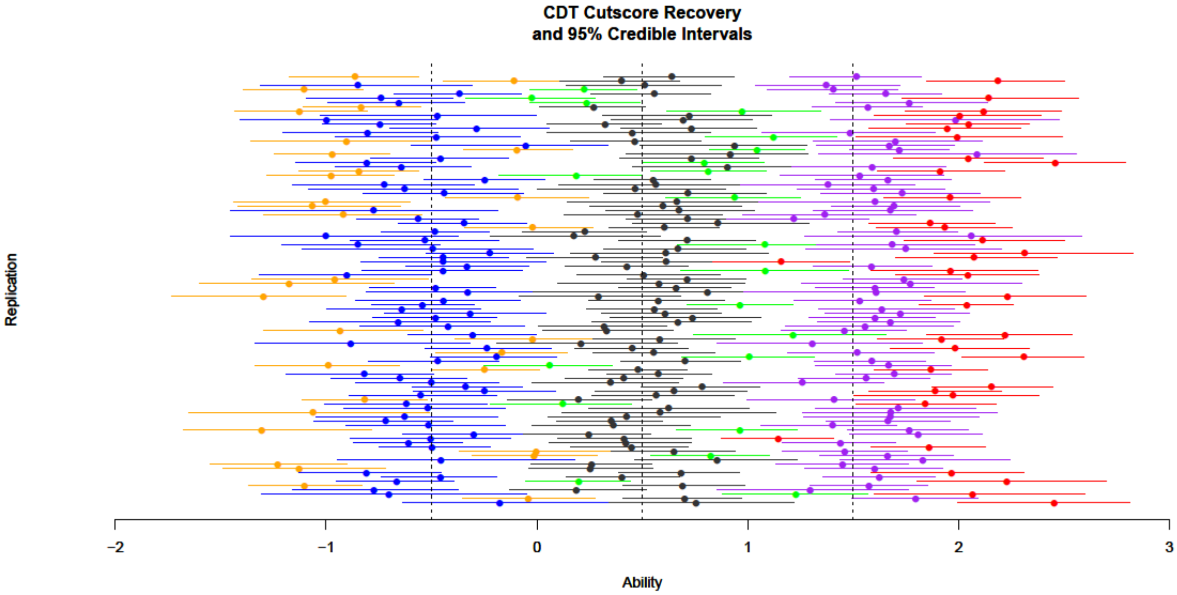

Figure A.58. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

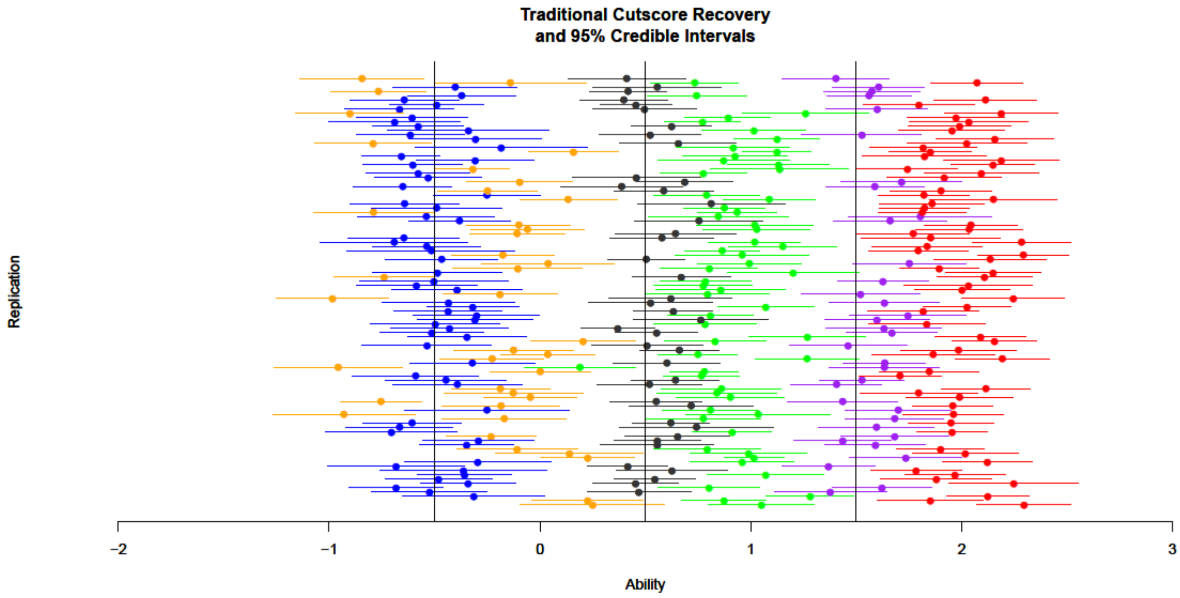

Figure A.59. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

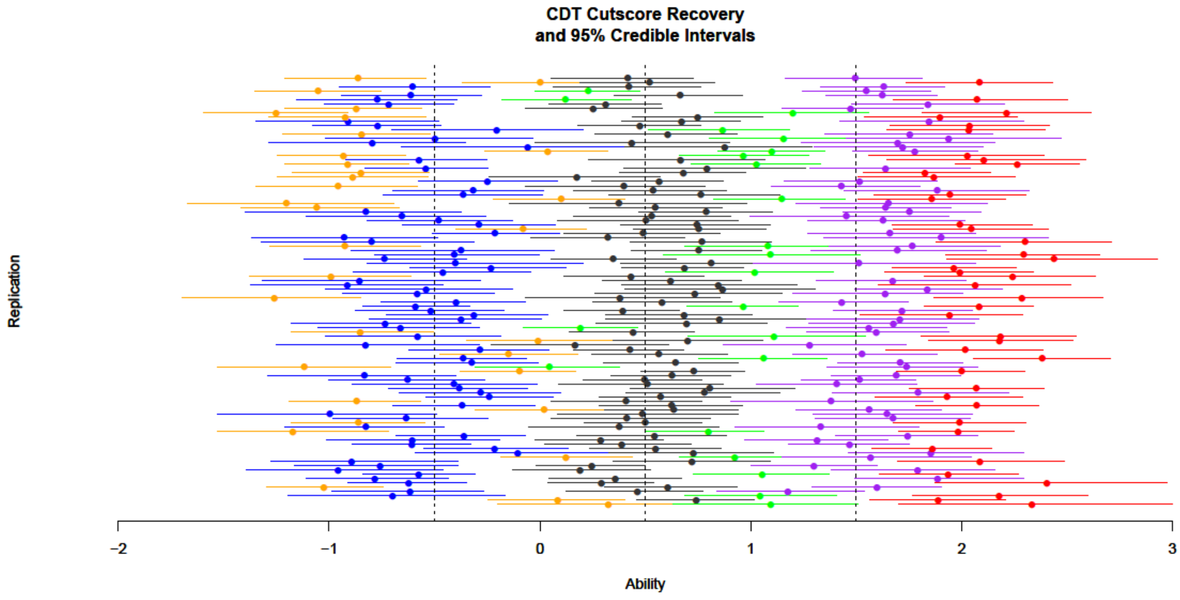

Figure A.60. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

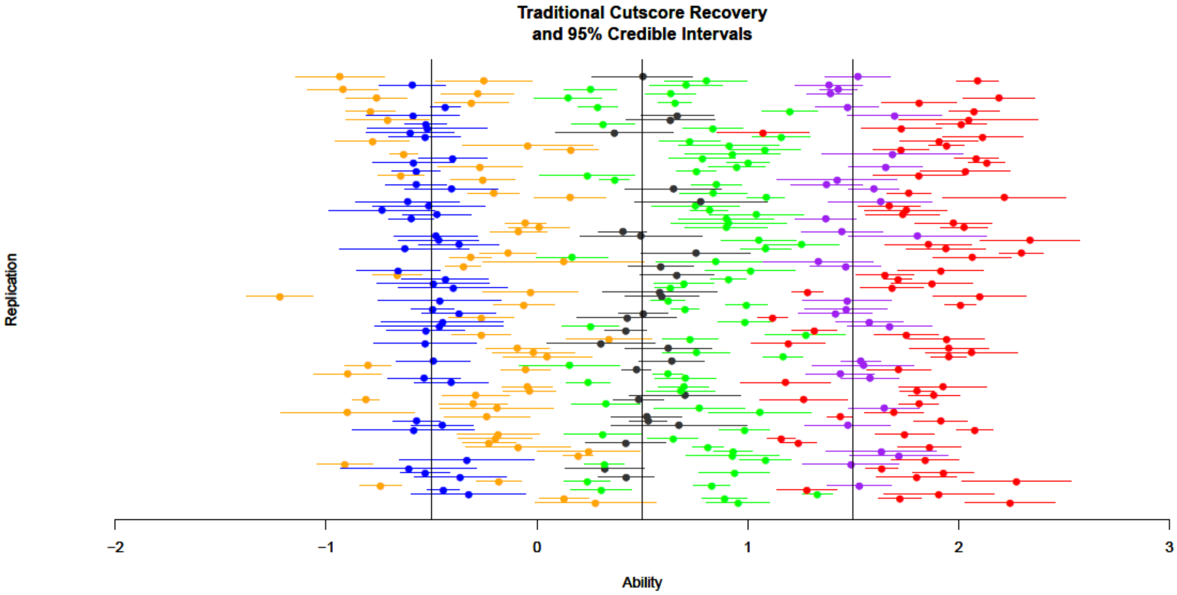

Figure A.61. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

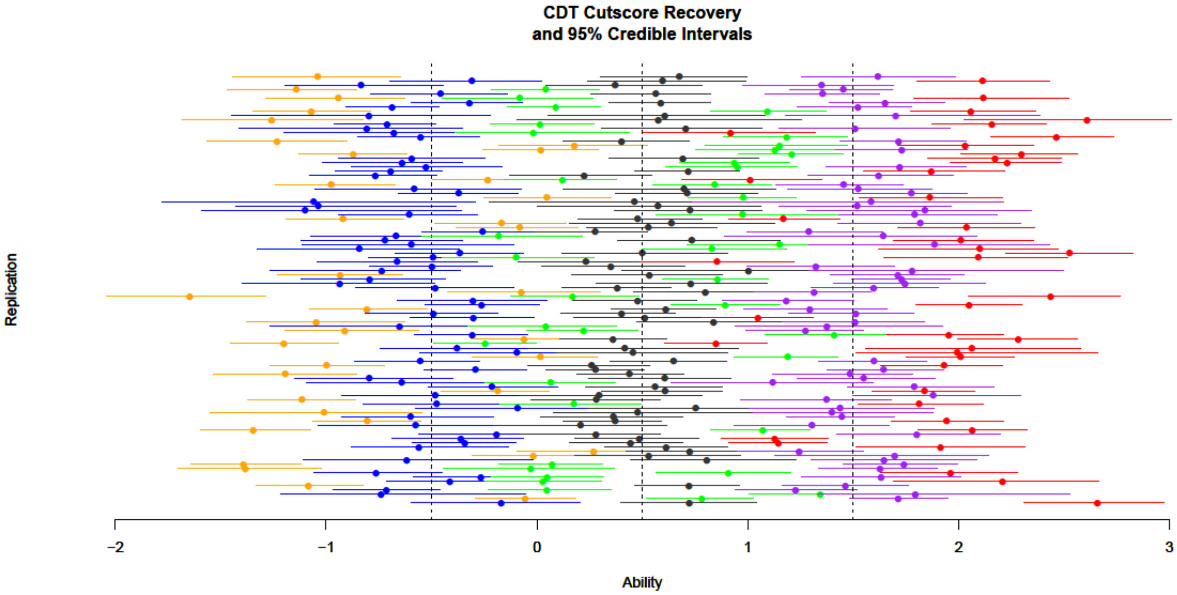

Figure A.62. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

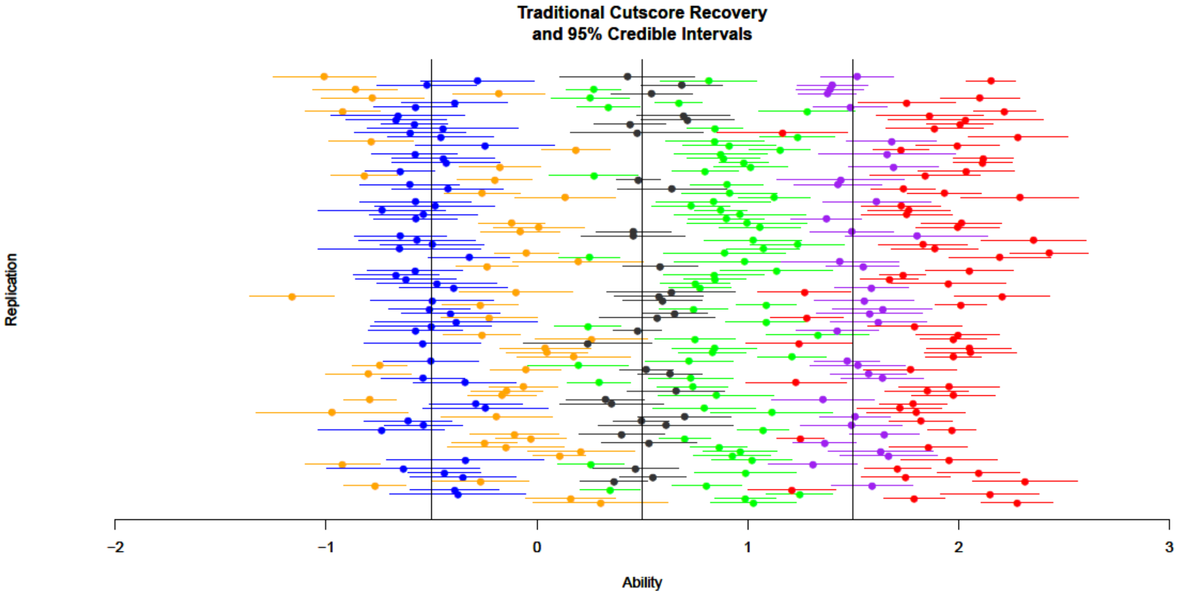

Figure A.63. N$_I$ = 8, N$_G$ = 3, σ^2$_B$ = 0.3, σ^2$_C$ = 0.1; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

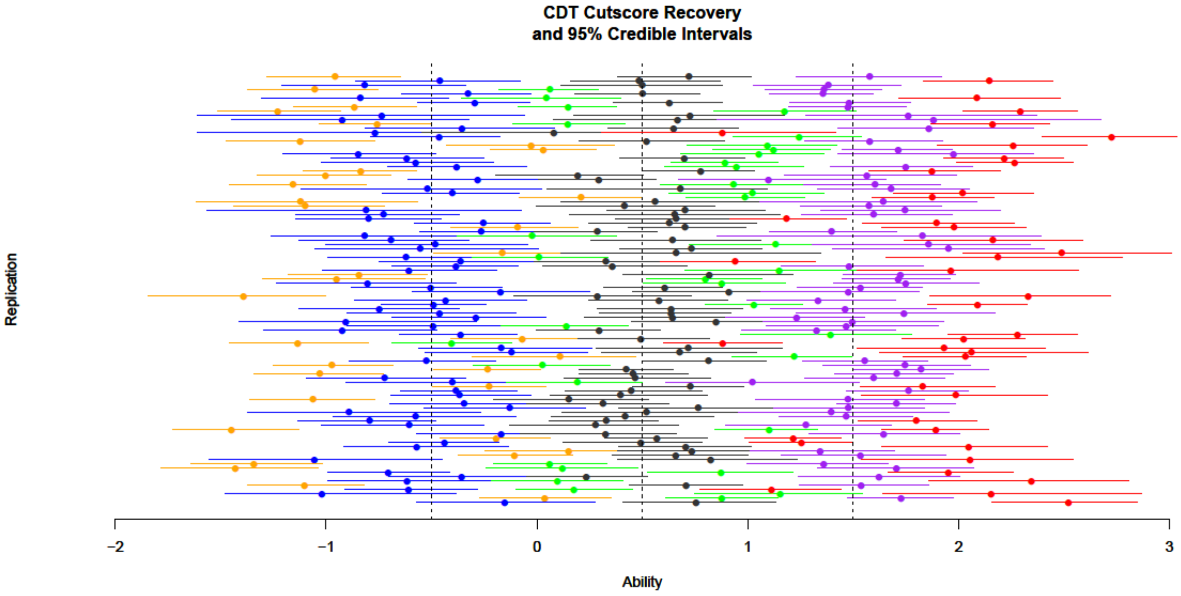

Figure A.64. N$_I$ = 8, N$_G$ = 3, σ^2$_B$ = 0.3, σ^2$_C$ = 0.1; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

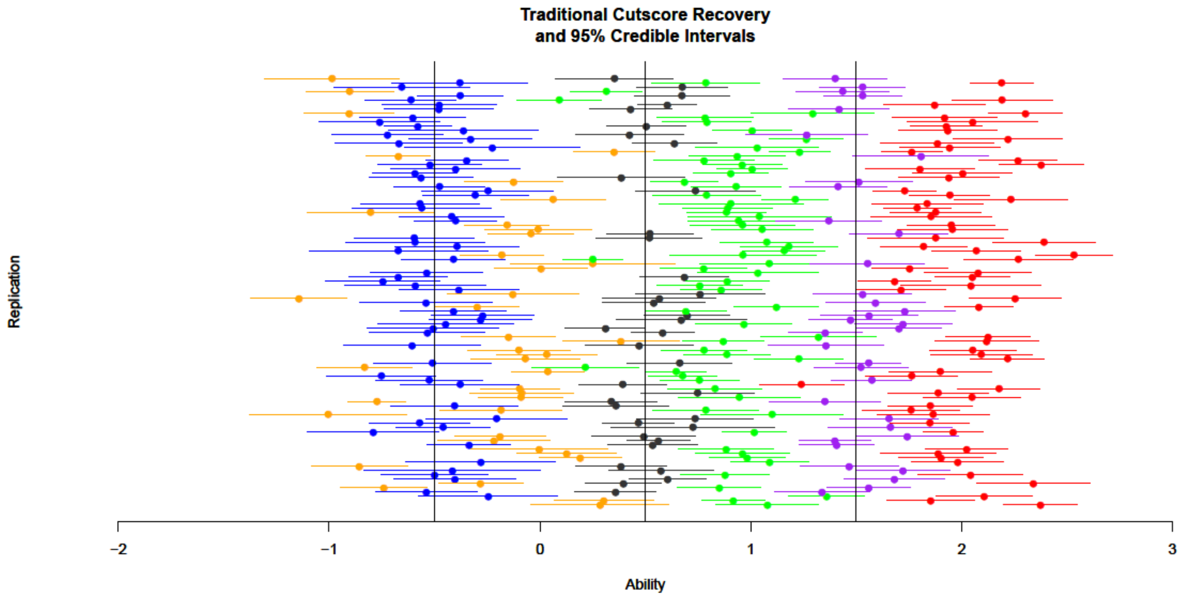

Figure A.65. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

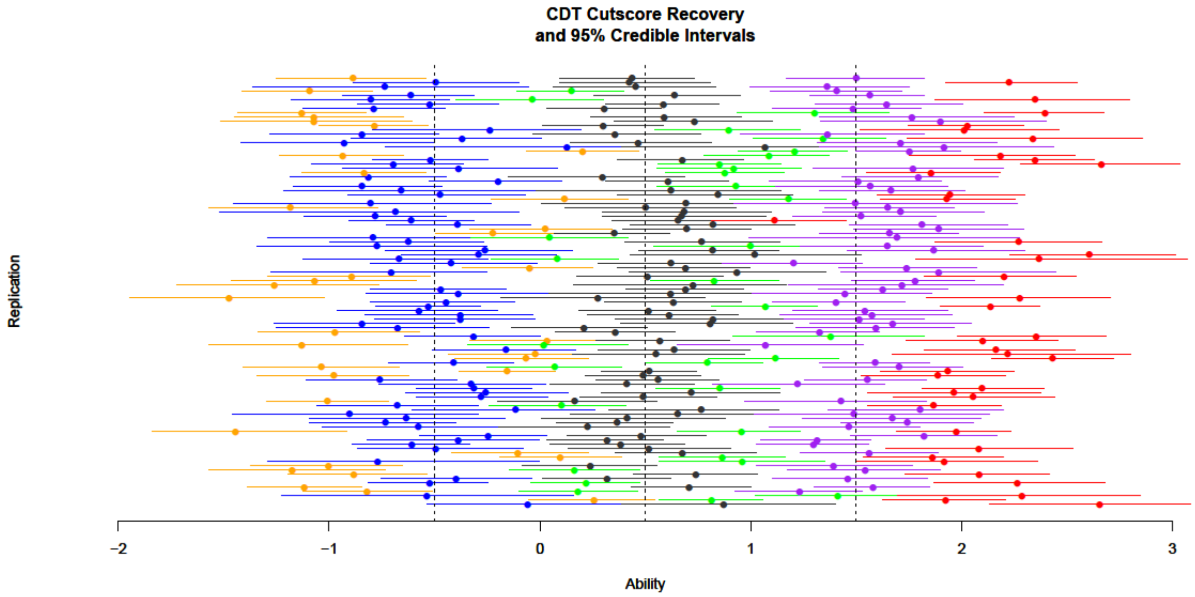

Figure A.66. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

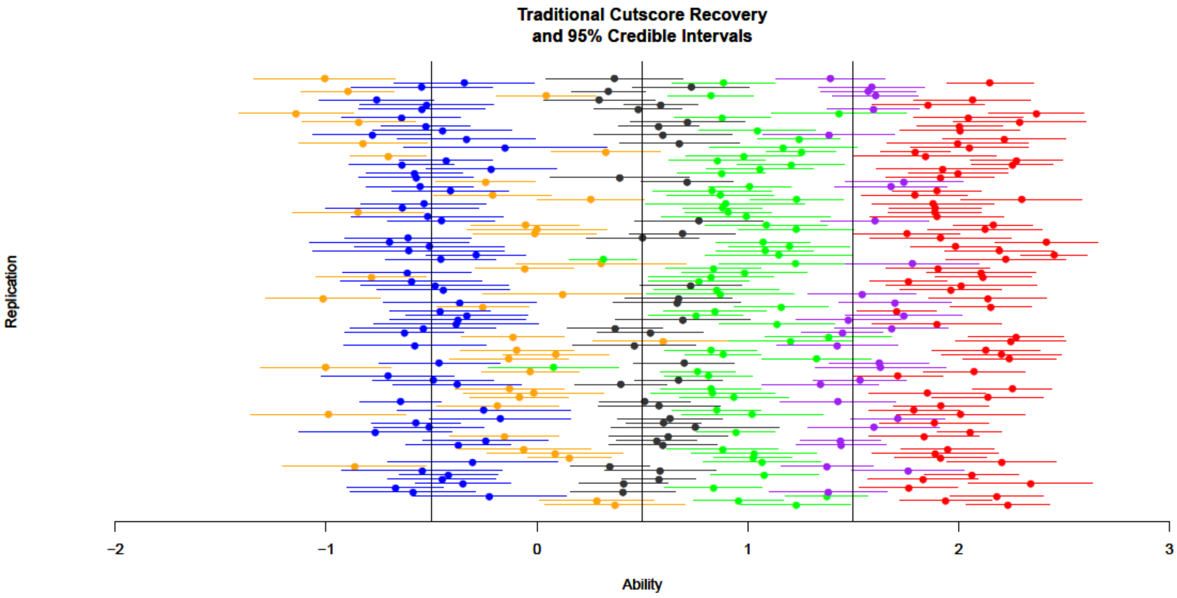

Figure A.67. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

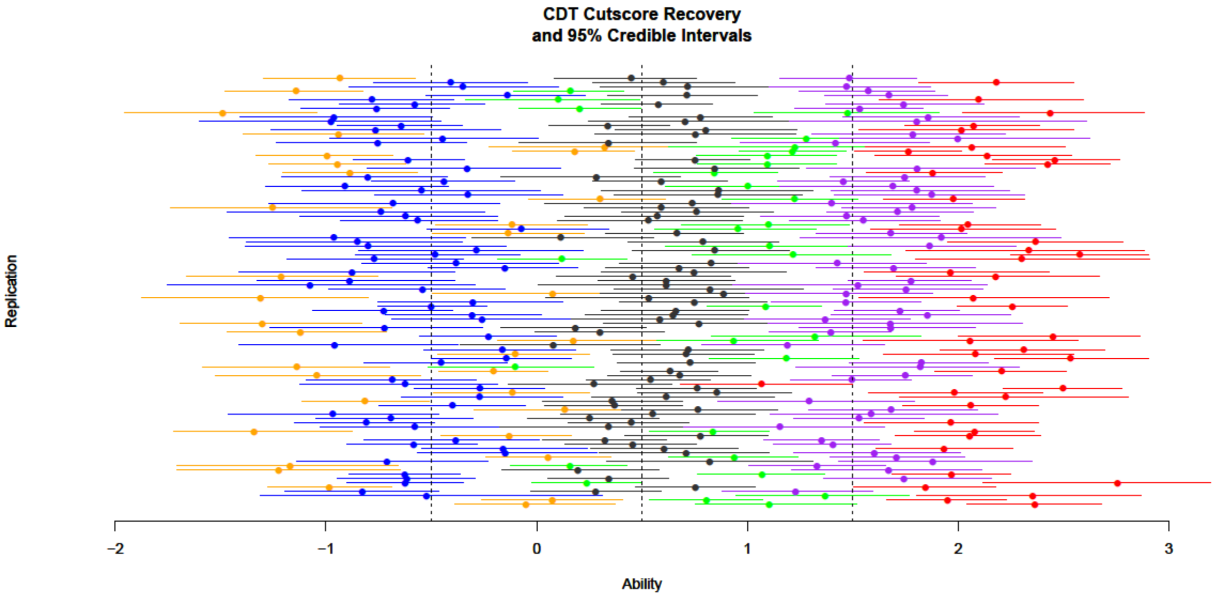

Figure A.68. $N_I = 8$, $N_G = 3$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

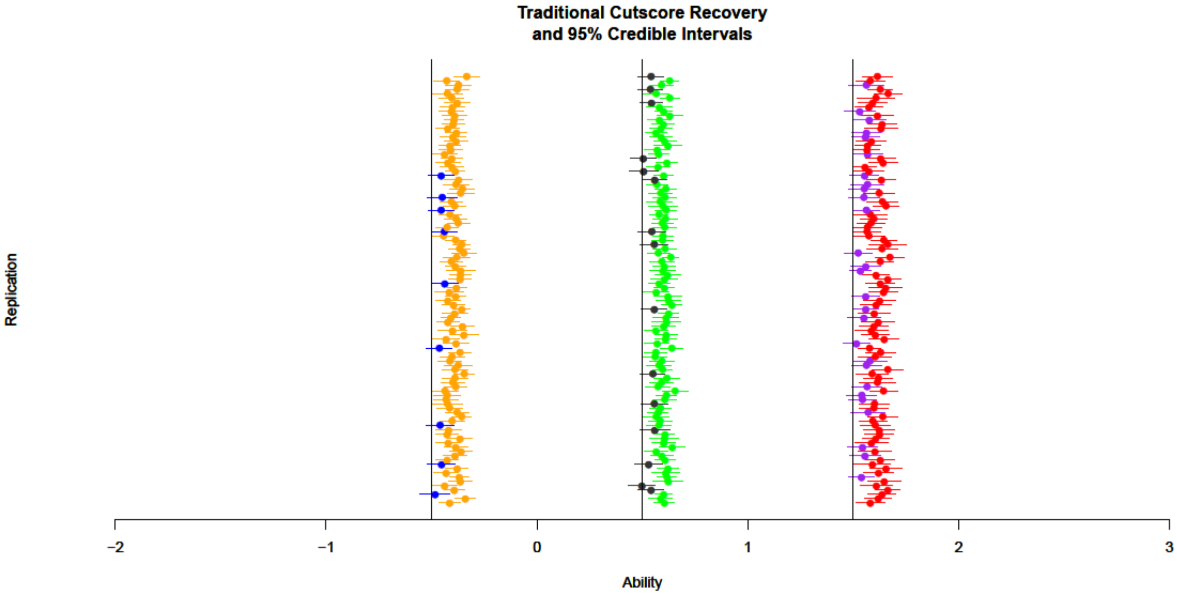

Figure A.69. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

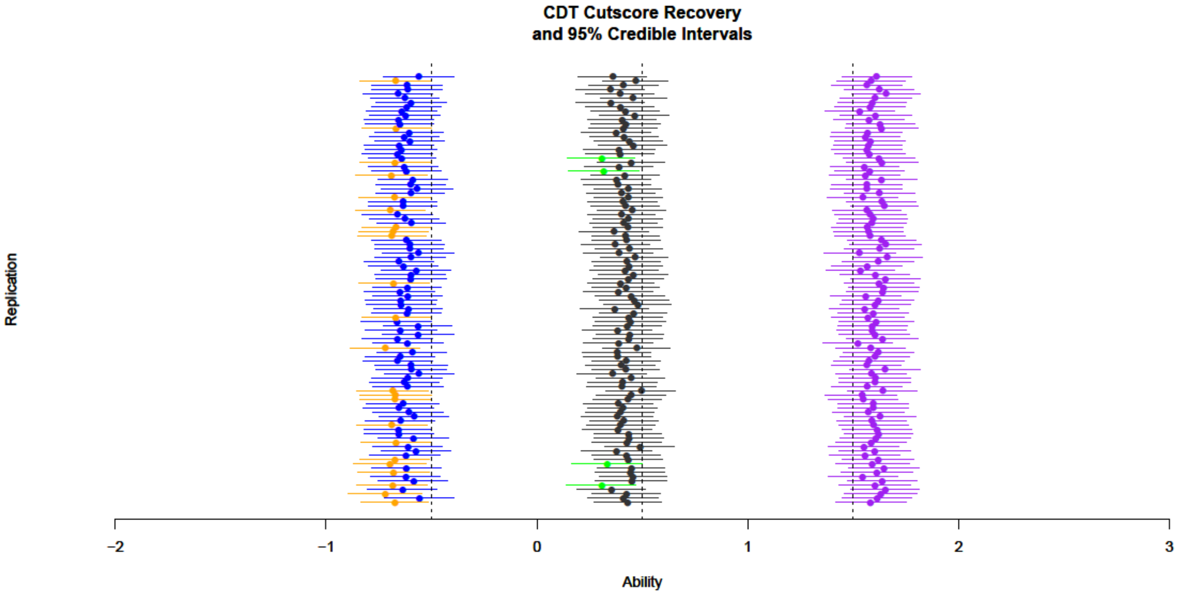

Figure A.70. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

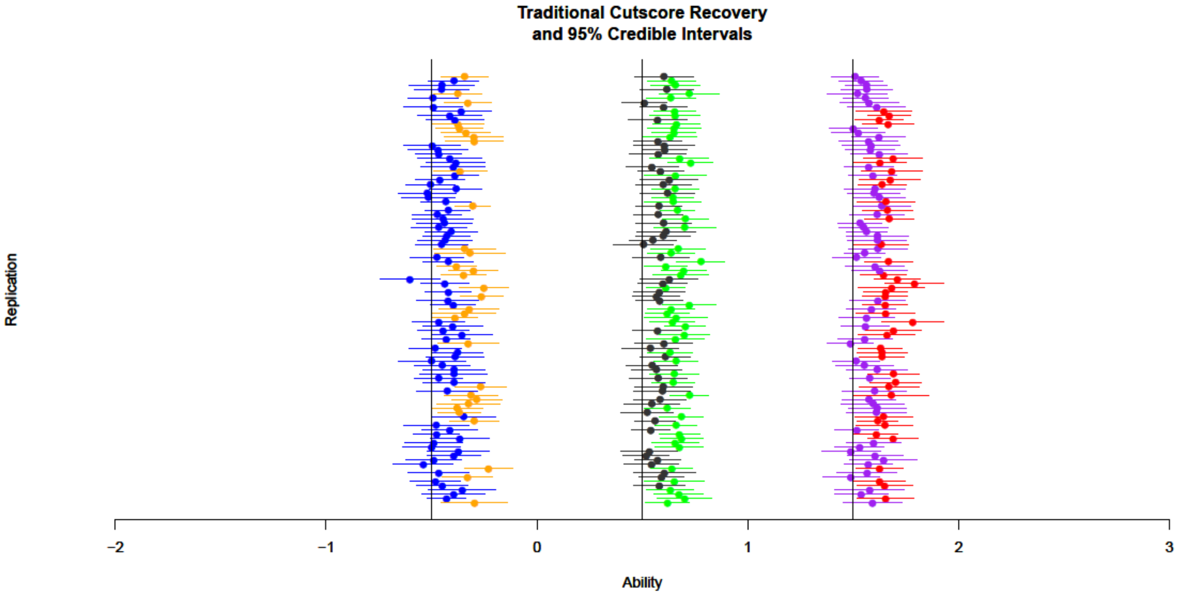

Figure A.71. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

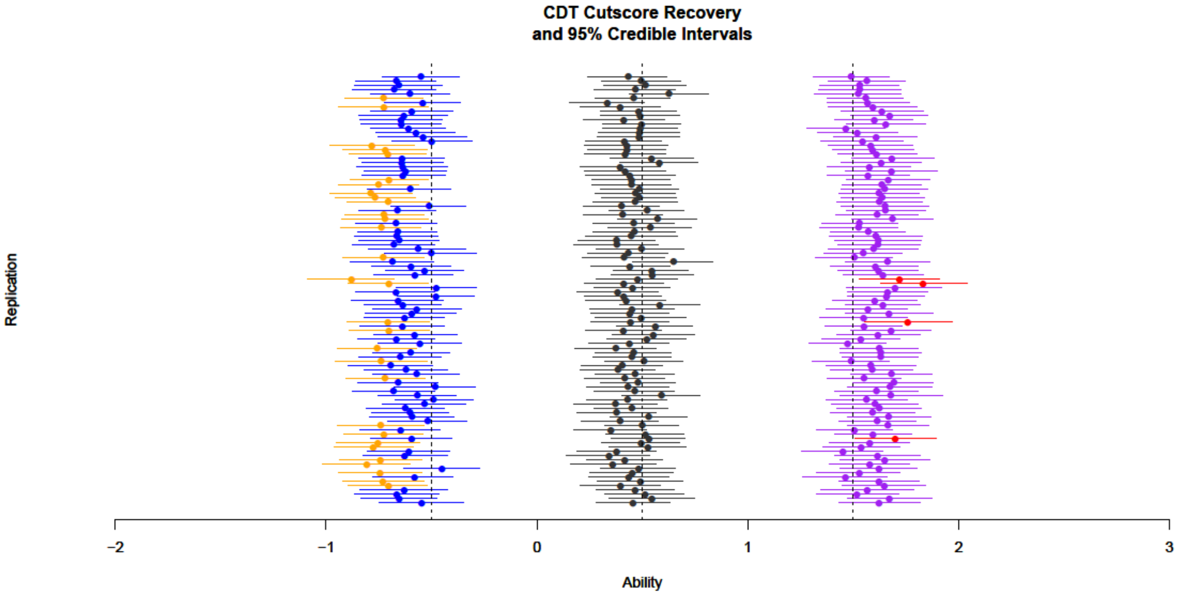

Figure A.72. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

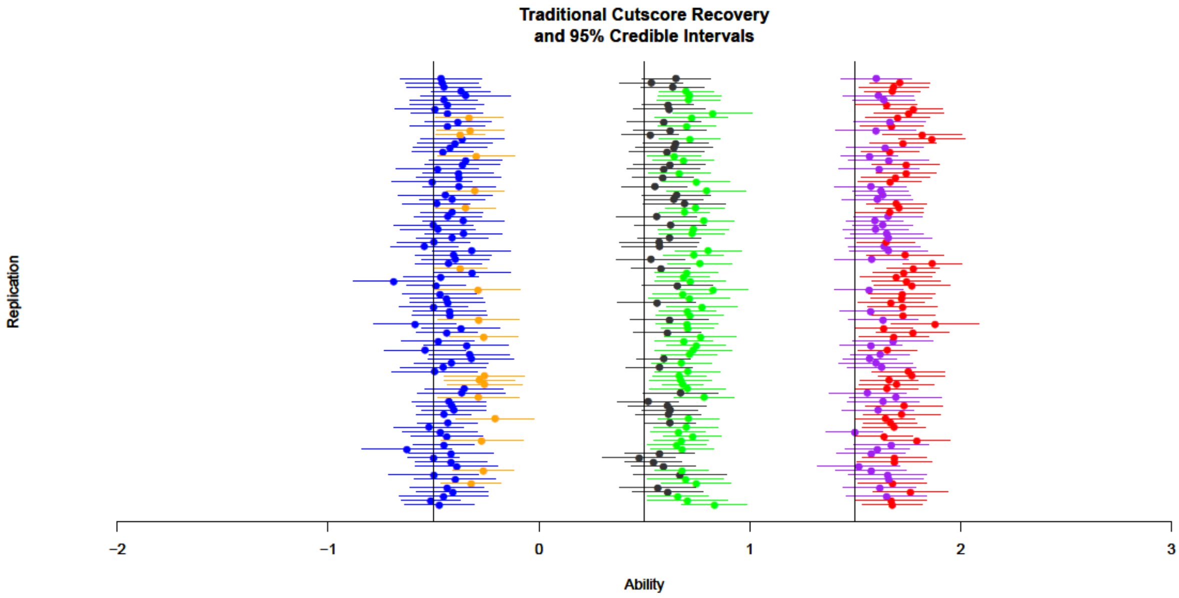

Figure A.73. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

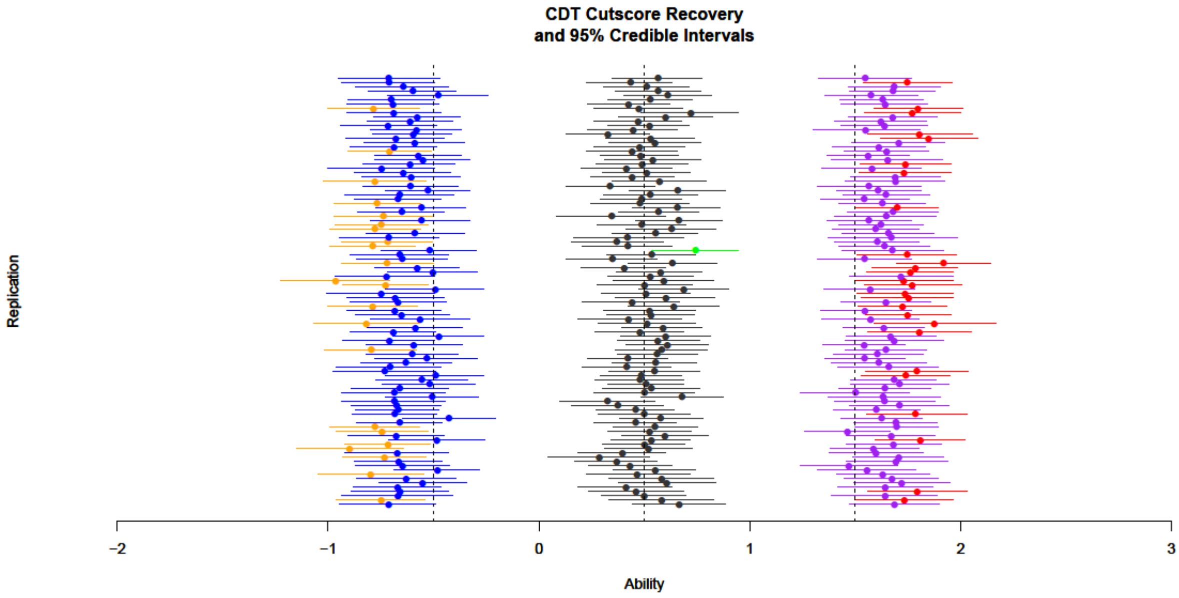

Figure A.74. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

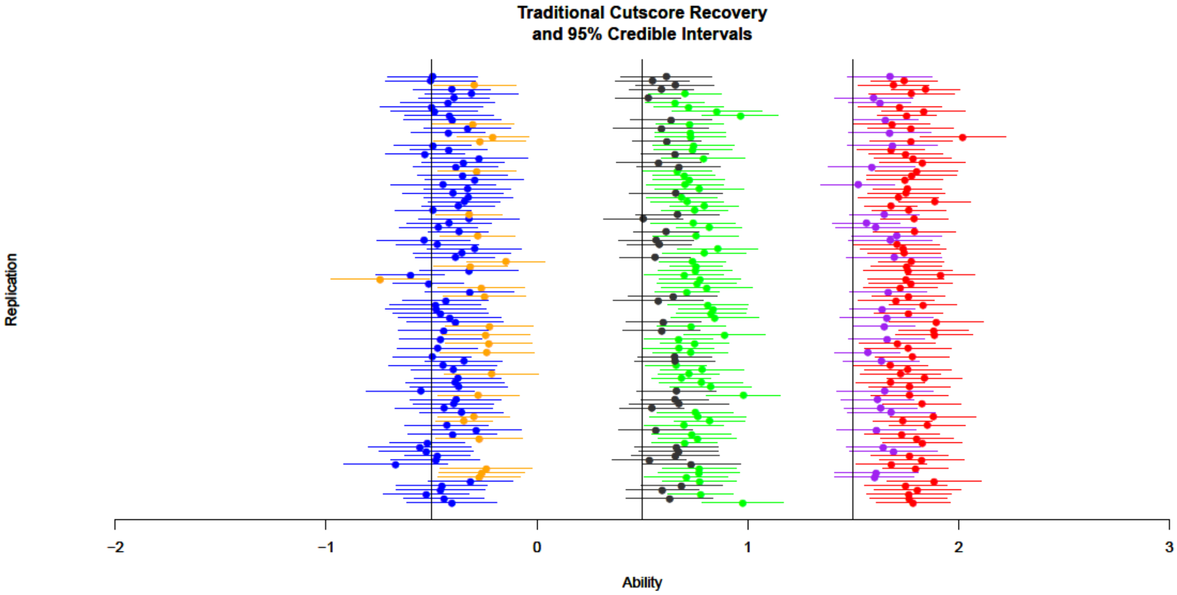

Figure A.75. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

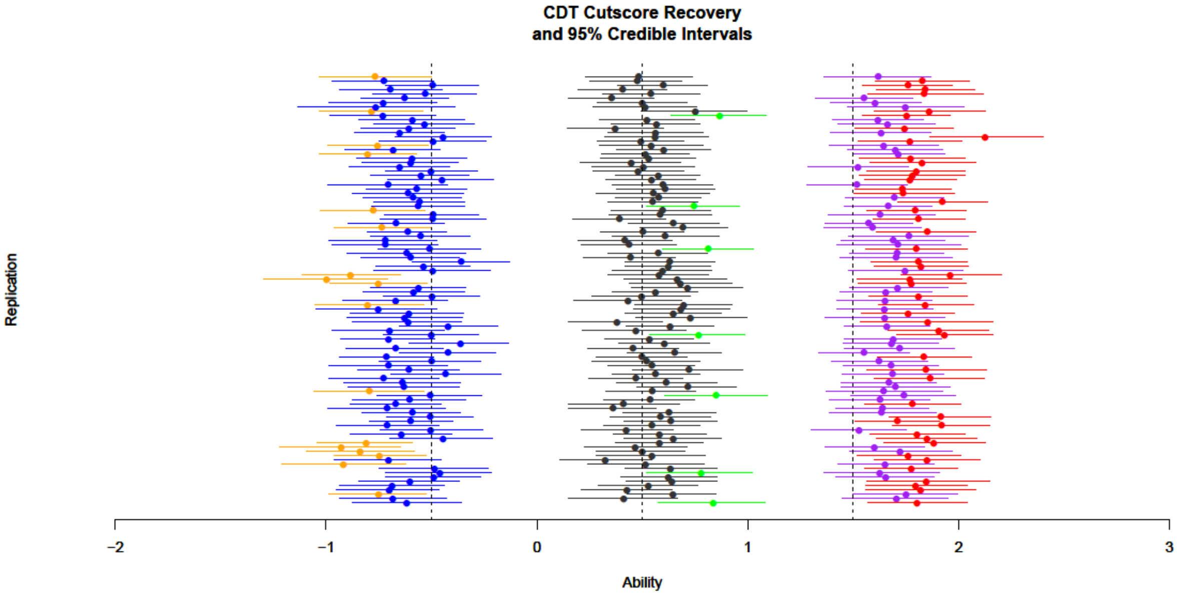

Figure A.76. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

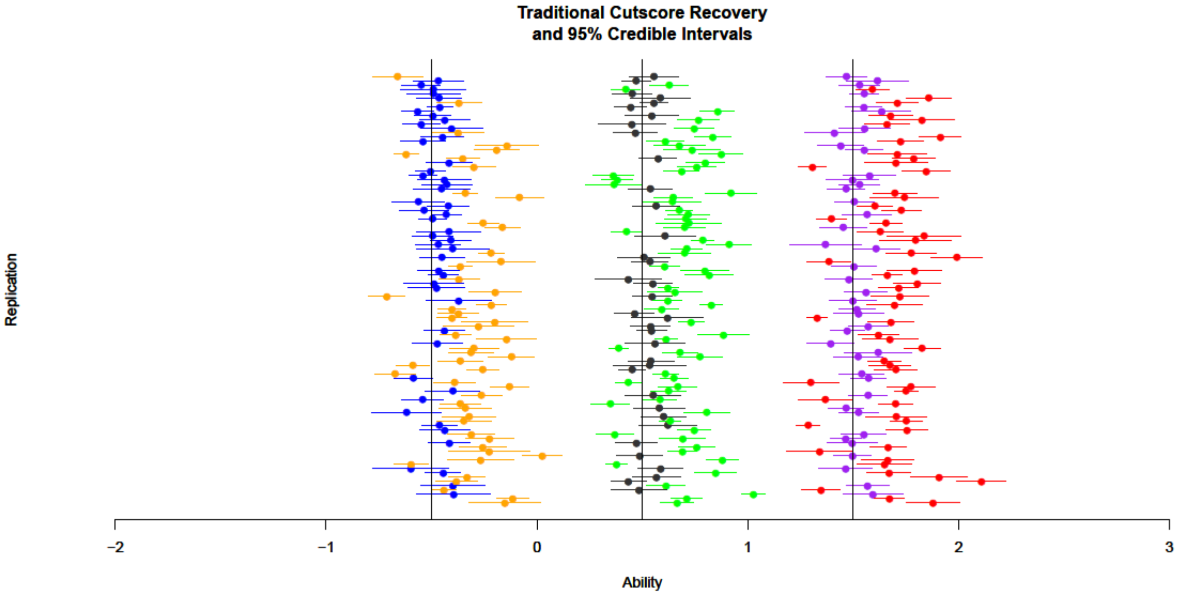

Figure A.77. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

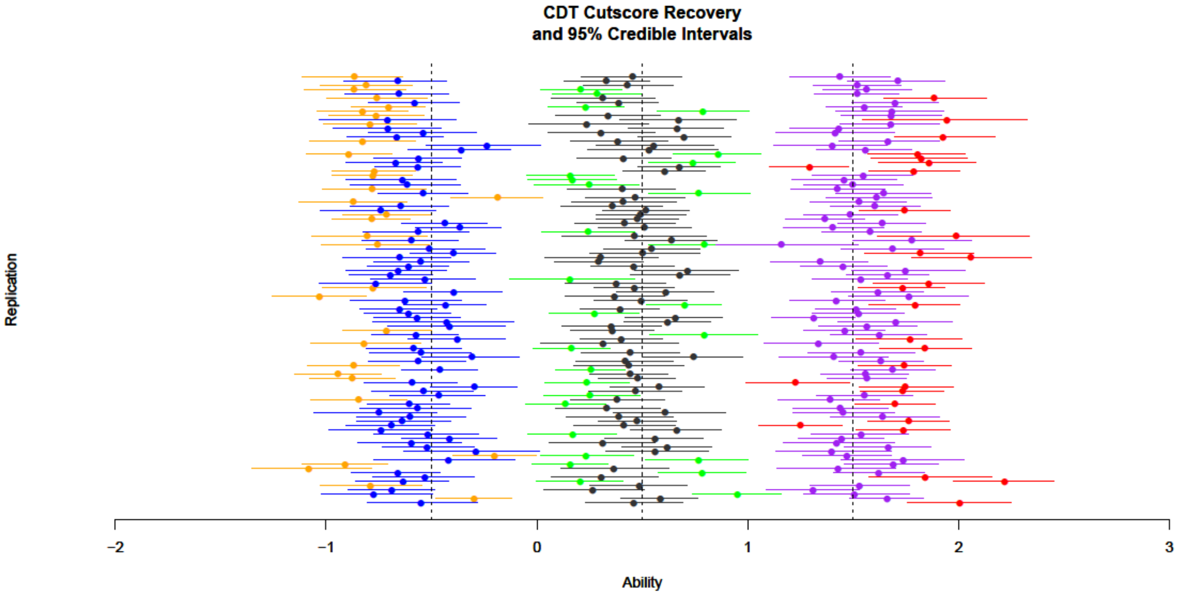

Figure A.78. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

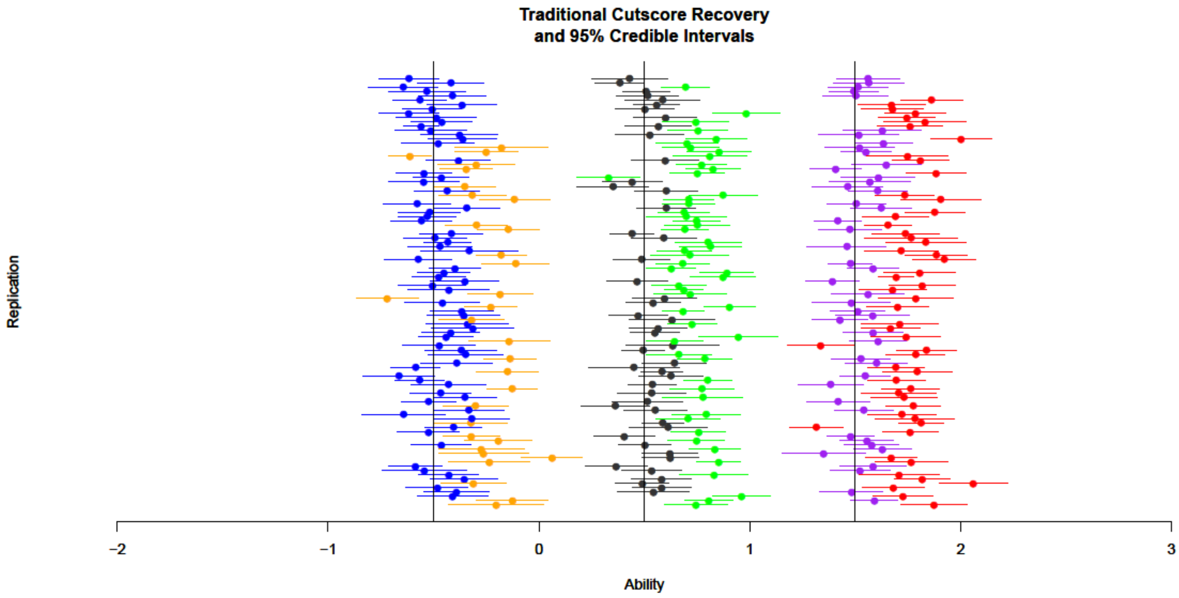

Figure A.79. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

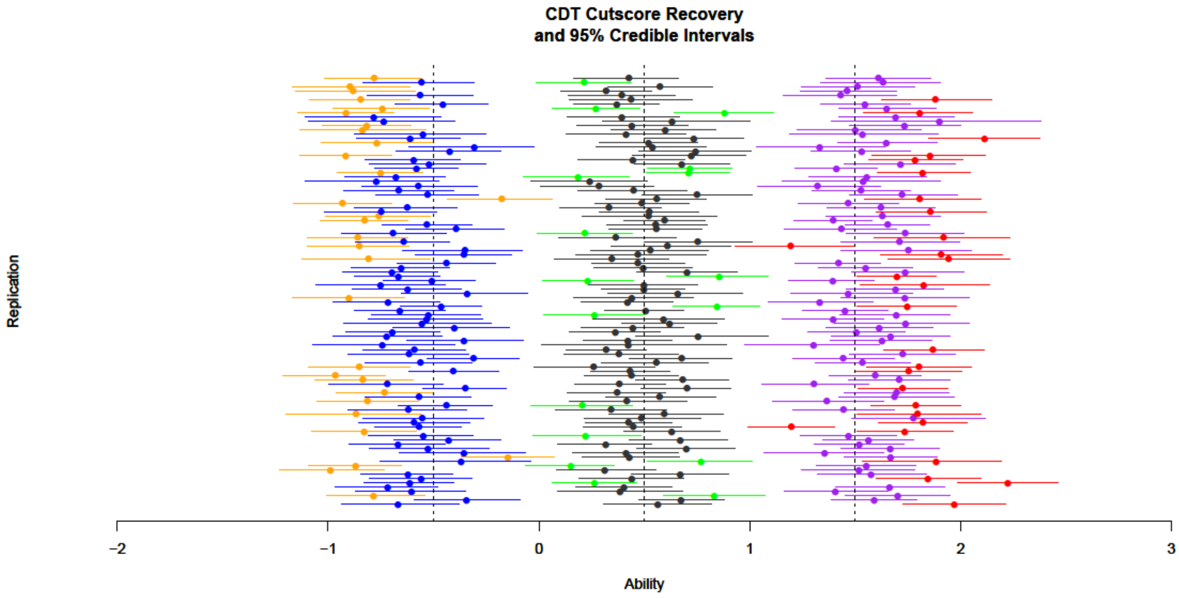

Figure A.80. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

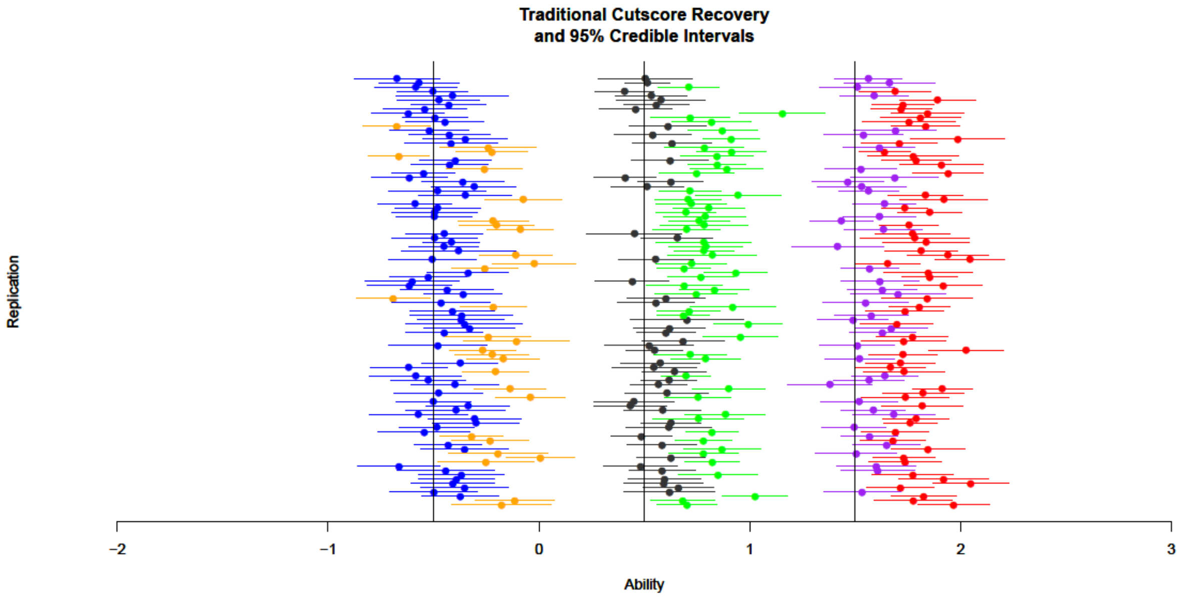

Figure A.81. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

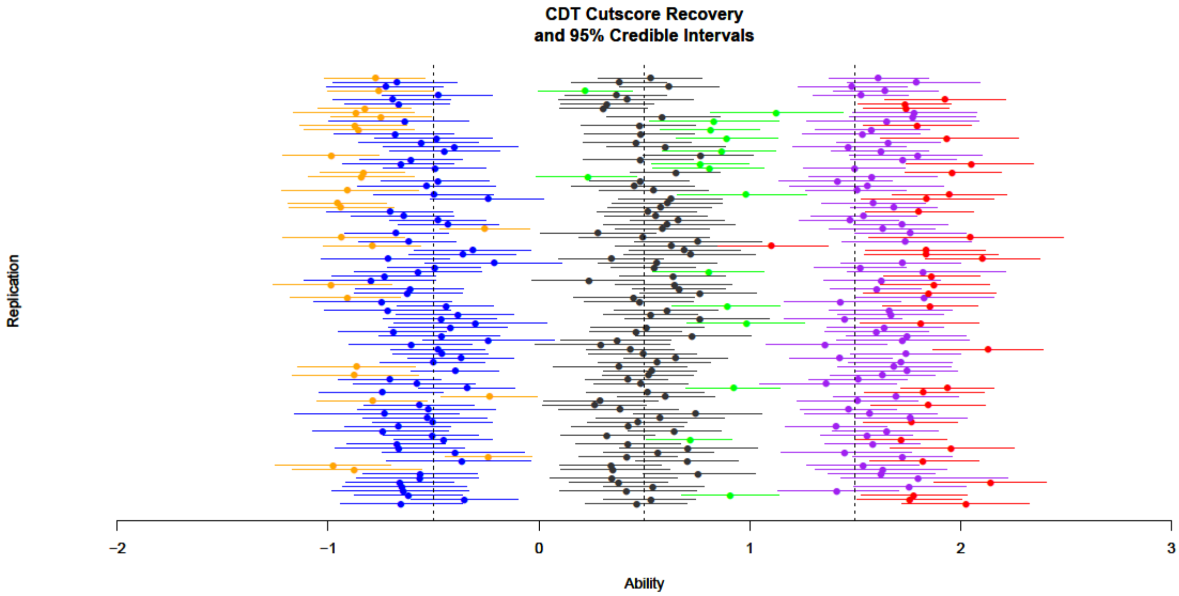

Figure A.82. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

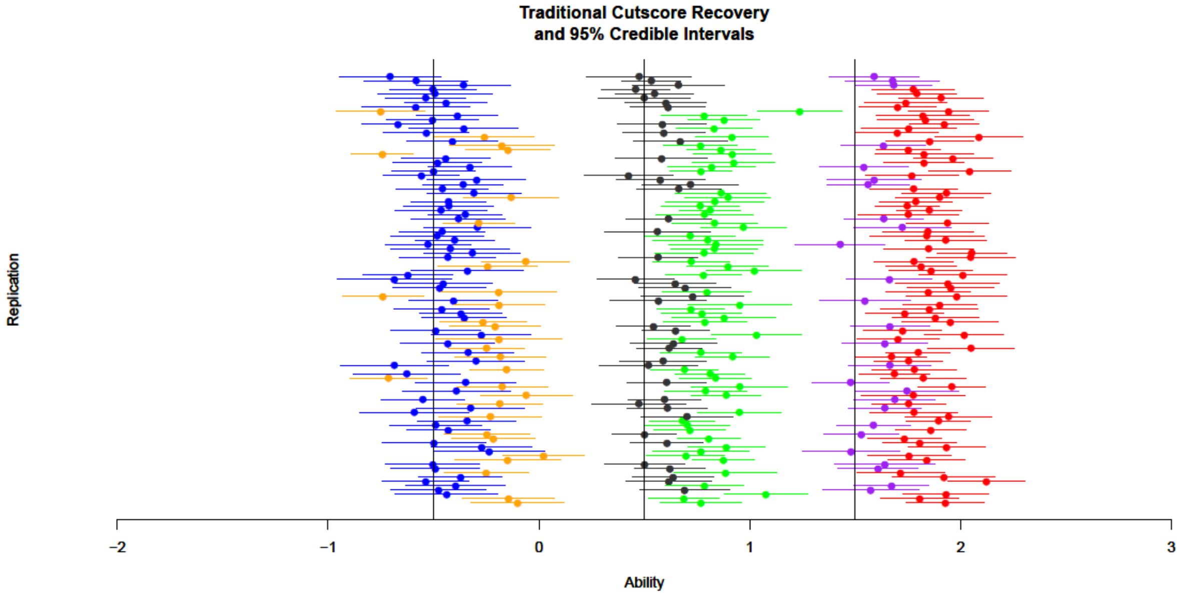

Figure A.83. N$_I$ = 8, N$_G$ = 4, σ^2_B = 0.1, σ^2_C = 0.3; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

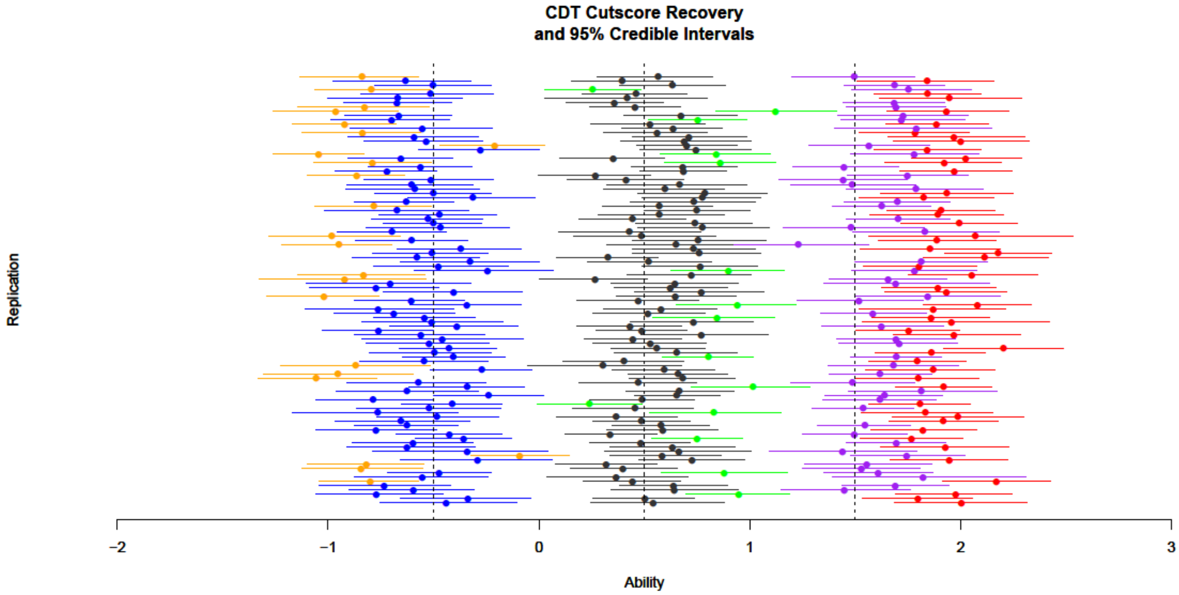

Figure A.84. N$_I$ = 8, N$_G$ = 4, σ^2_B = 0.1, σ^2_C = 0.3; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

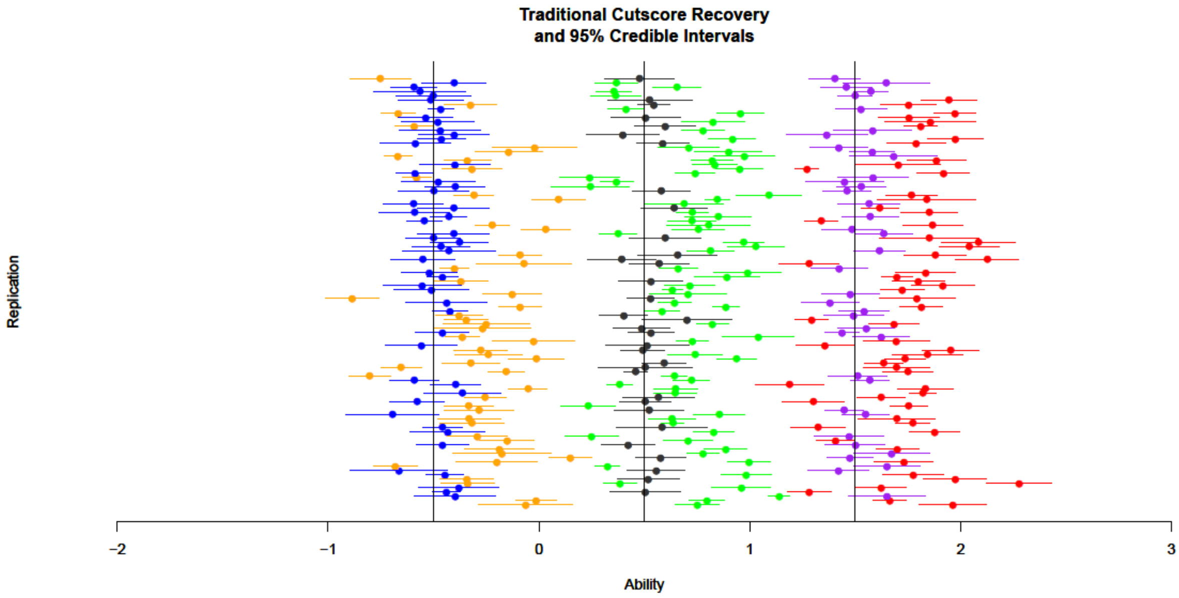

Figure A.85. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

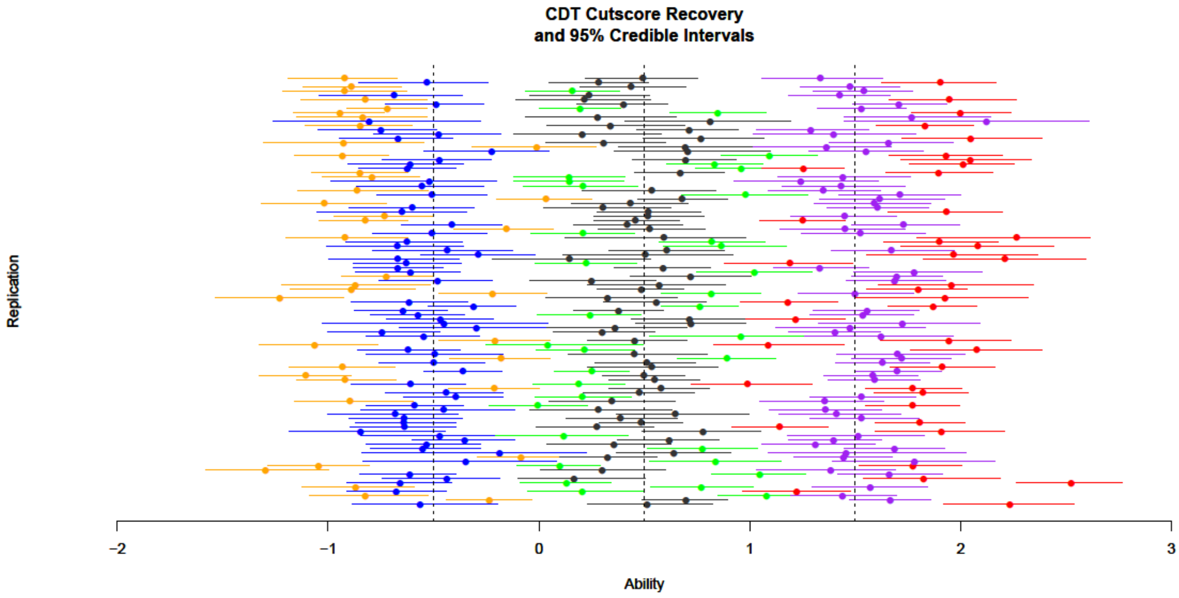

Figure A.86. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

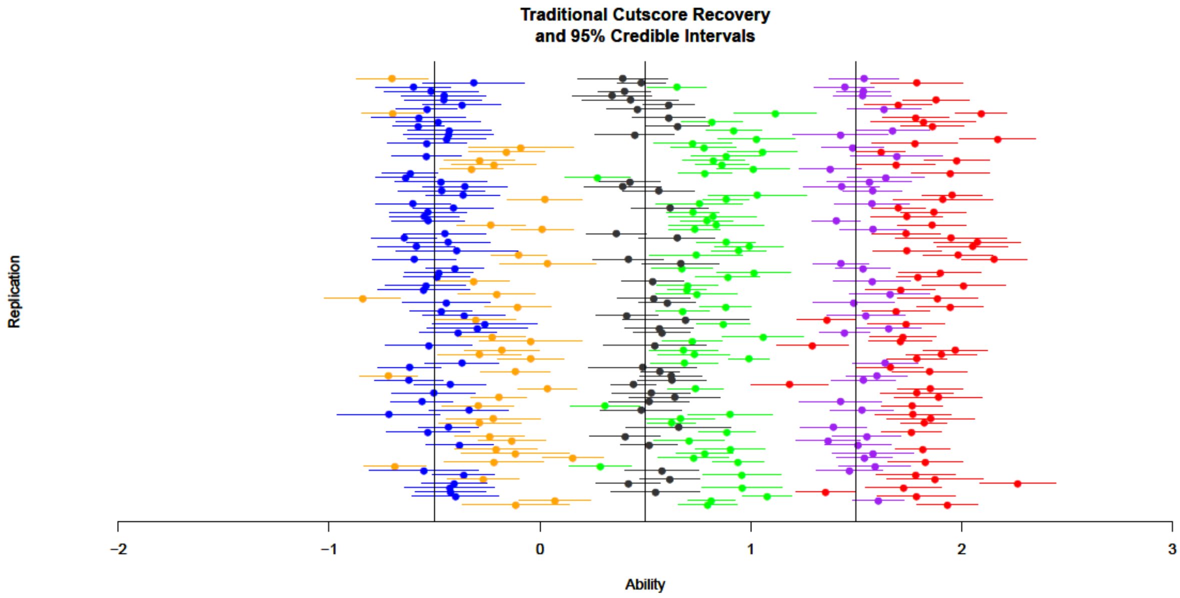

Figure A.87. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

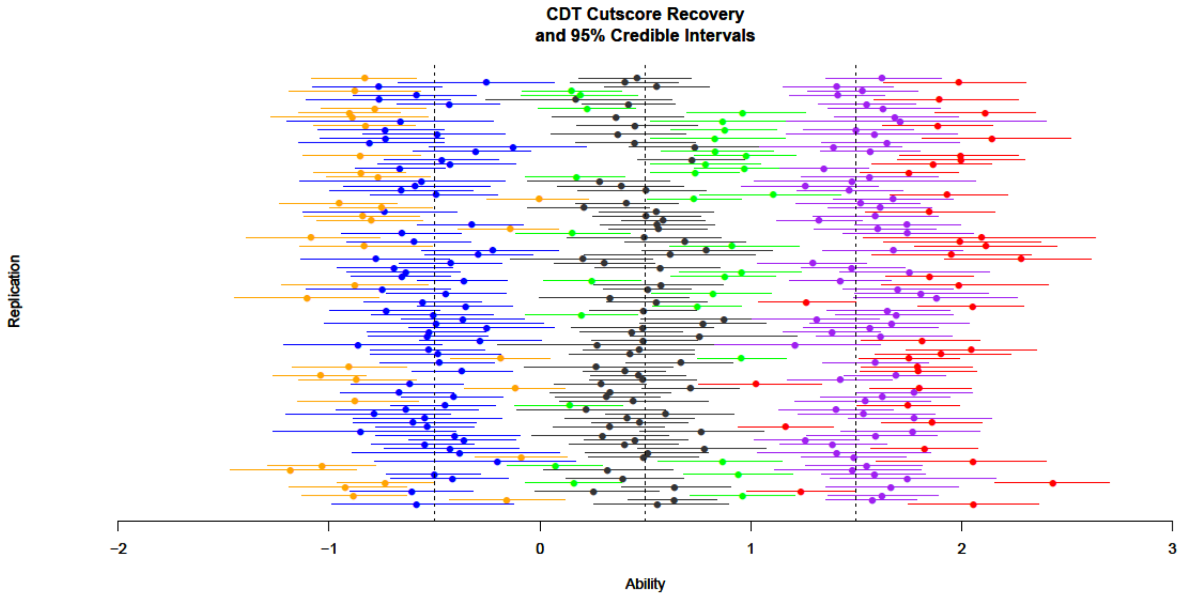

Figure A.88. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

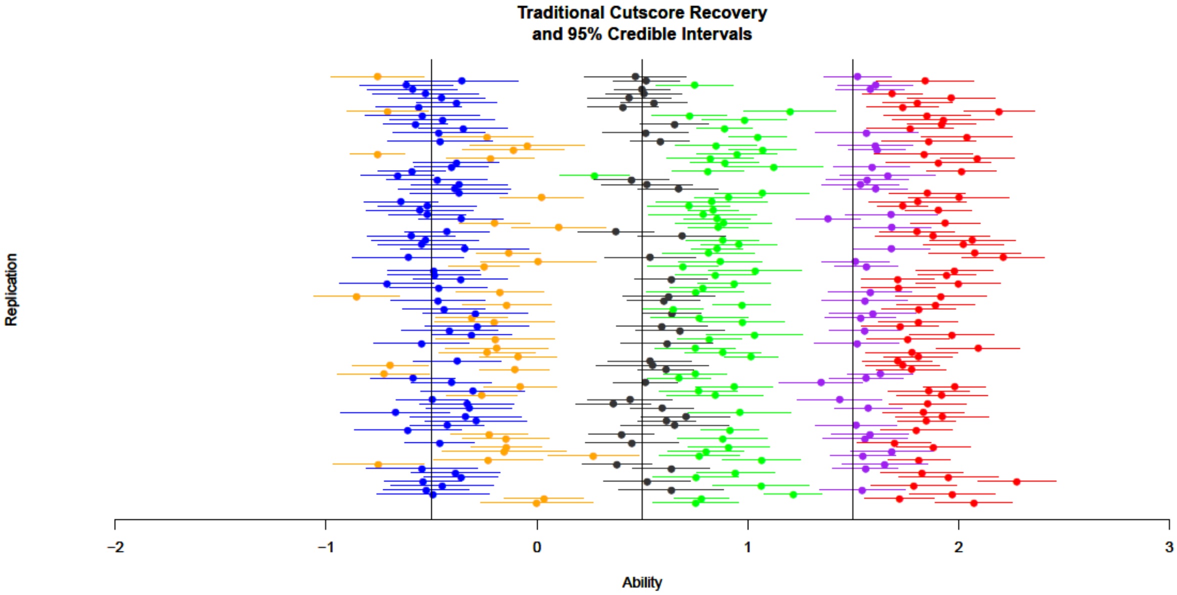

Figure A.89. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

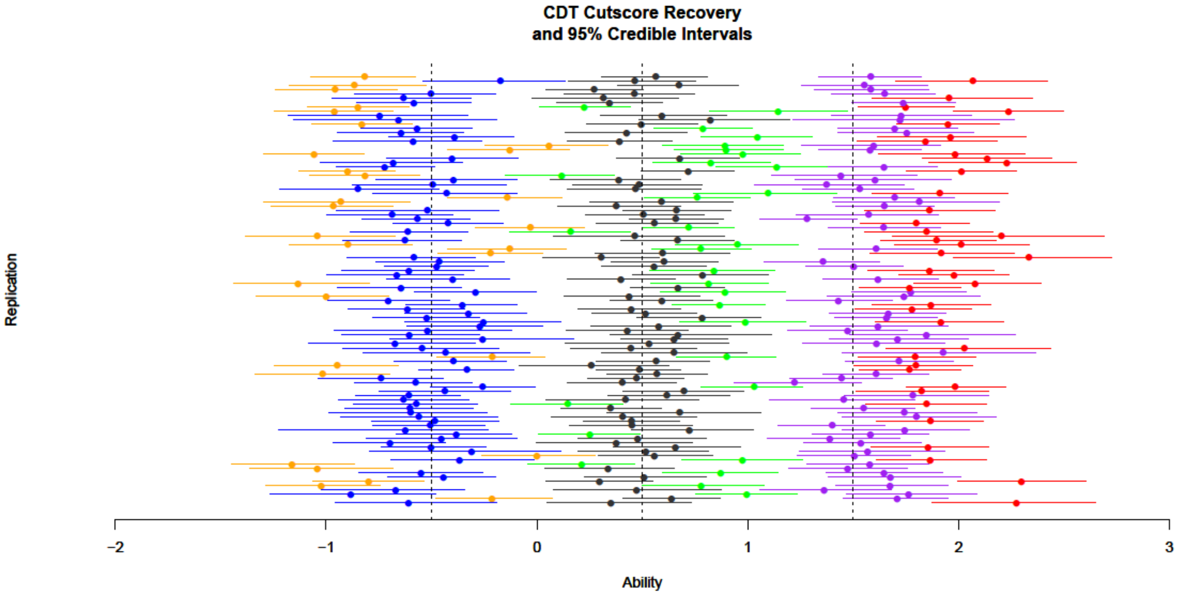

Figure A.90. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

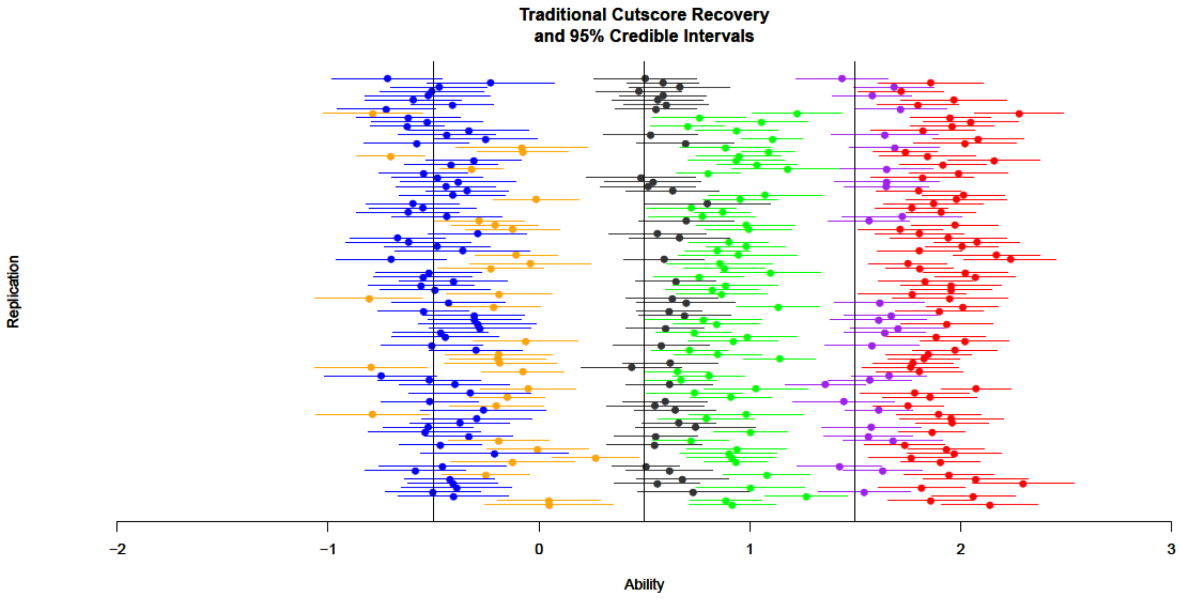

Figure A.91. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

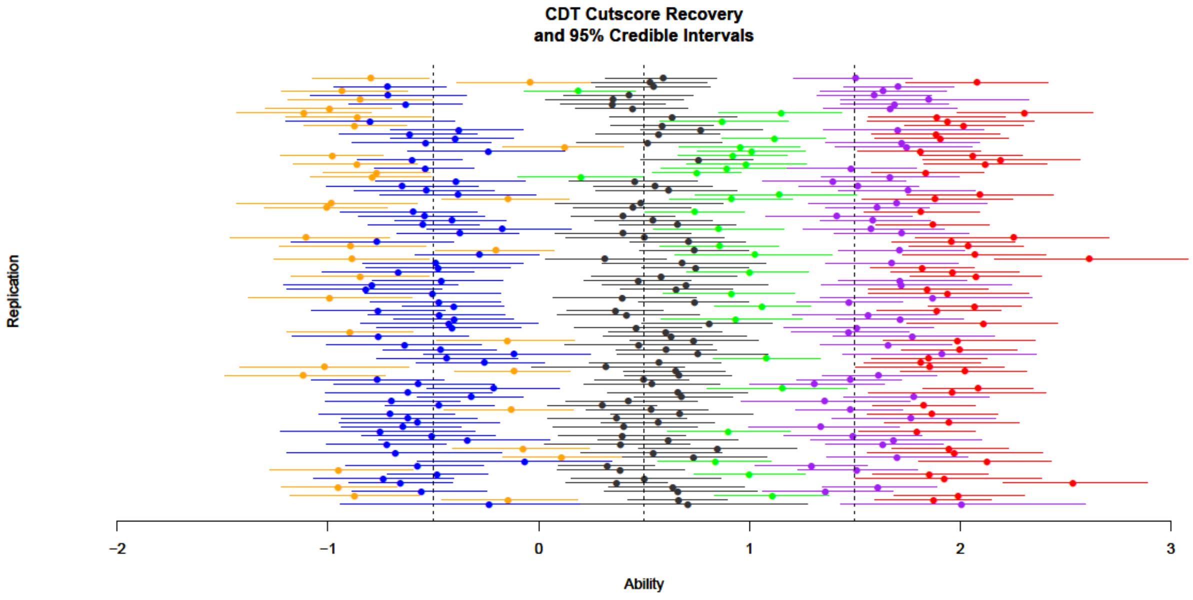

Figure A.92. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

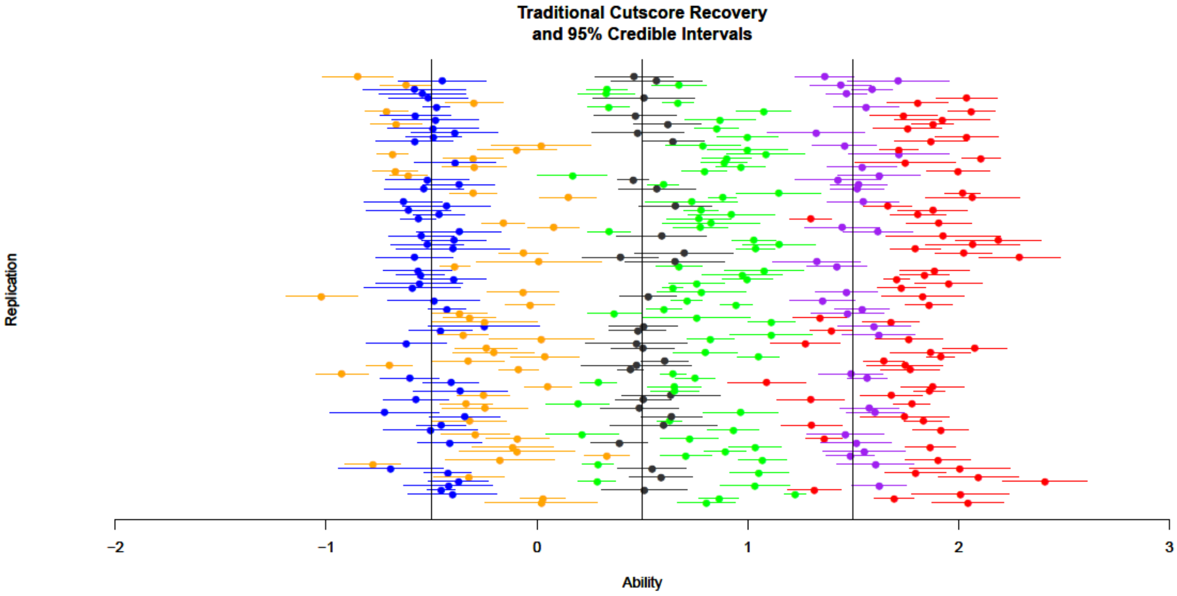

Figure A.93. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

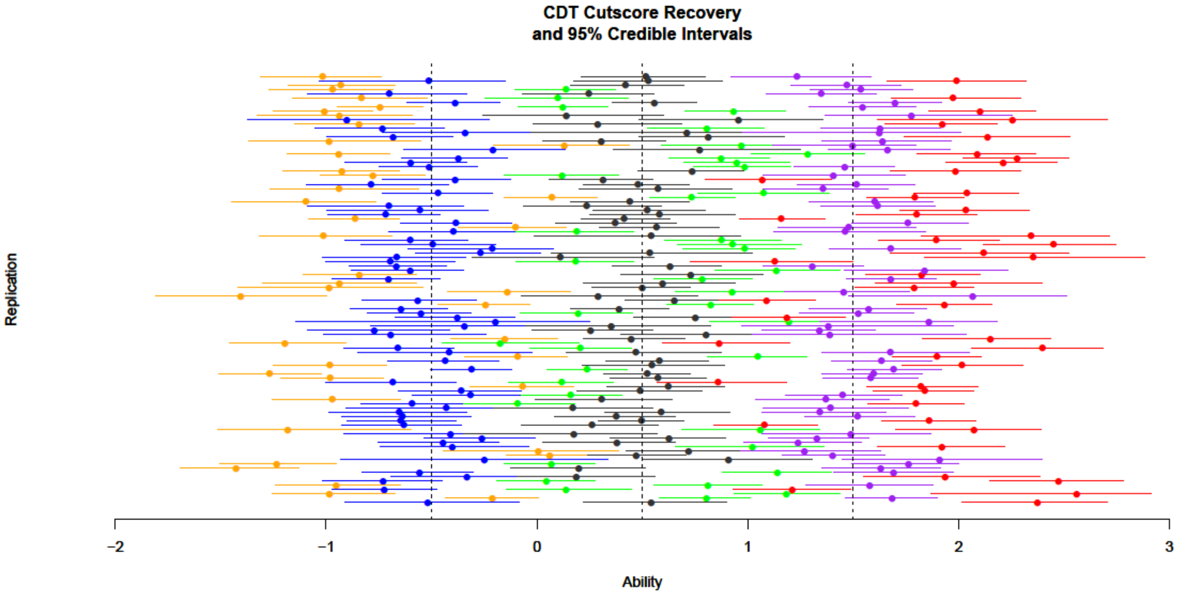

Figure A.94. $N_I = 8$, $N_G = 4$ $\sigma^2_B = 0.3$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

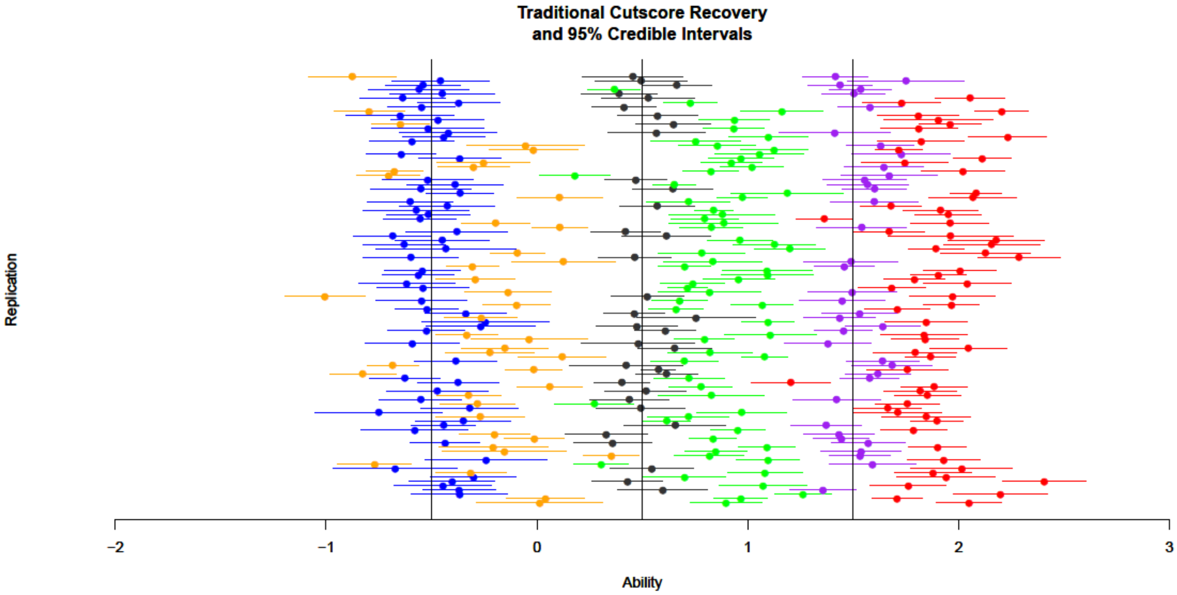

Figure A.95. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

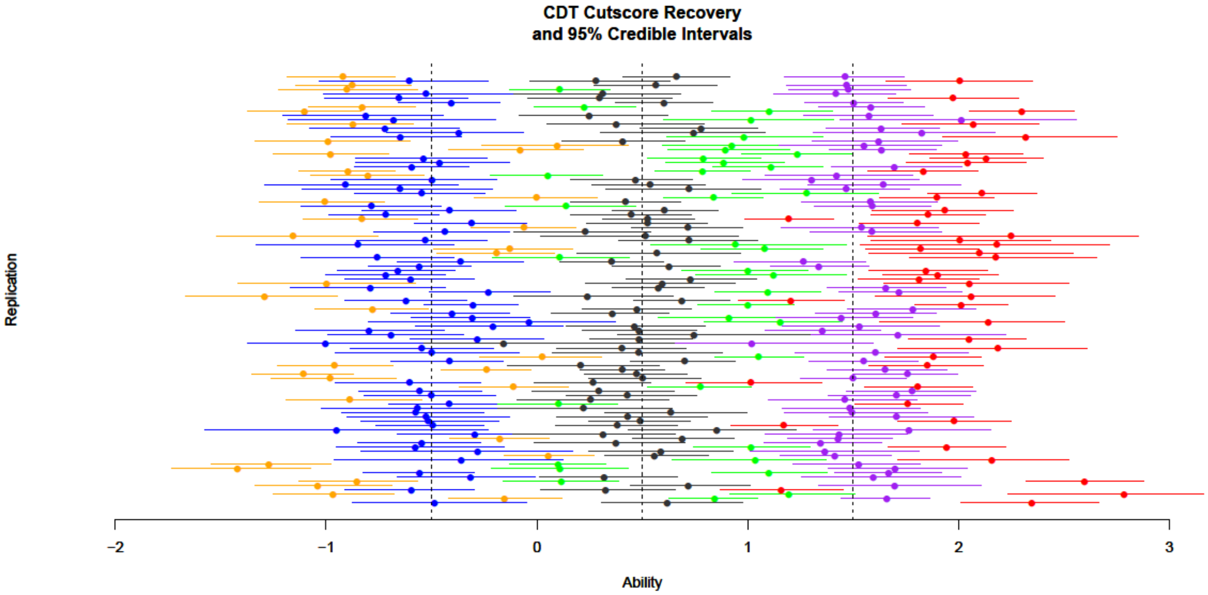

Figure A.96. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

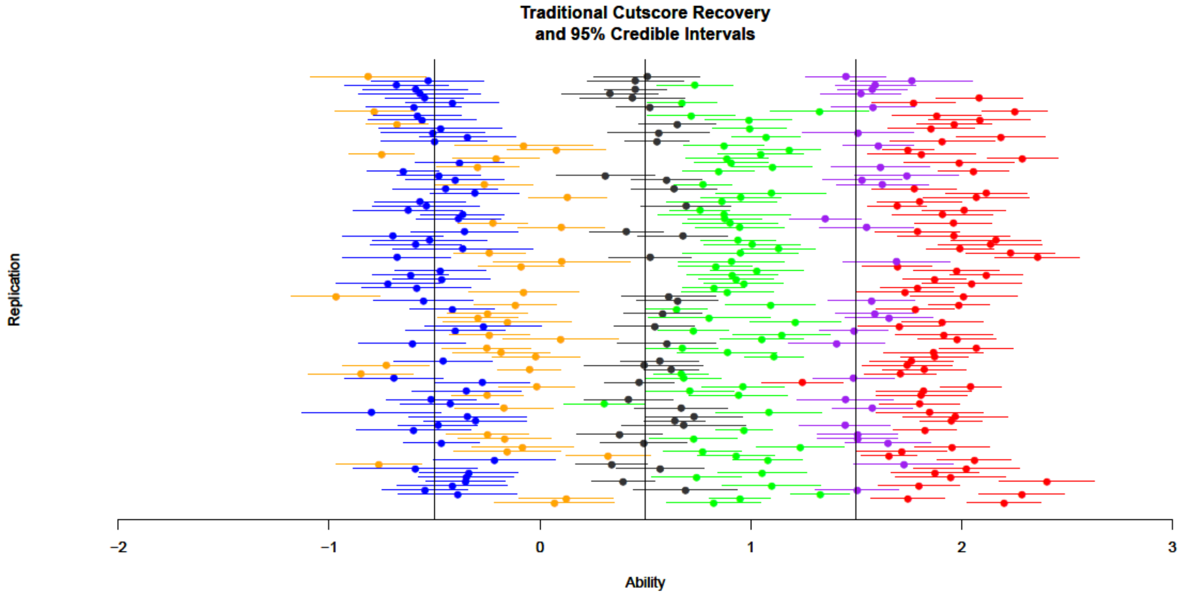

Figure A.97. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

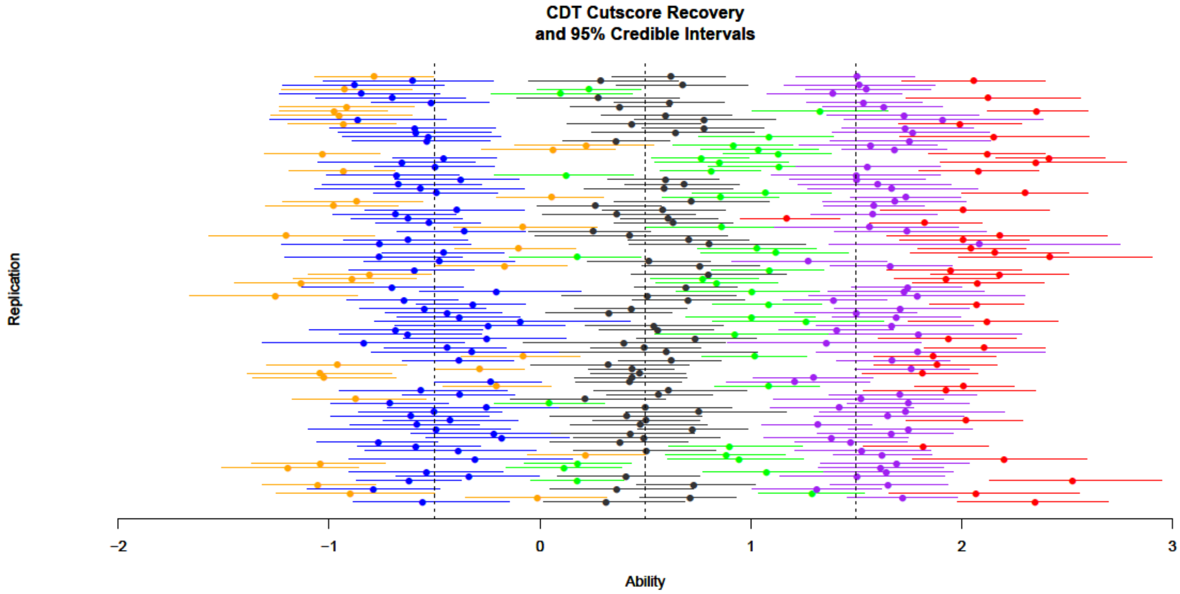

Figure A.98. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

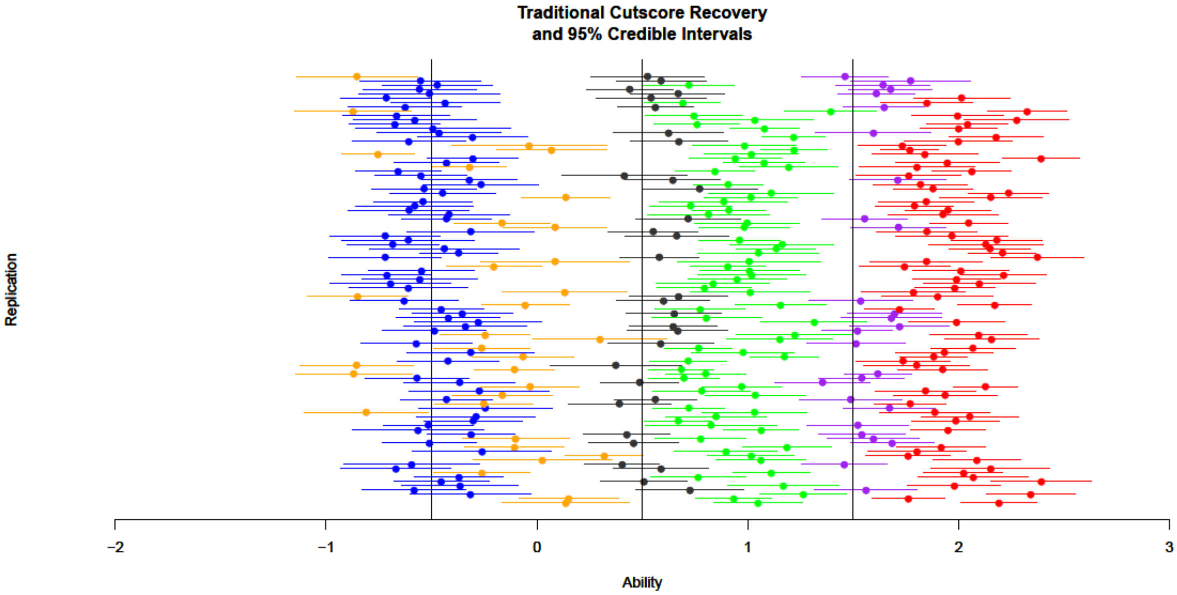

Figure A.99. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

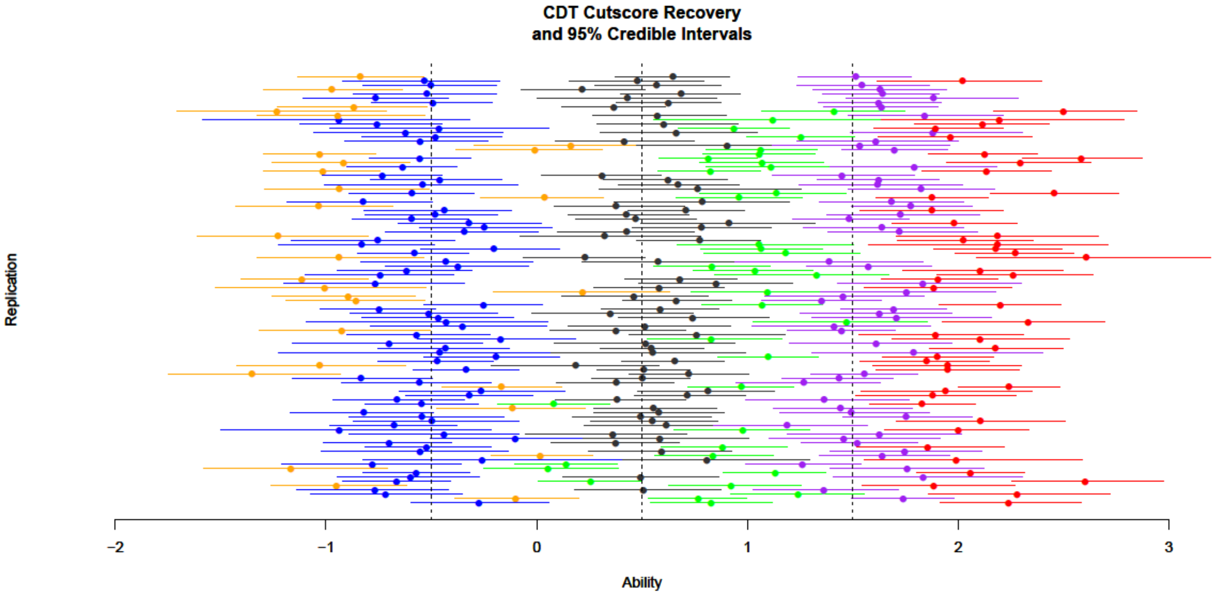

Figure A.100. $N_I = 8$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

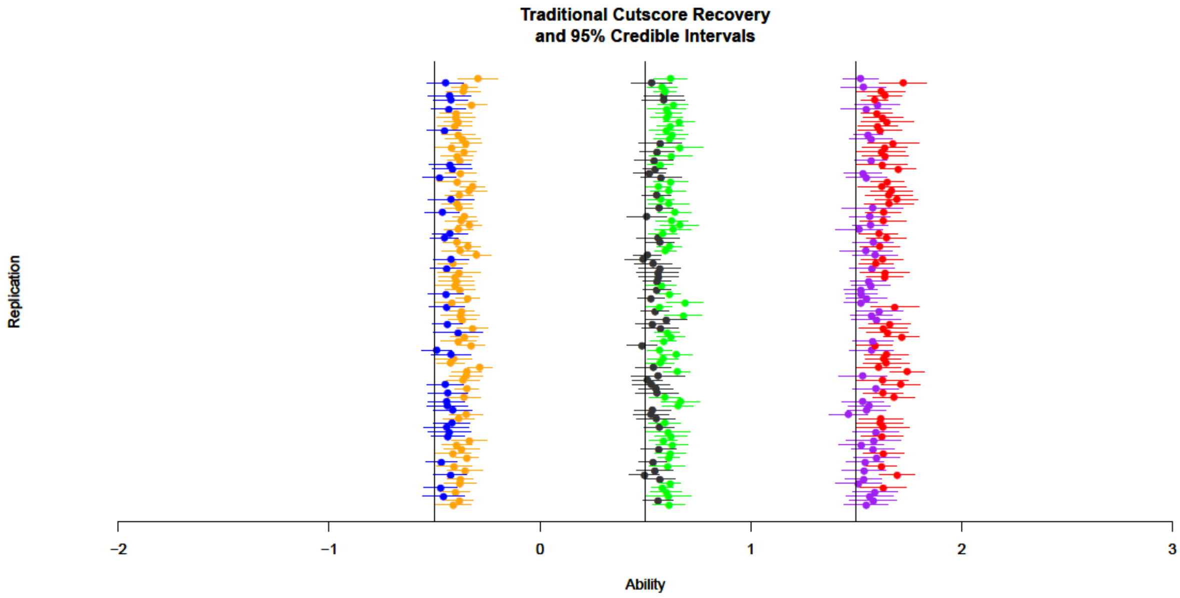

Figure A.101. $N_I = 16$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

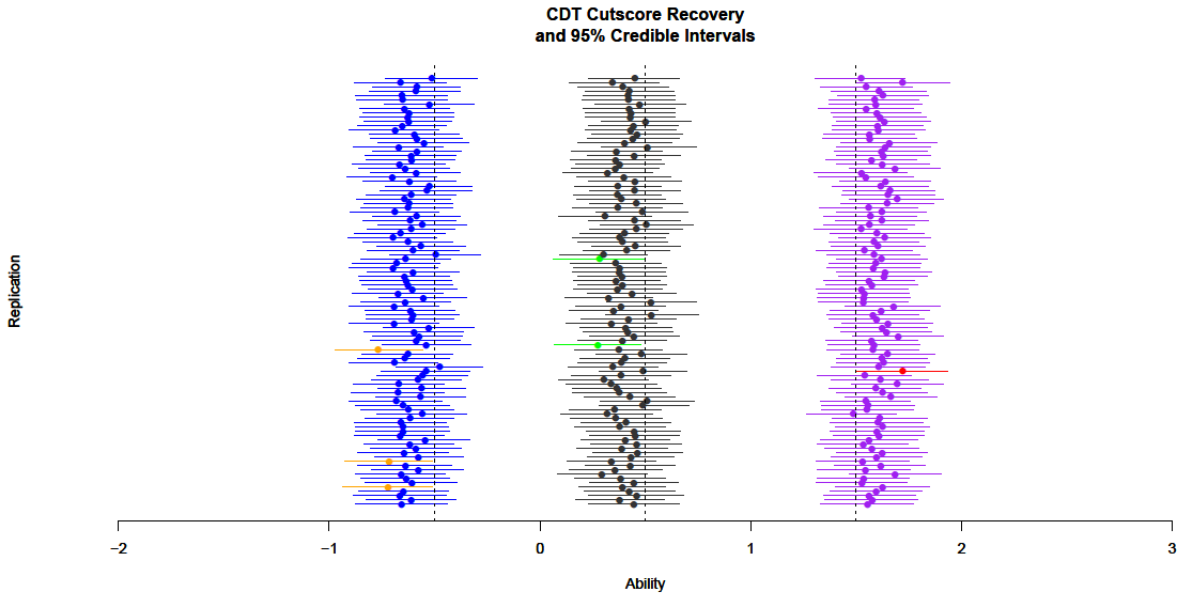

Figure A.102. $N_I = 16$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

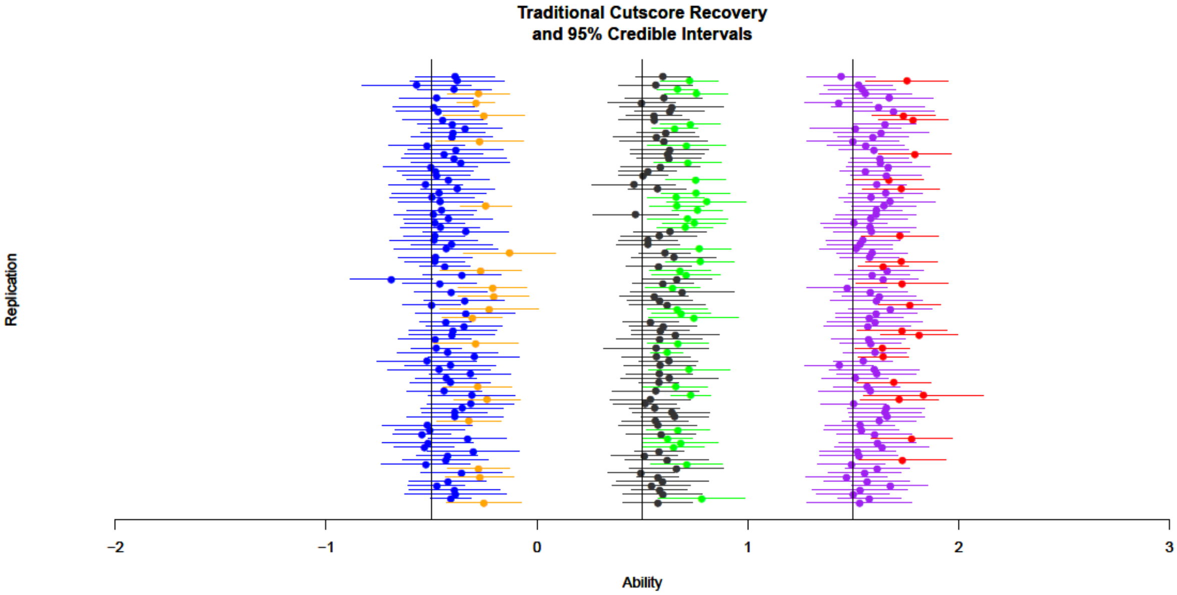

Figure A.103. $N_I = 16$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

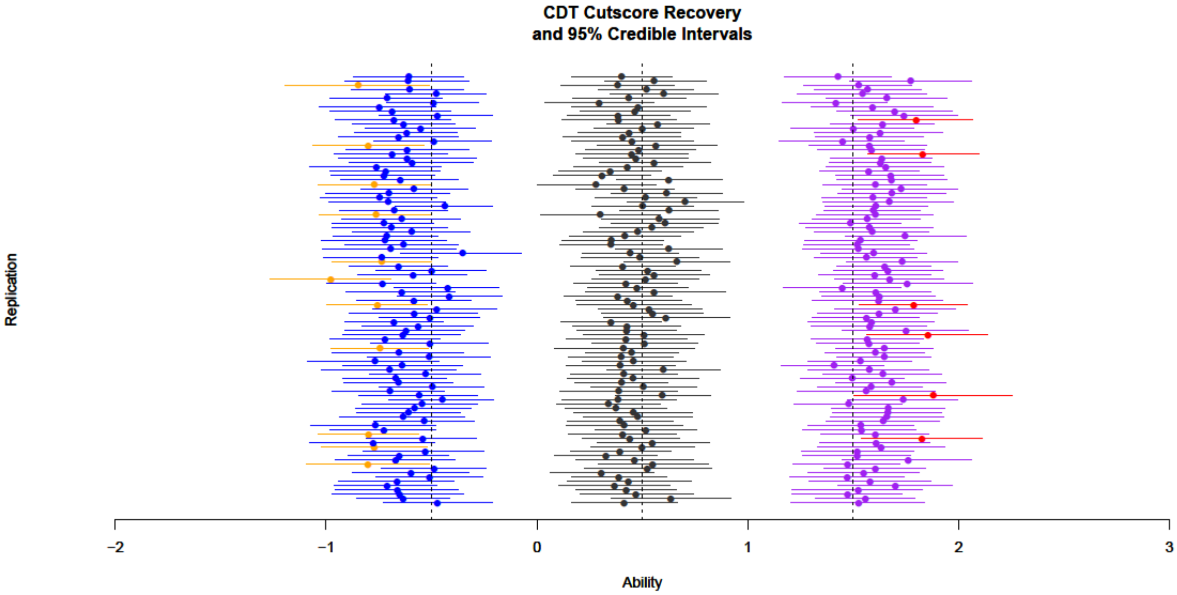

Figure A.104. $N_I = 16$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

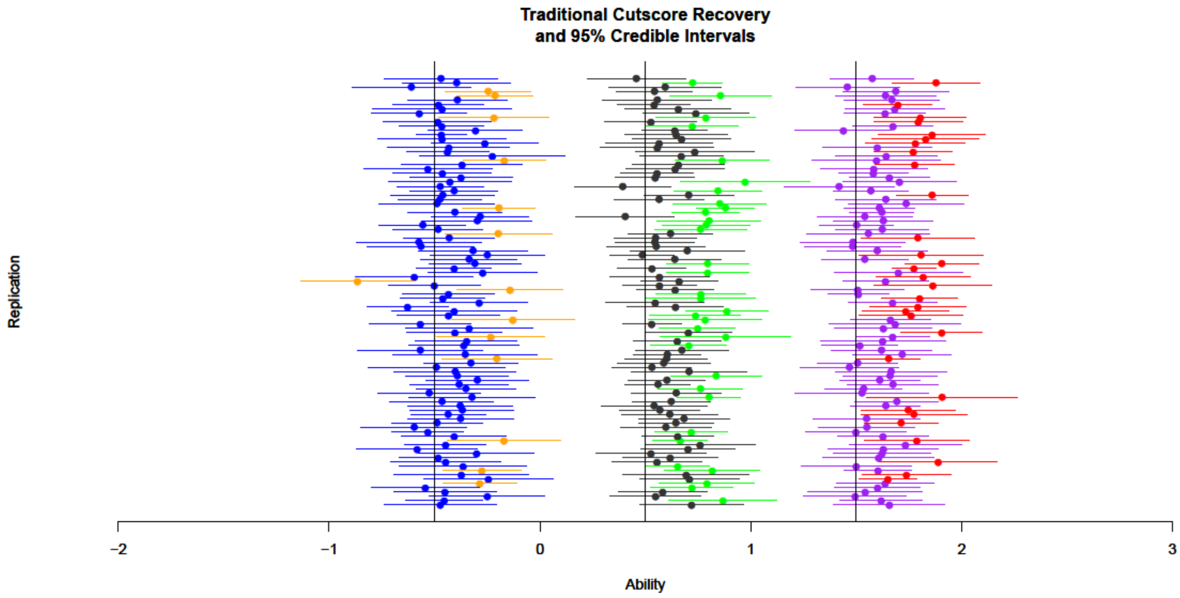

Figure A.105. $N_I = 16$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

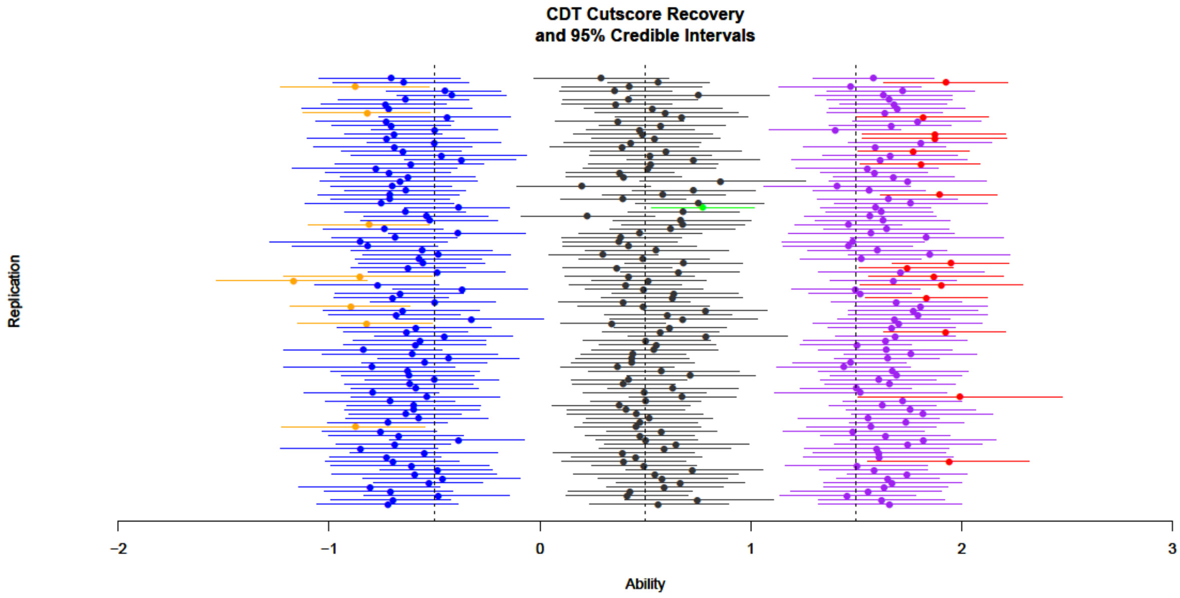

Figure A.106. $N_I = 16$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

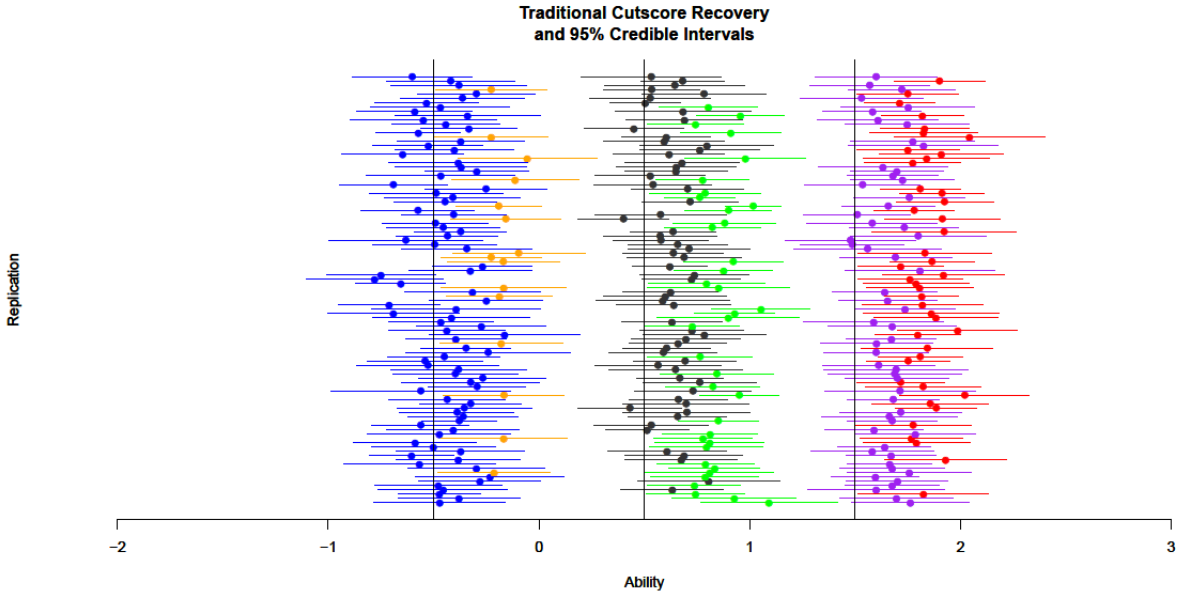

Figure A.107. $N_I = 16$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

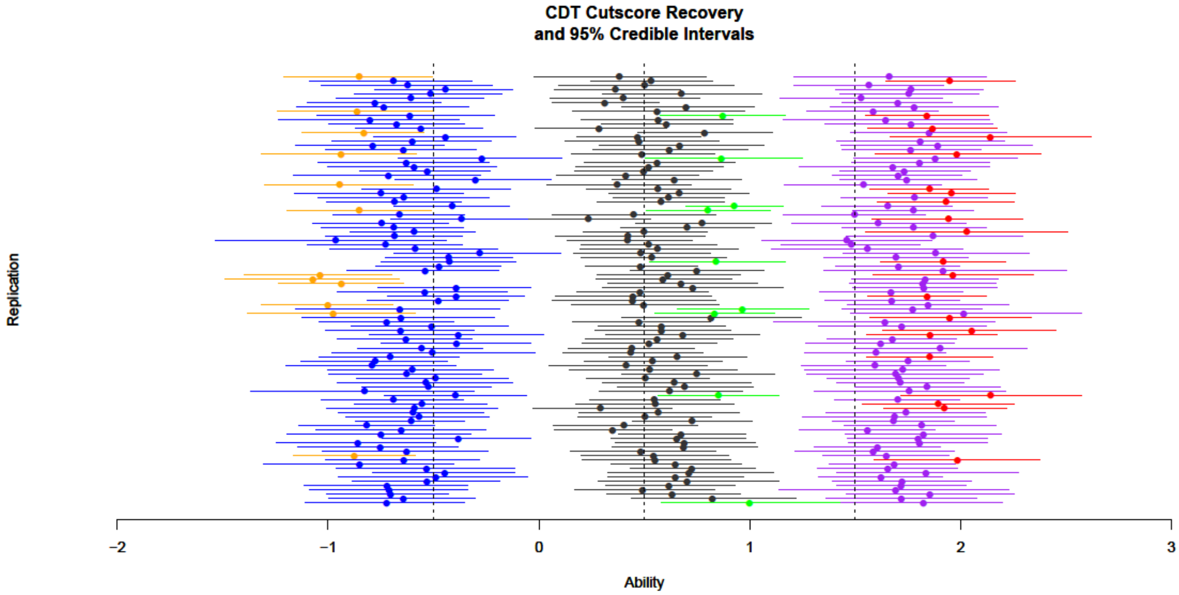

Figure A.108. $N_I = 16$, $N_G = 1$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

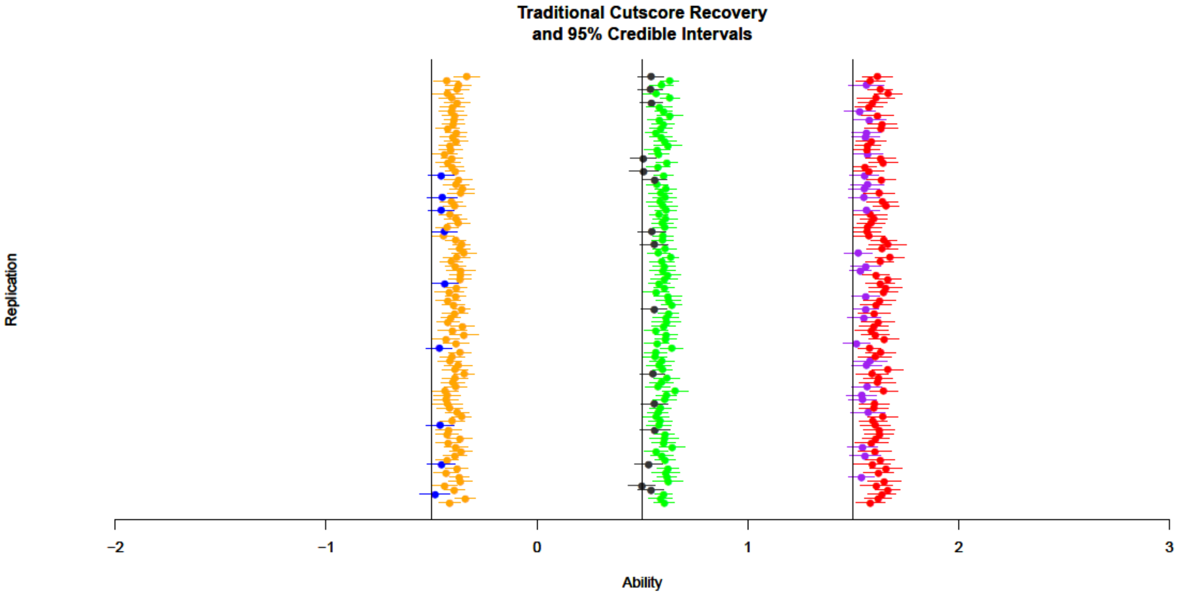

Figure A.109. N_I = 16, N_G = 2, σ^2_B = 0.0, σ^2_C = 0.0; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

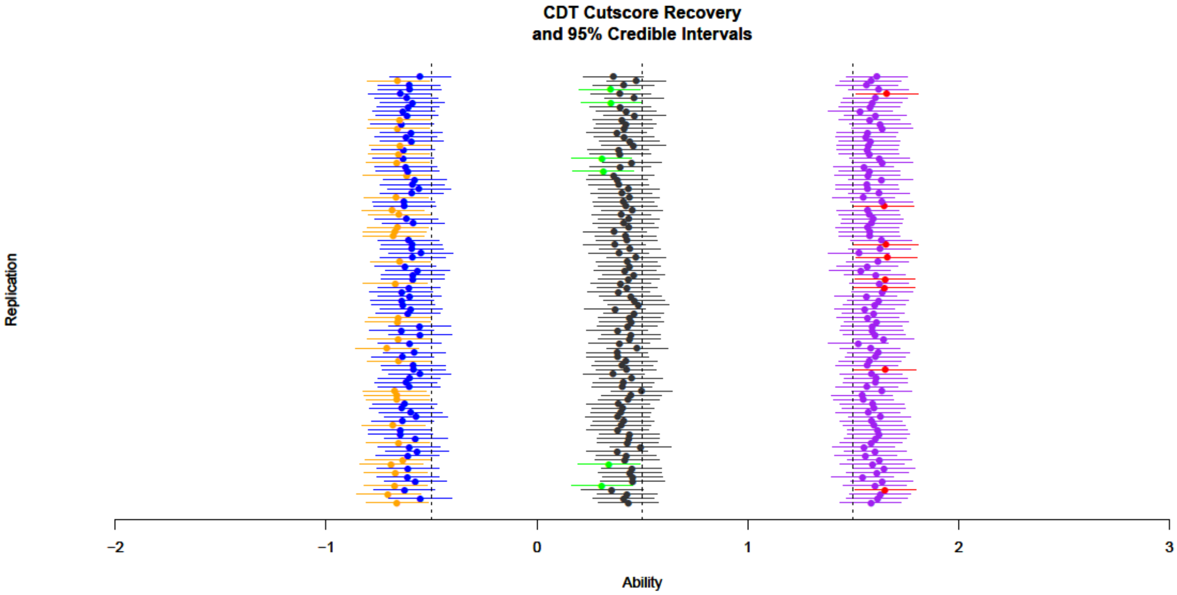

Figure A.110. N_I = 16, N_G = 2, σ^2_B = 0.0, σ^2_C = 0.0; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

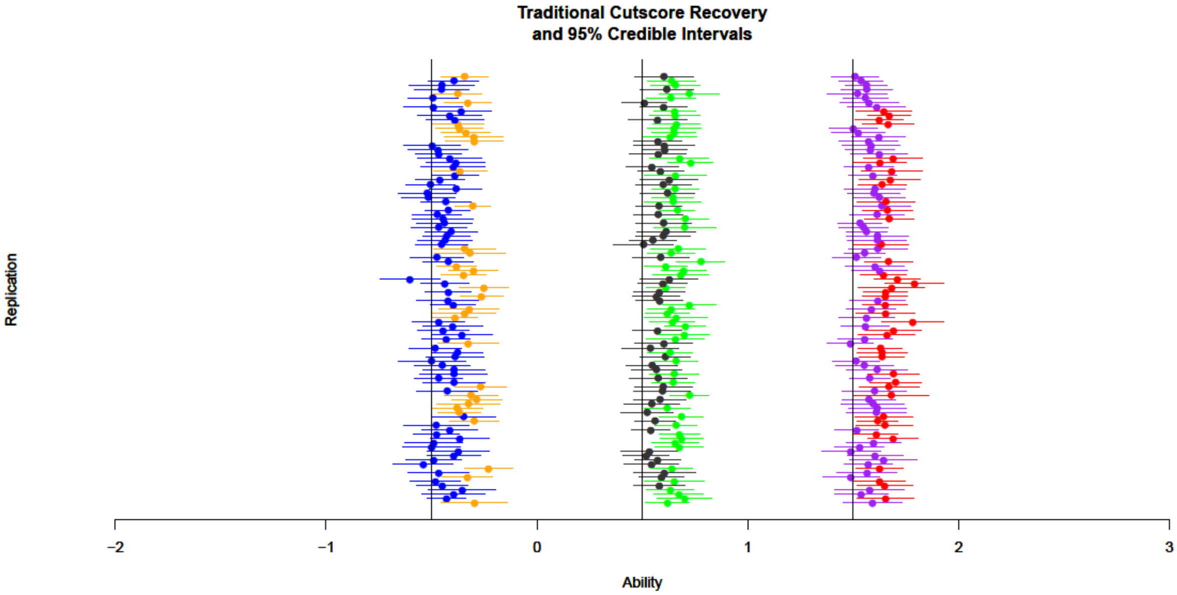

Figure A.111. N_I = 16, N_G = 2, σ^2_B = 0.0, σ^2_C = 0.1; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

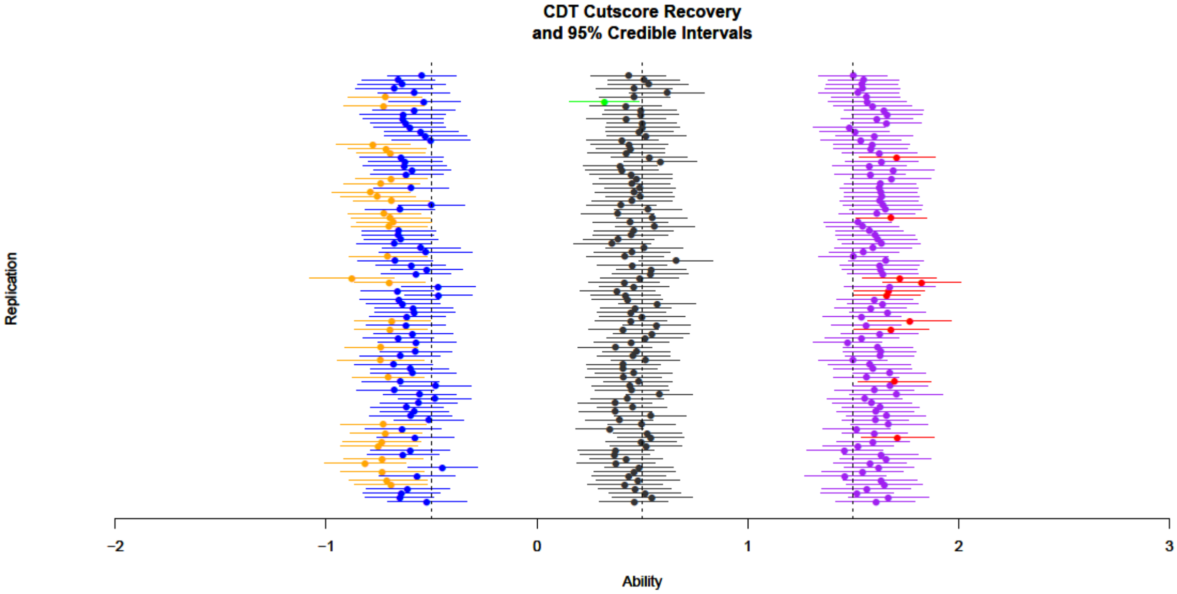

Figure A.112. N_I = 16, N_G = 2, σ^2_B = 0.0, σ^2_C = 0.1; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

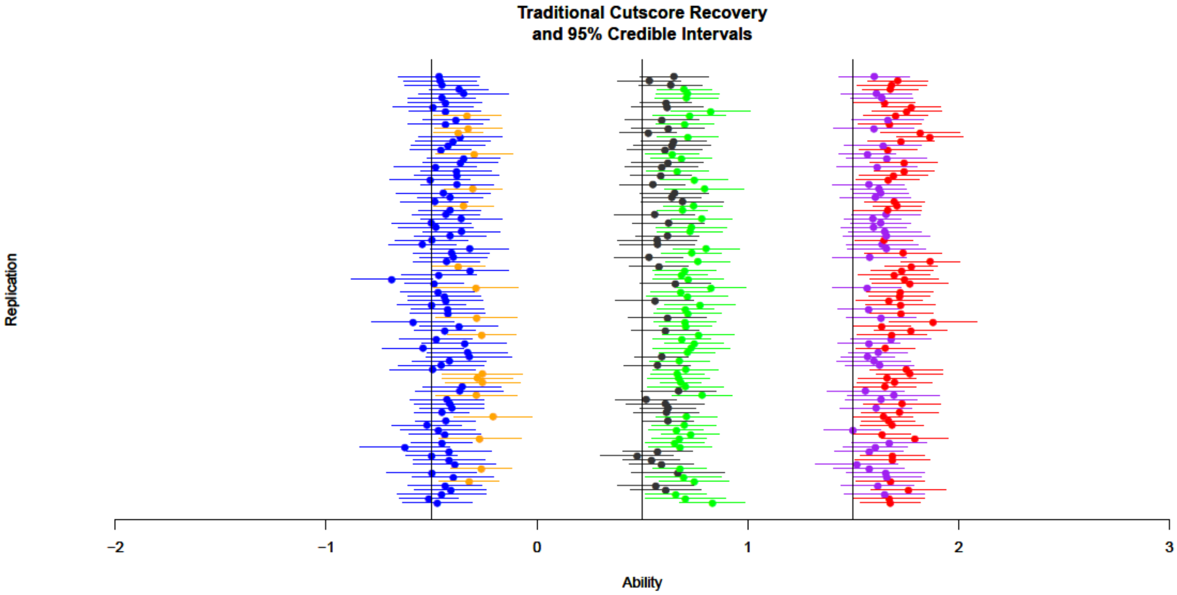

Figure A.113. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

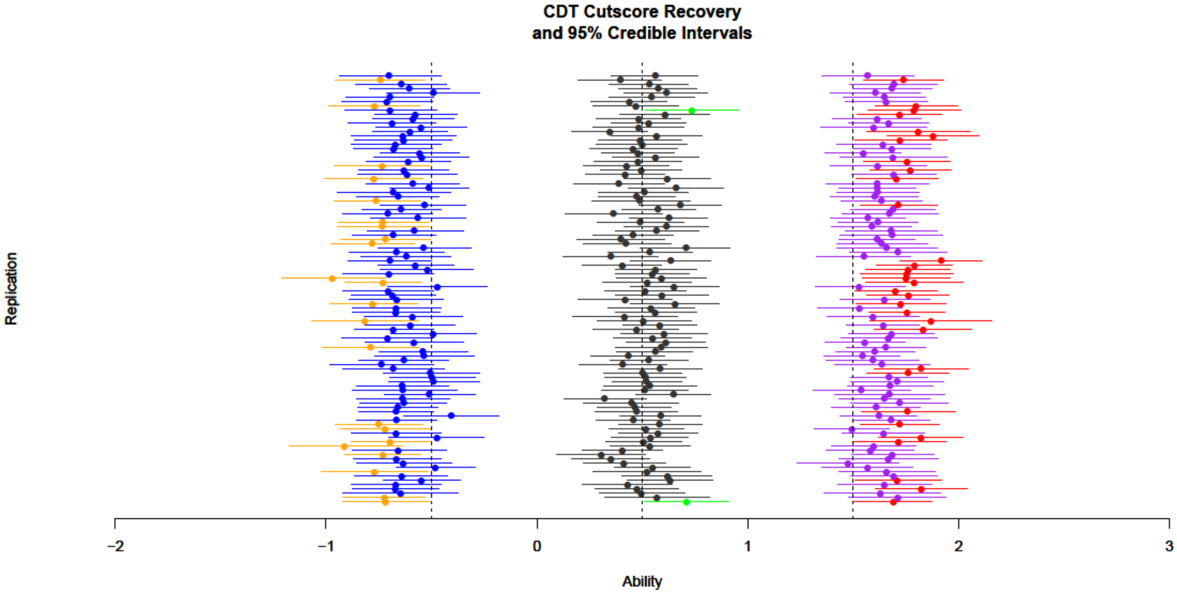

Figure A.114. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

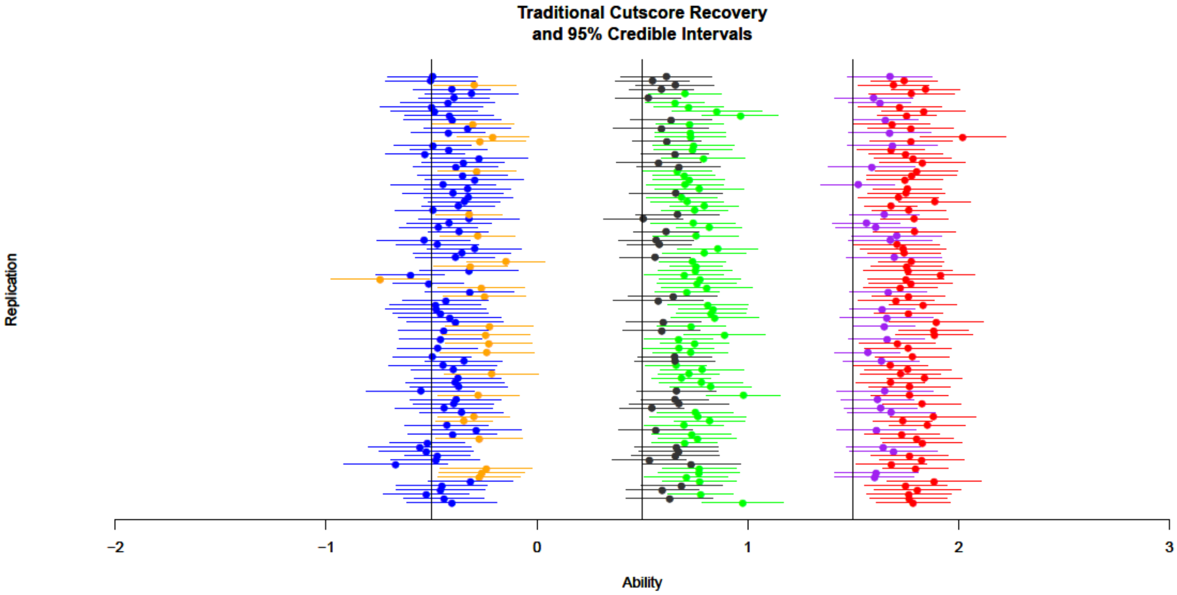

Figure A.115. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

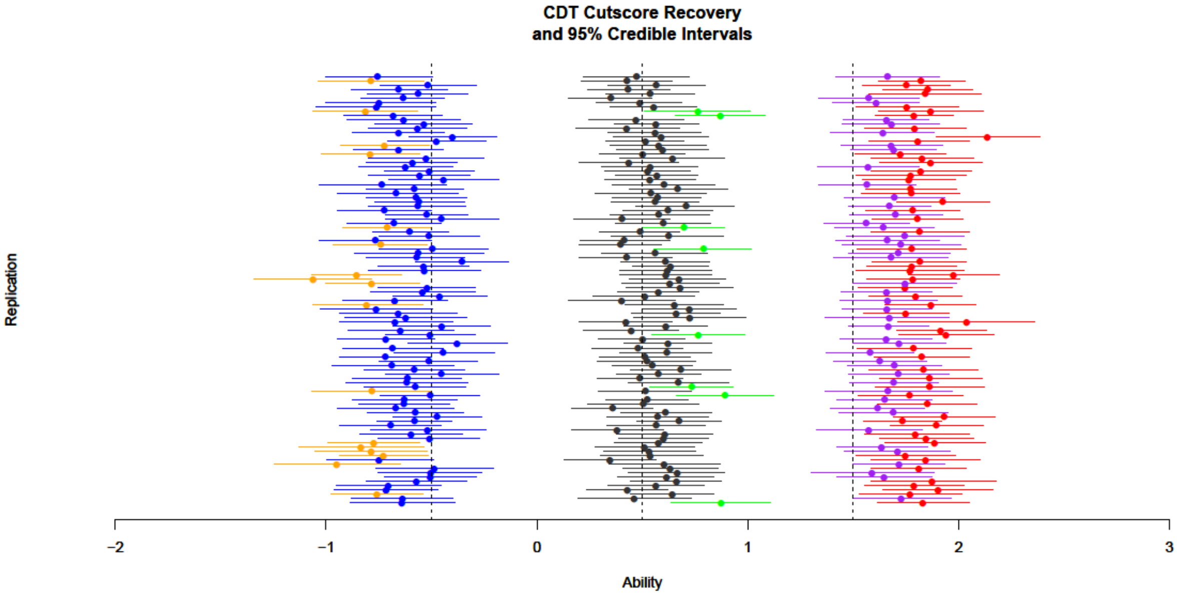

Figure A.116. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

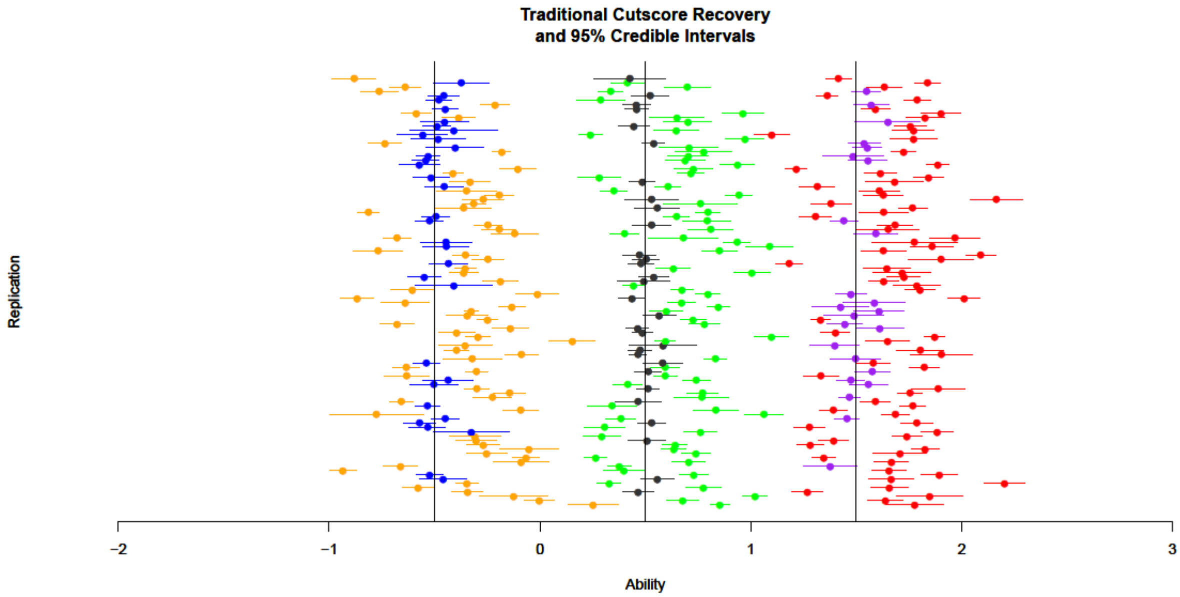

Figure A.117. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

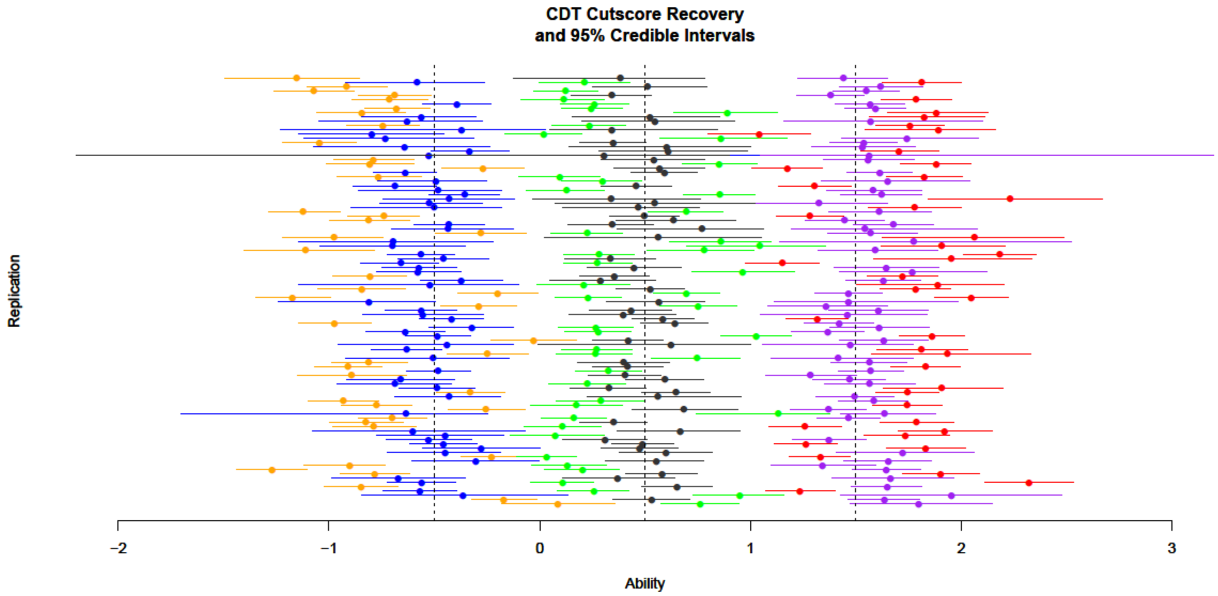

Figure A.118. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

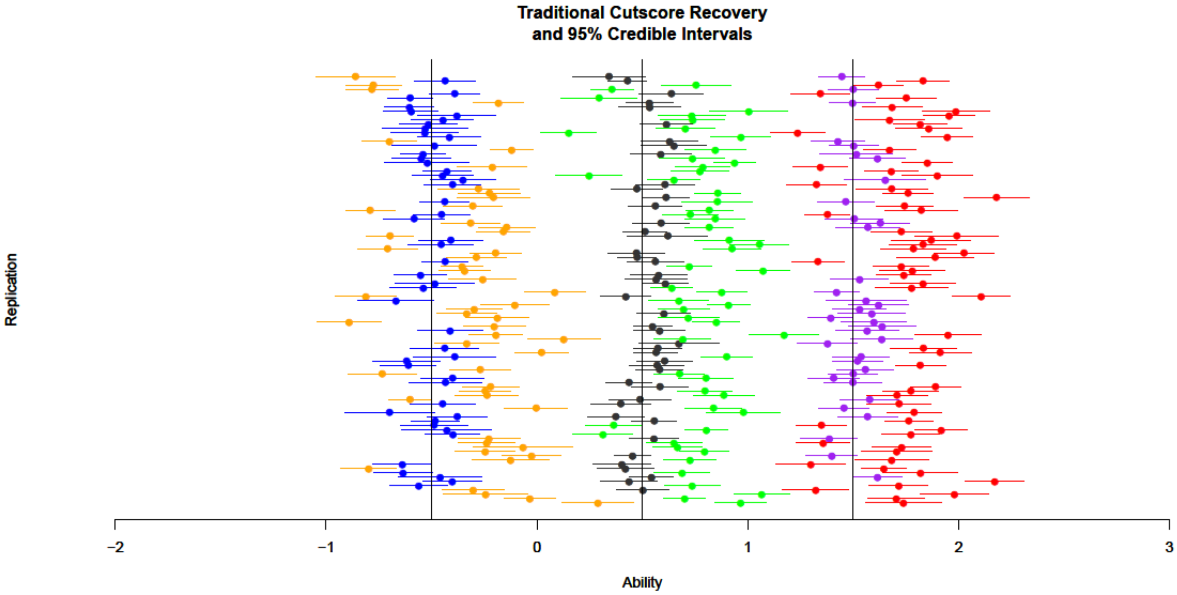

Figure A.119. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

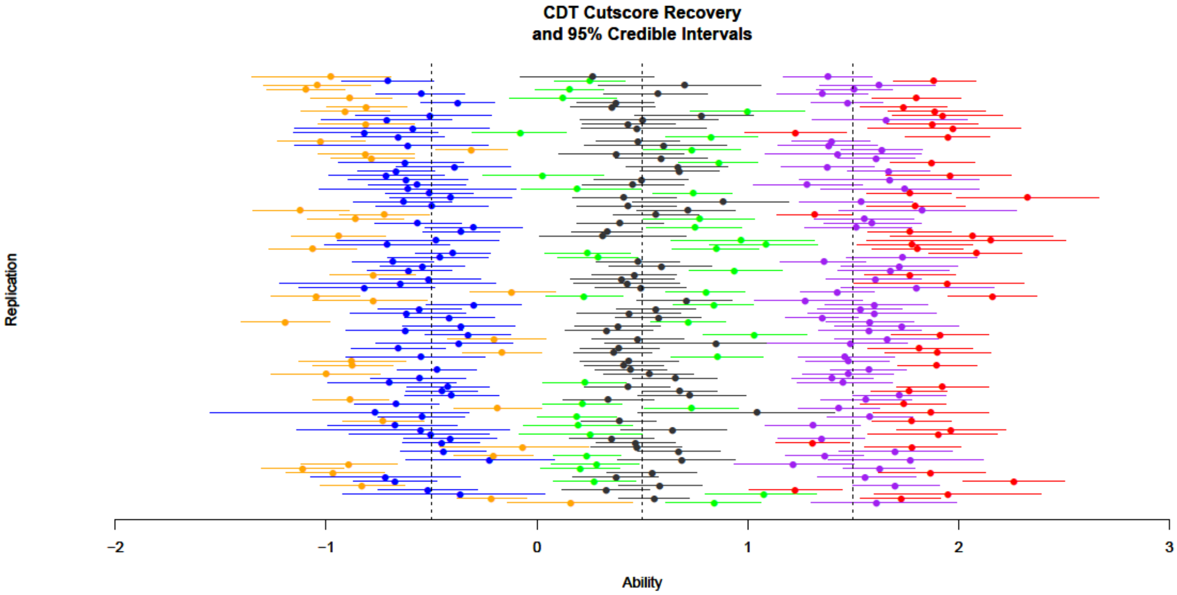

Figure A.120. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

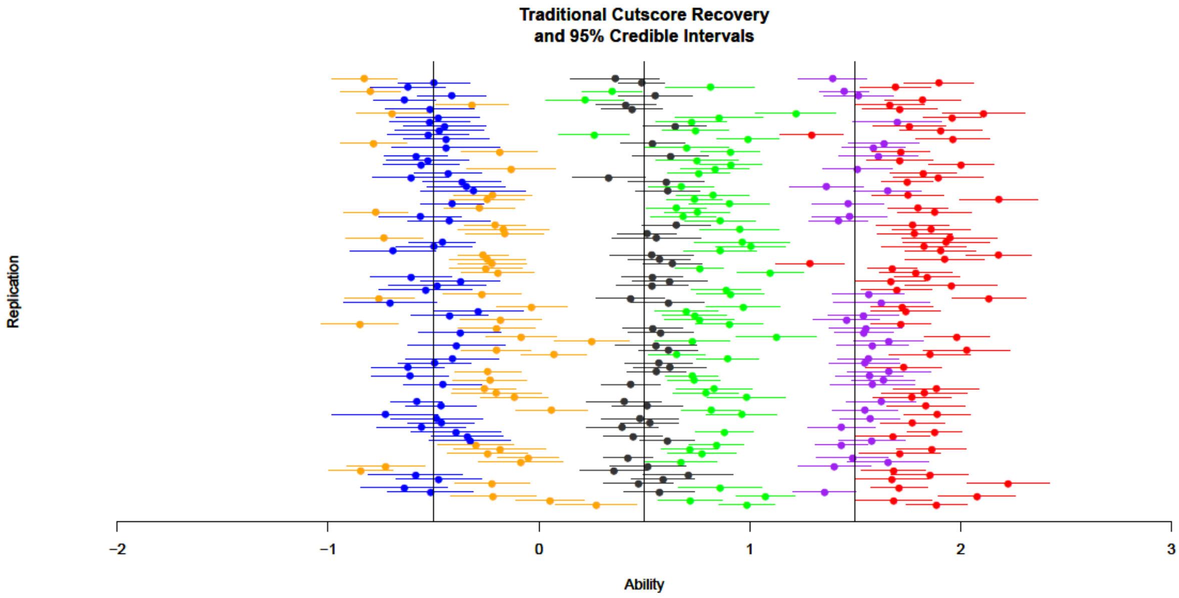

Figure A.121. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

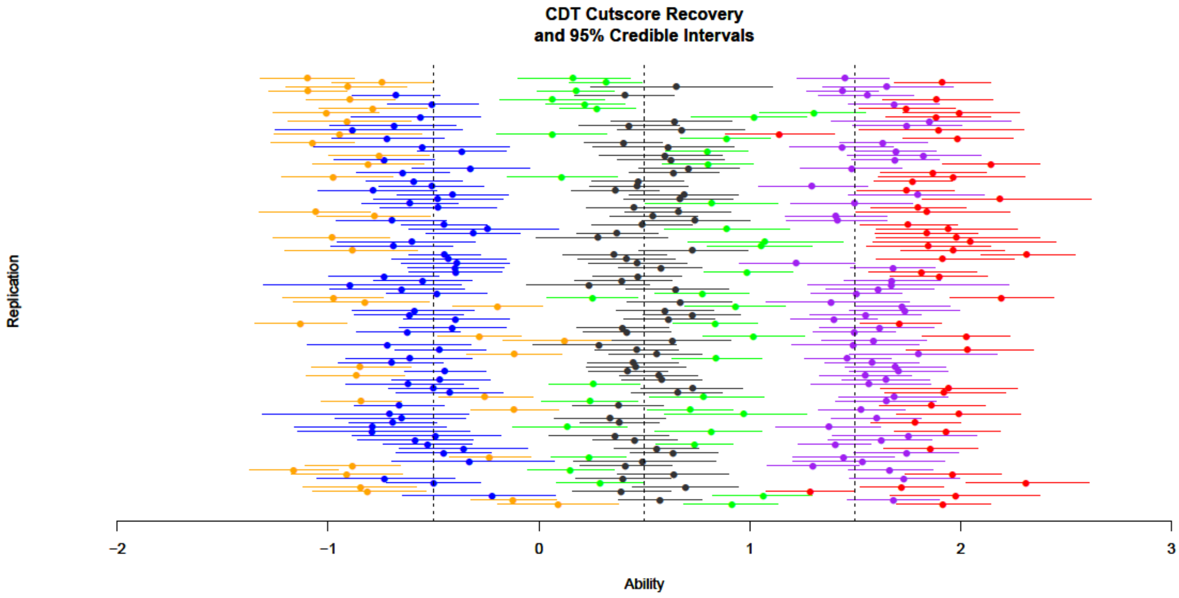

Figure A.122. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

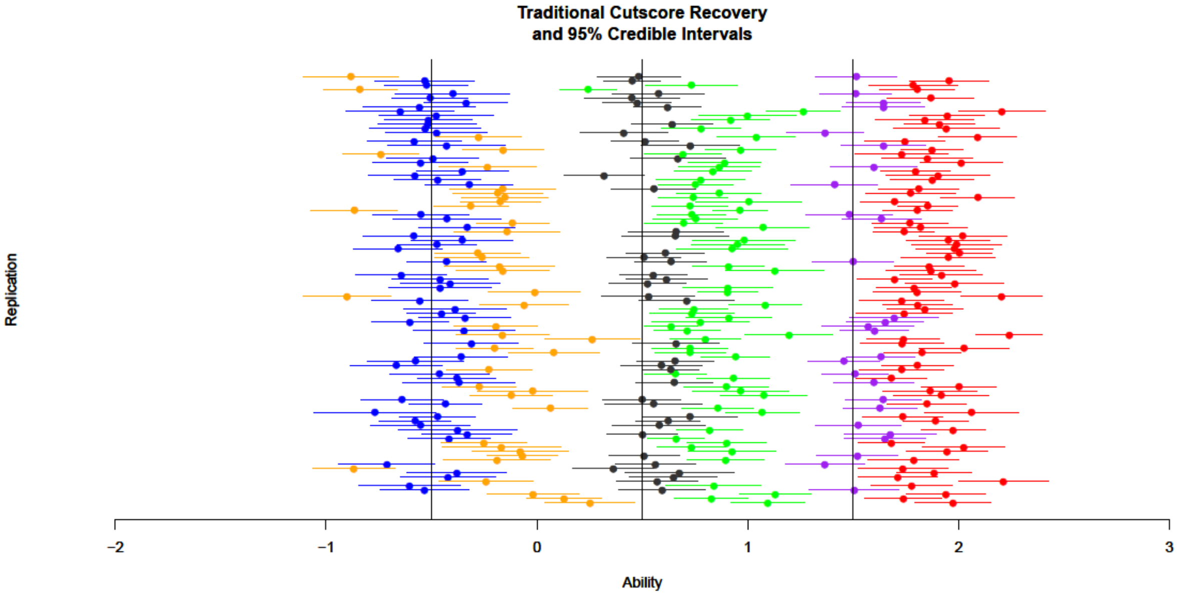

Figure A.123. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

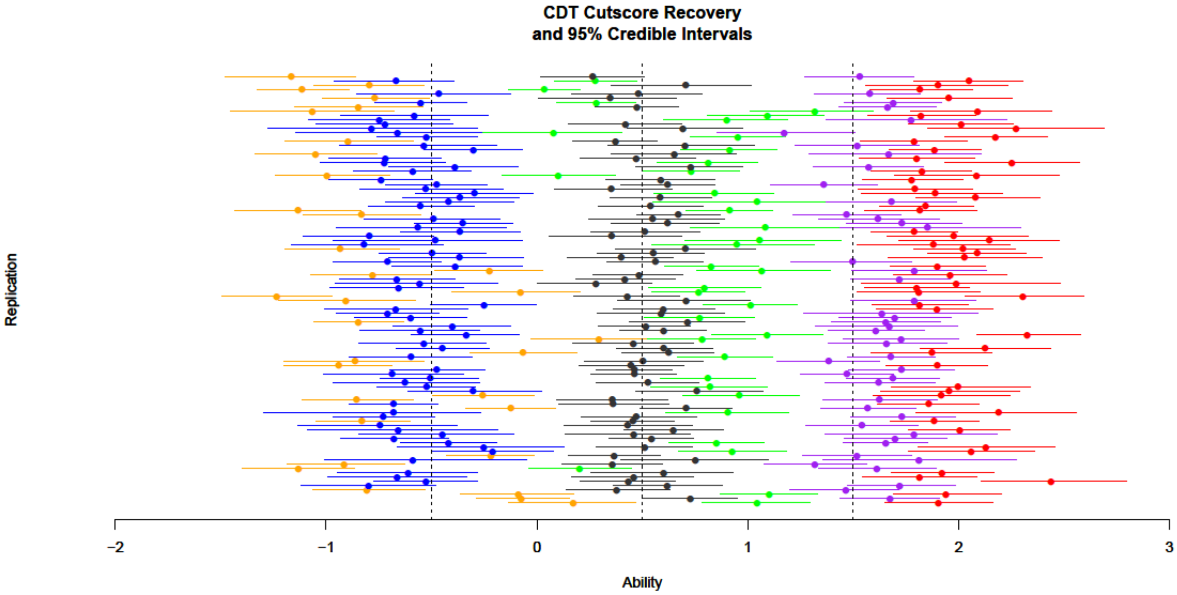

Figure A.124. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

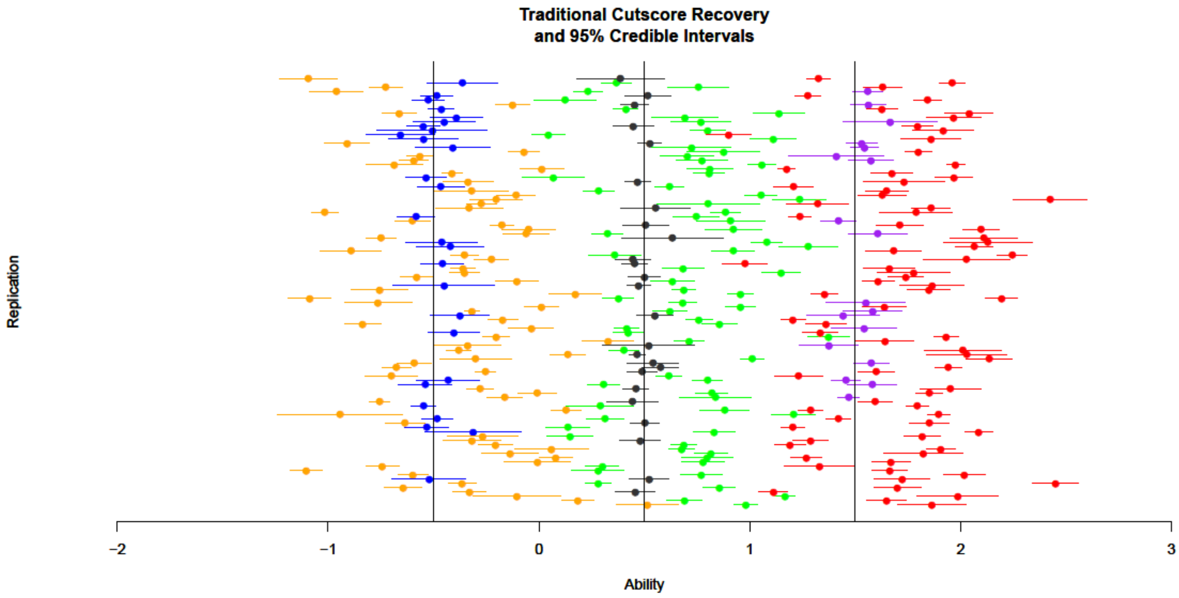

Figure A.125. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

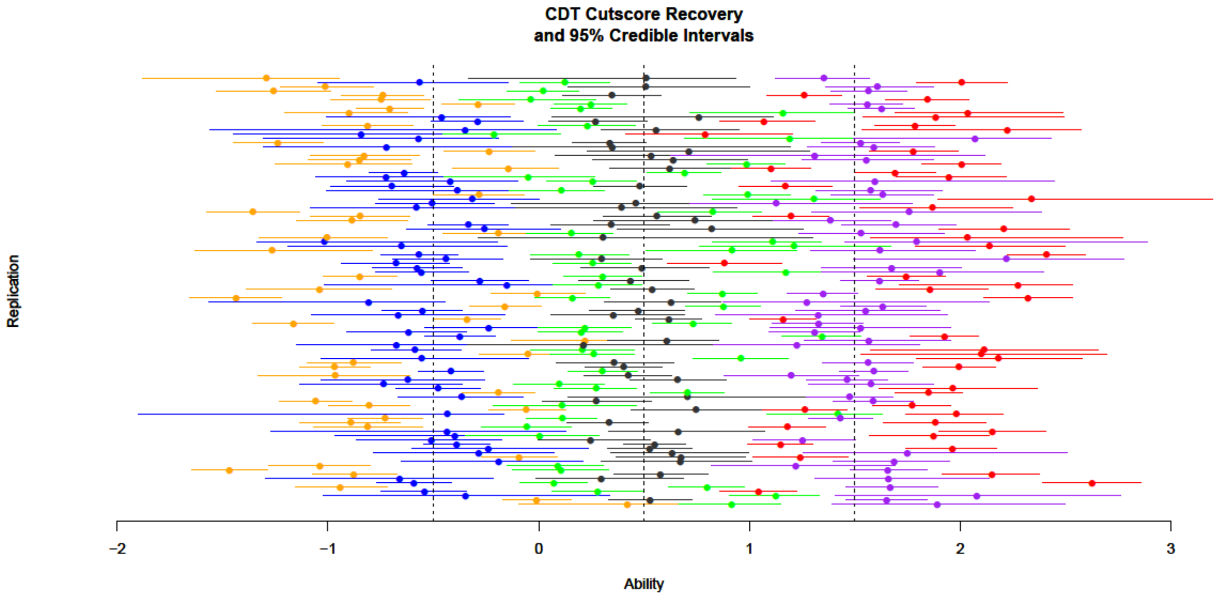

Figure A.126. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

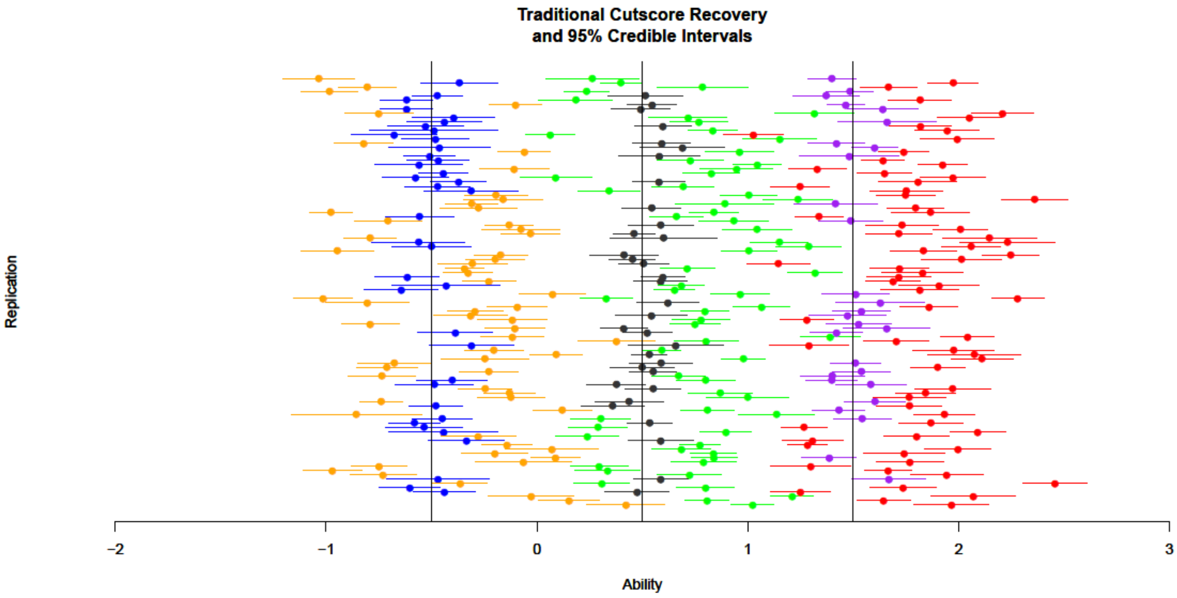

Figure A.127. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

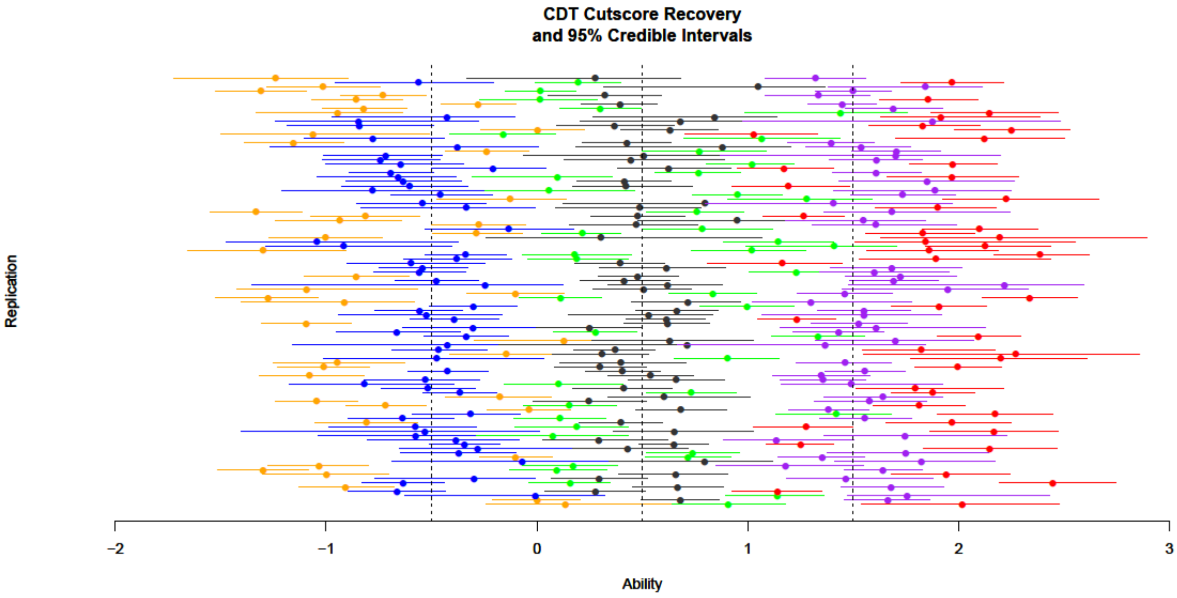

Figure A.128. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

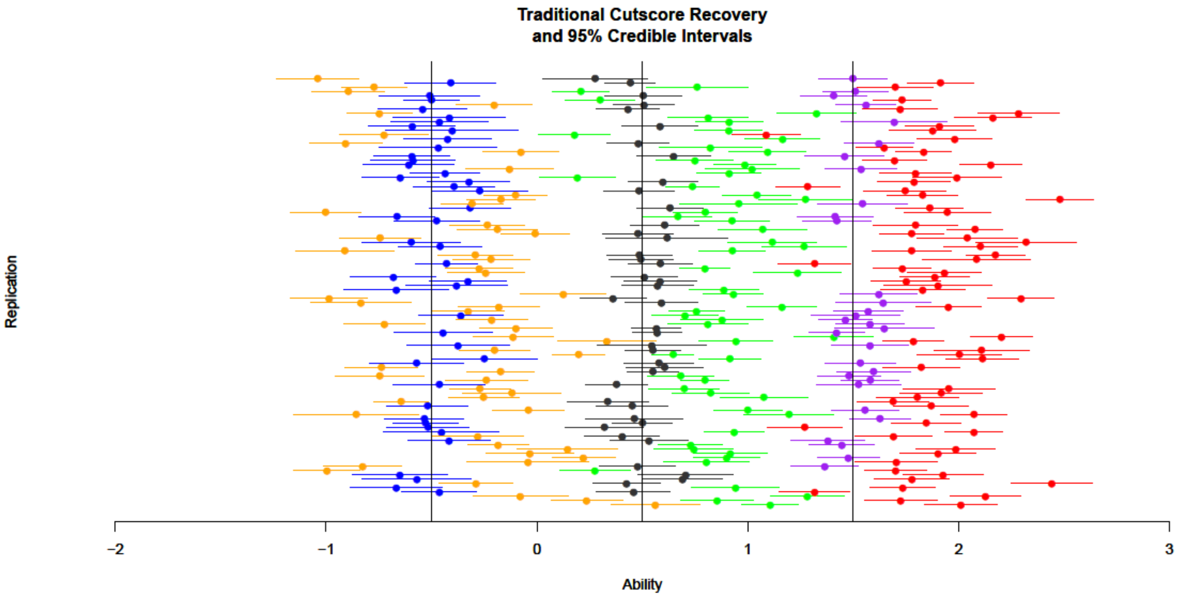

Figure A.129. $N_I = 16$, $N_G = 2$, $\sigma^2{}_B = 0.2$, $\sigma^2{}_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

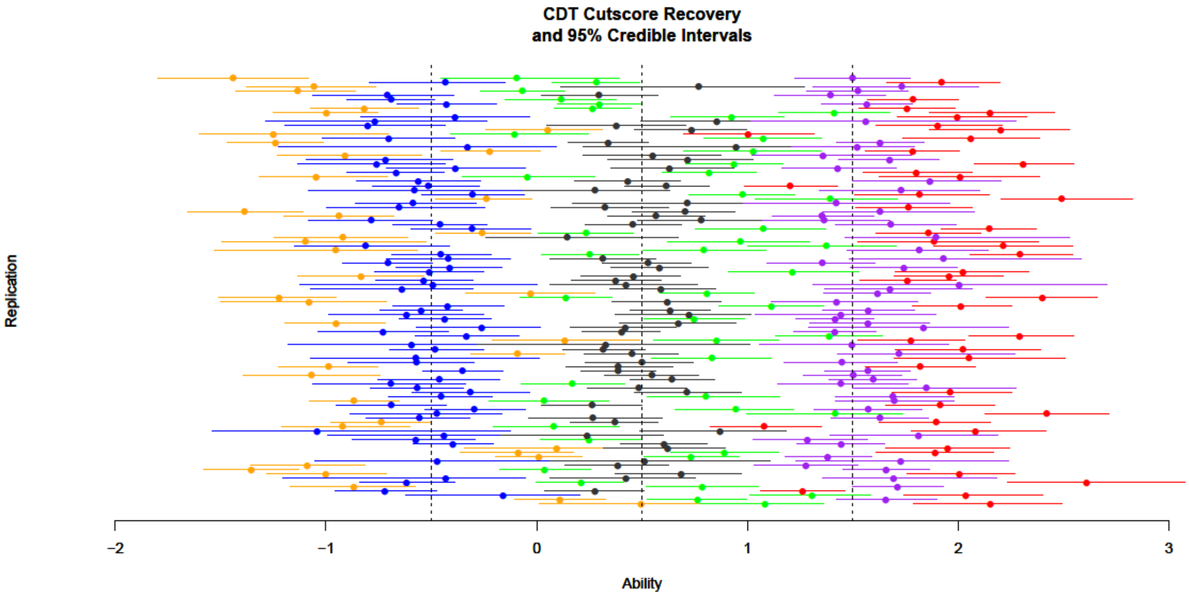

Figure A.130. $N_I = 16$, $N_G = 2$ $\sigma^2{}_B = 0.2$, $\sigma^2{}_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

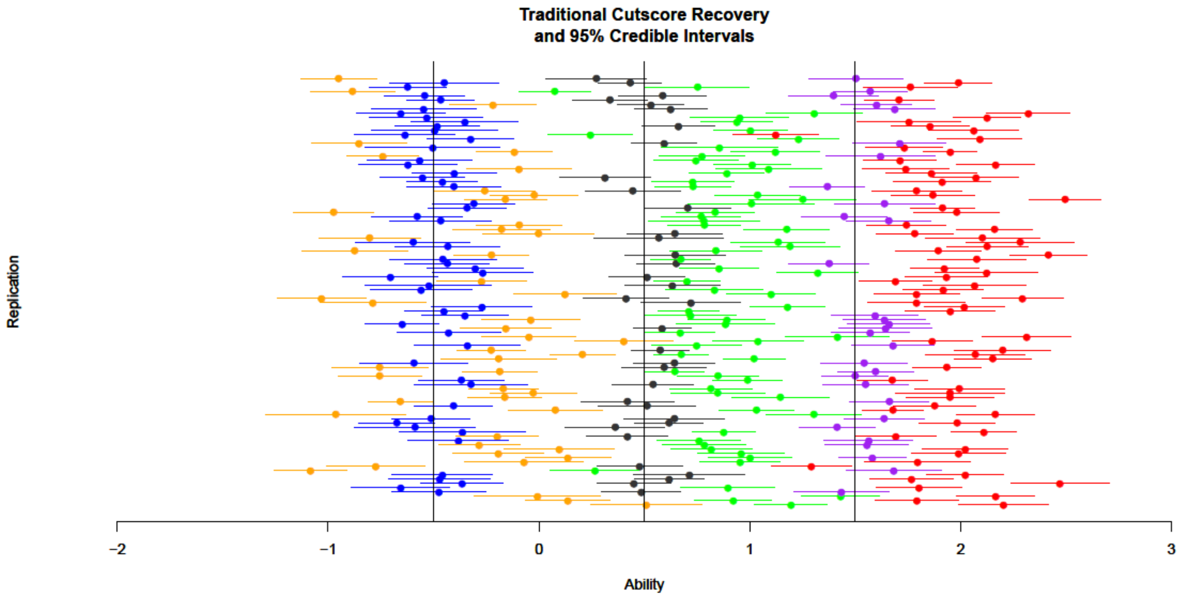

Figure A.131. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

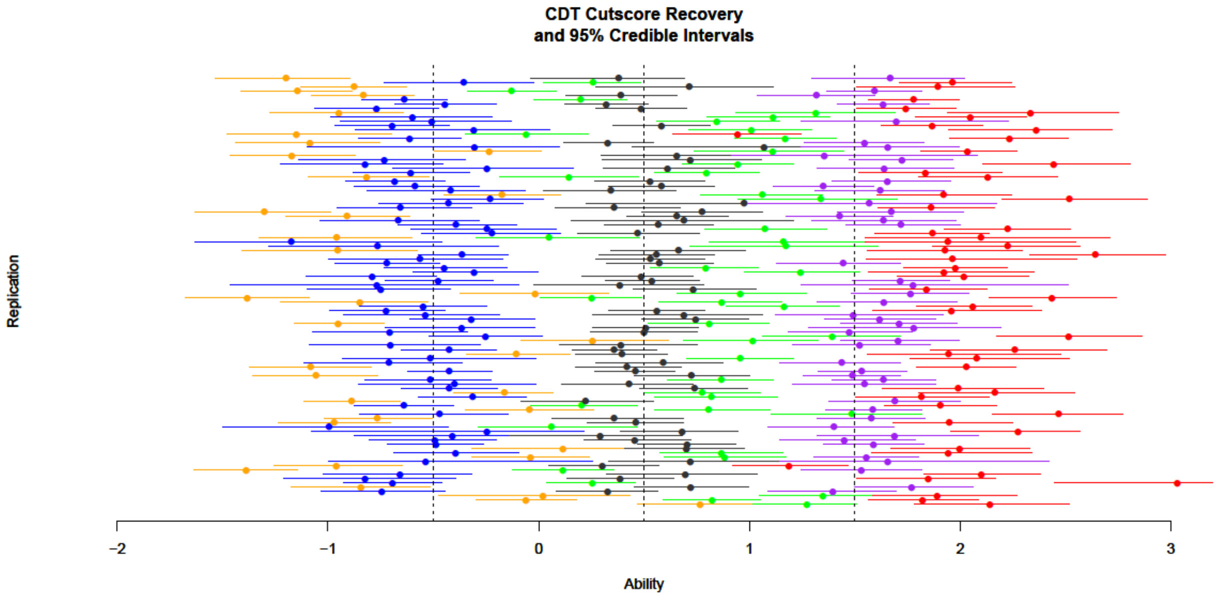

Figure A.132. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

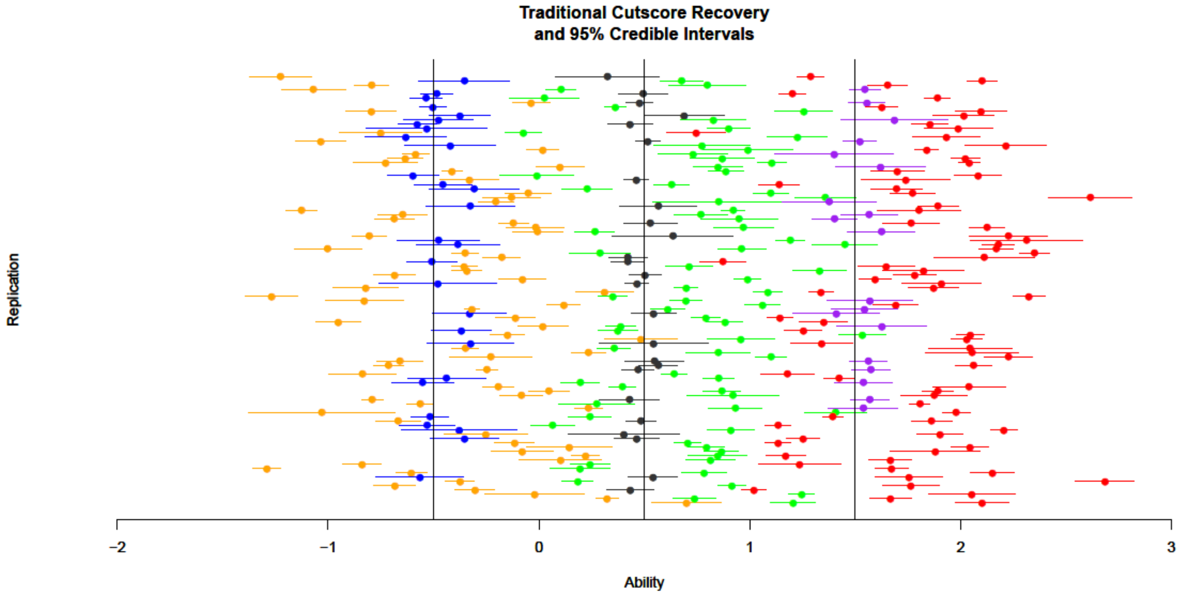

Figure A.133. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

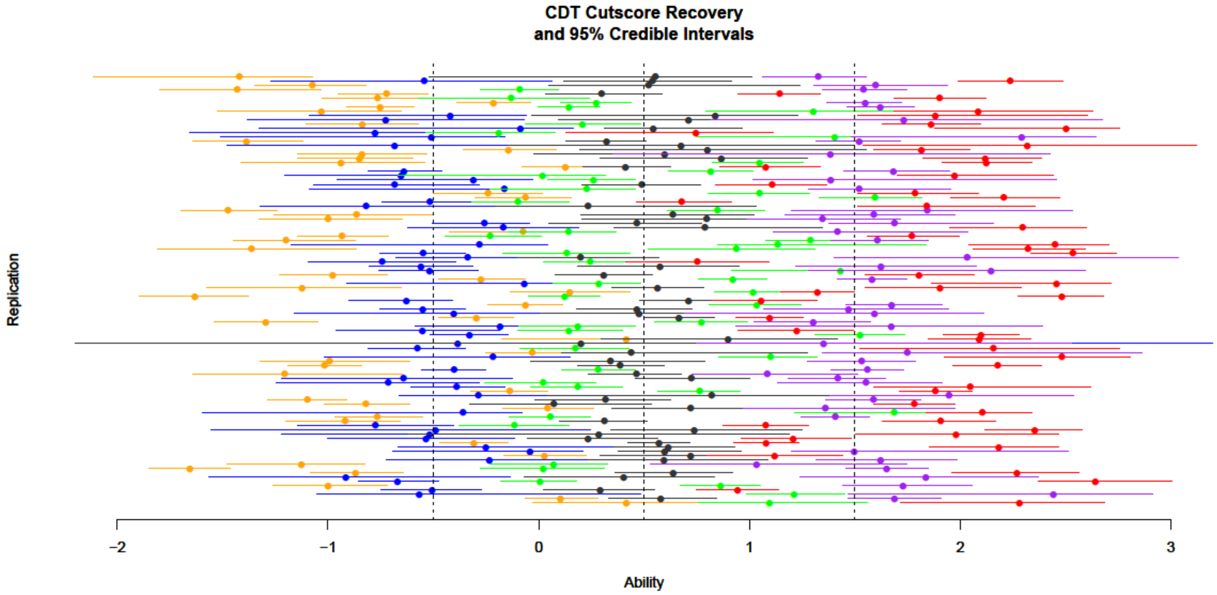

Figure A.134. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

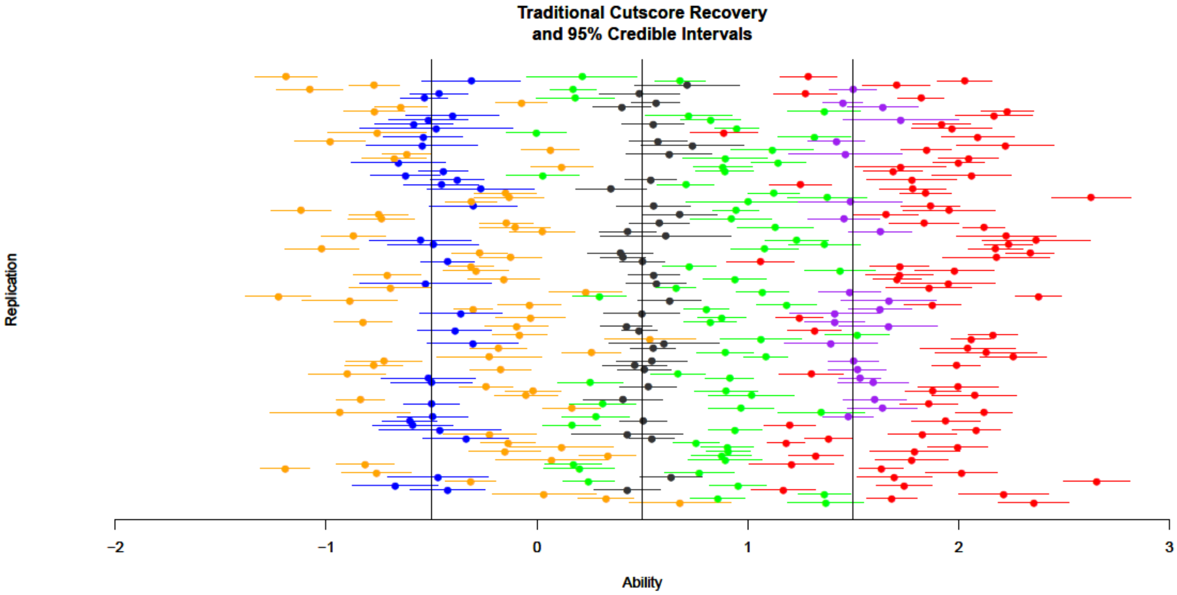

Figure A.135. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

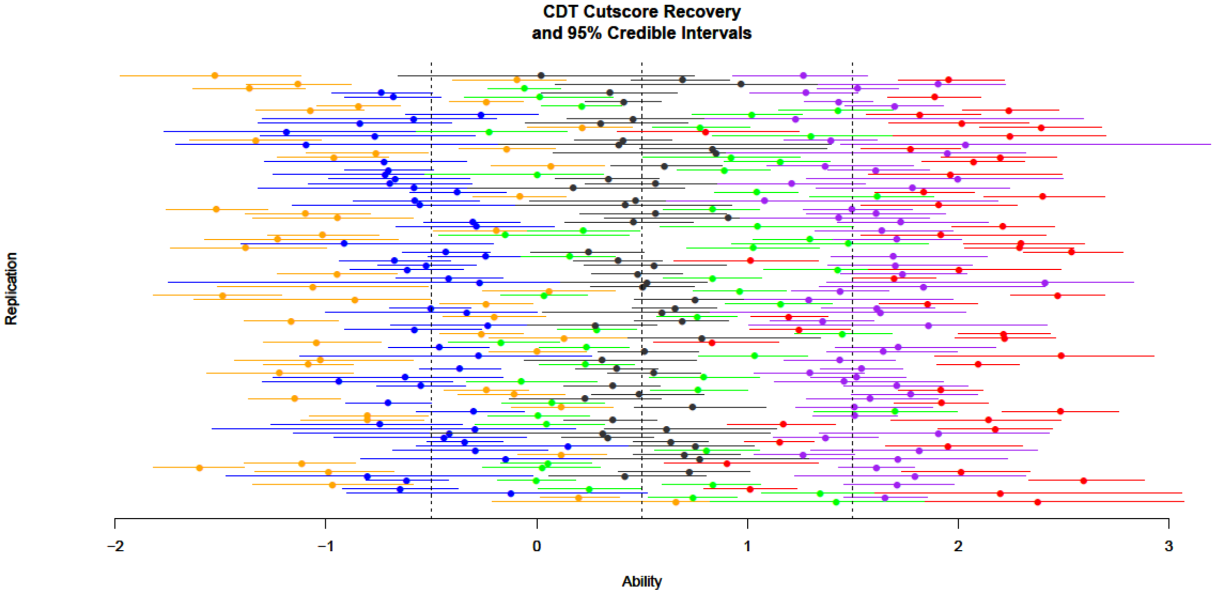

Figure A.136. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

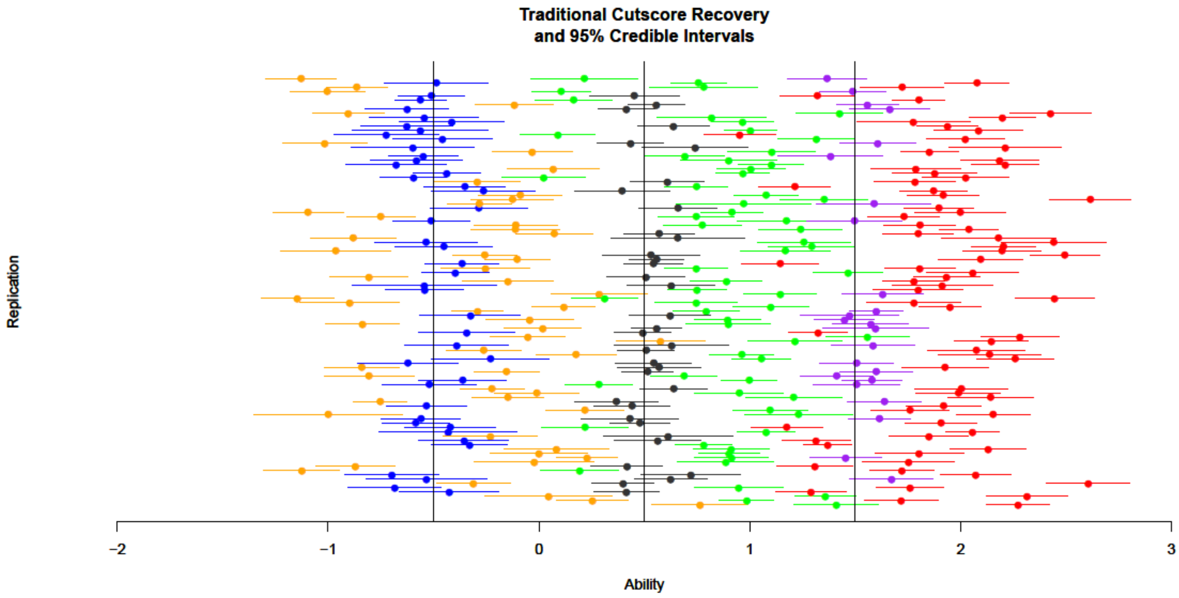

Figure A.137. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

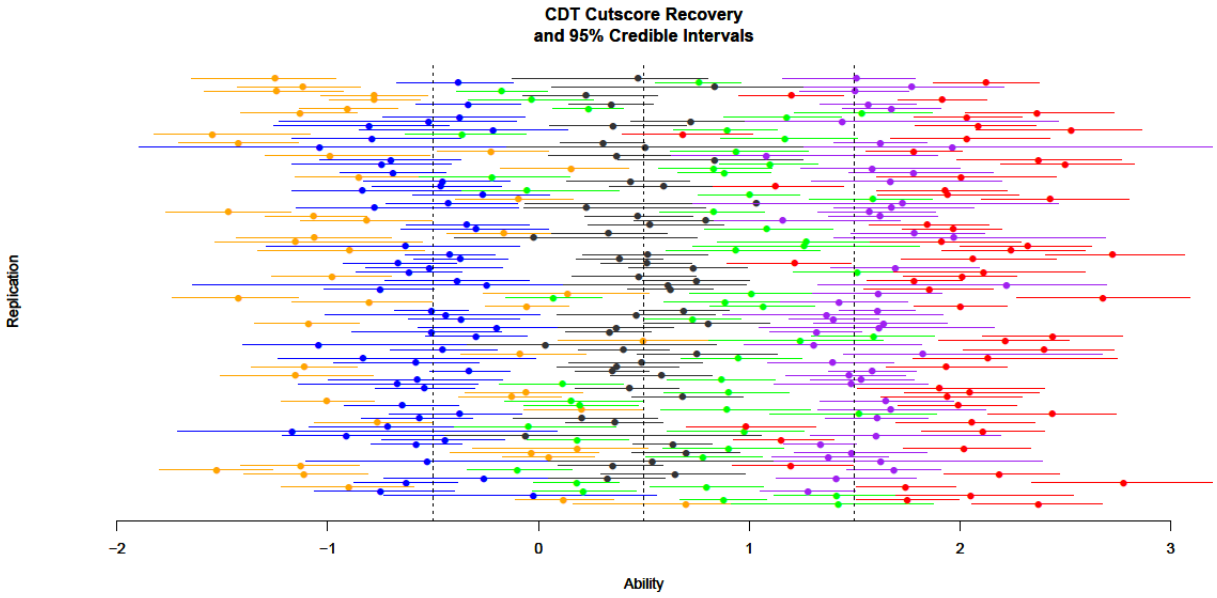

Figure A.138. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

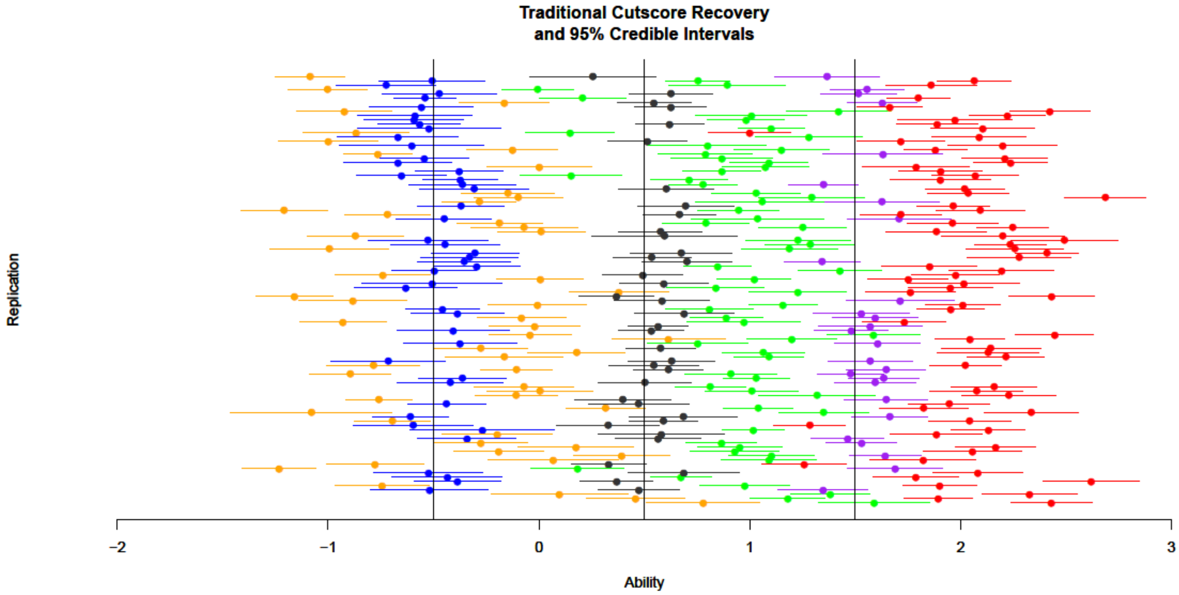

Figure A.139. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

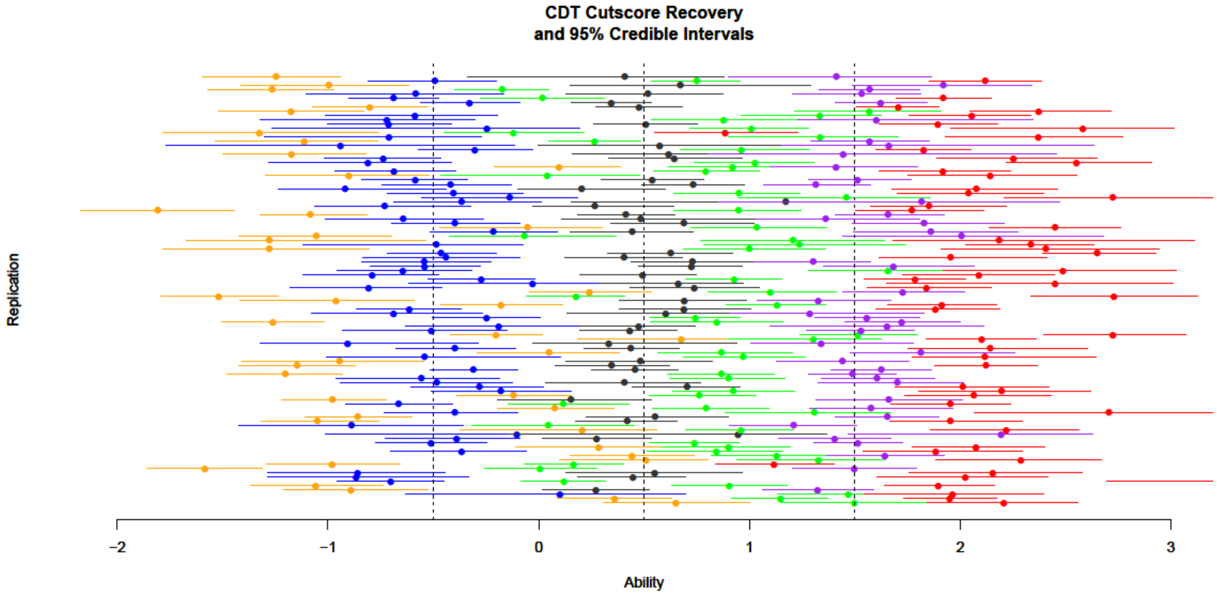

Figure A.140. $N_I = 16$, $N_G = 2$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

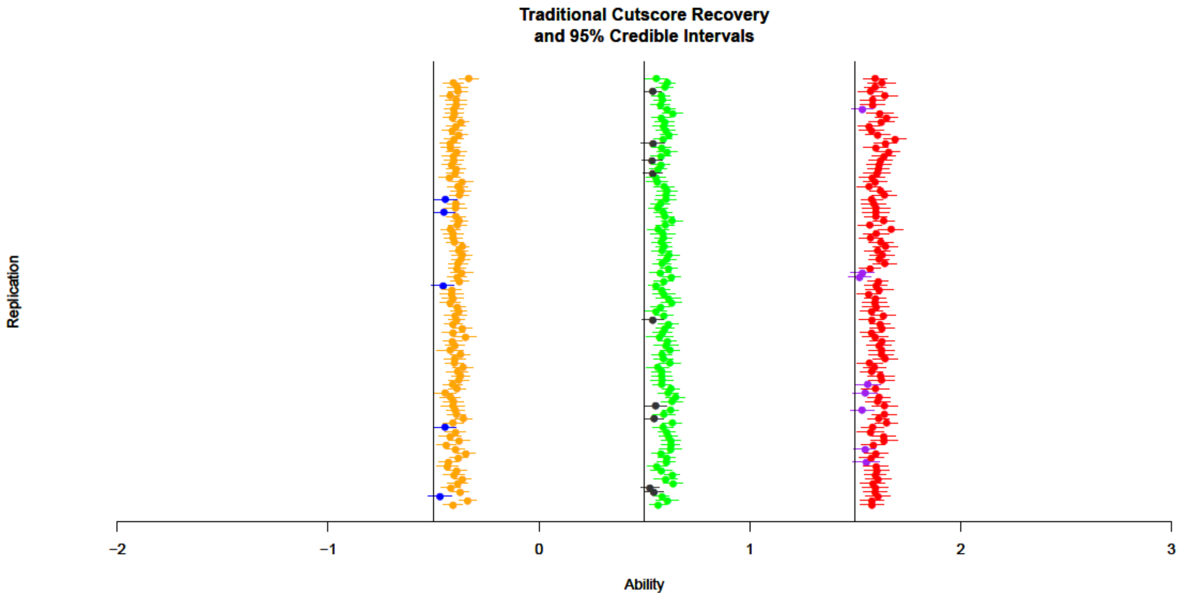

Figure A.141. N_I = 16, N_G = 3, σ^2_B = 0.0, σ^2_C = 0.0; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

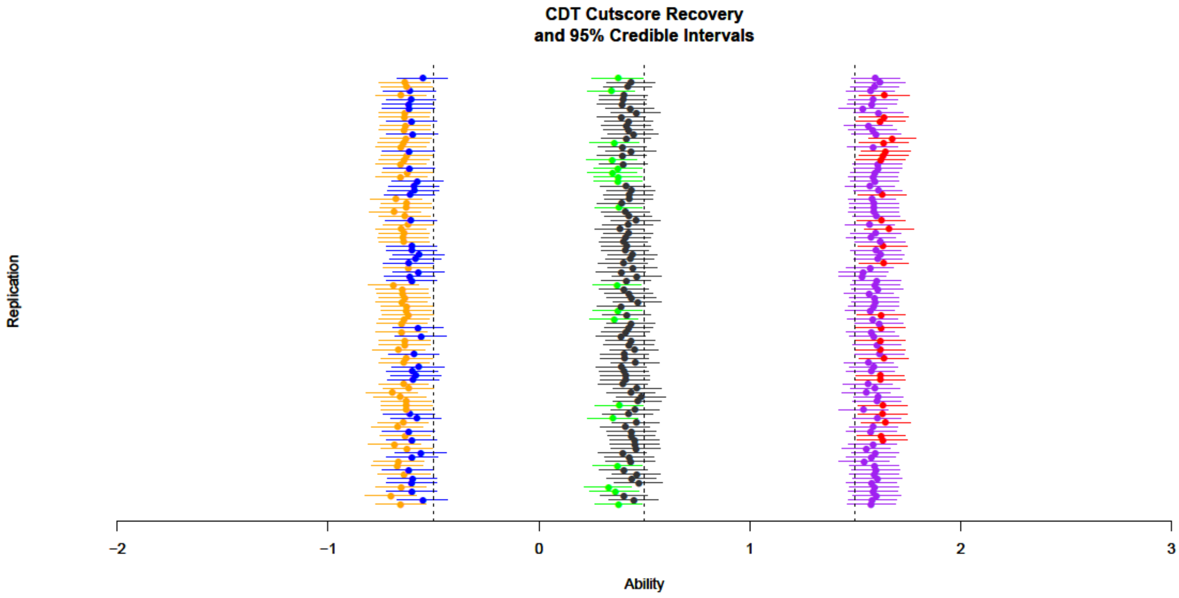

Figure A.142. N_I = 16, N_G = 3, σ^2_B = 0.0, σ^2_C = 0.0; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

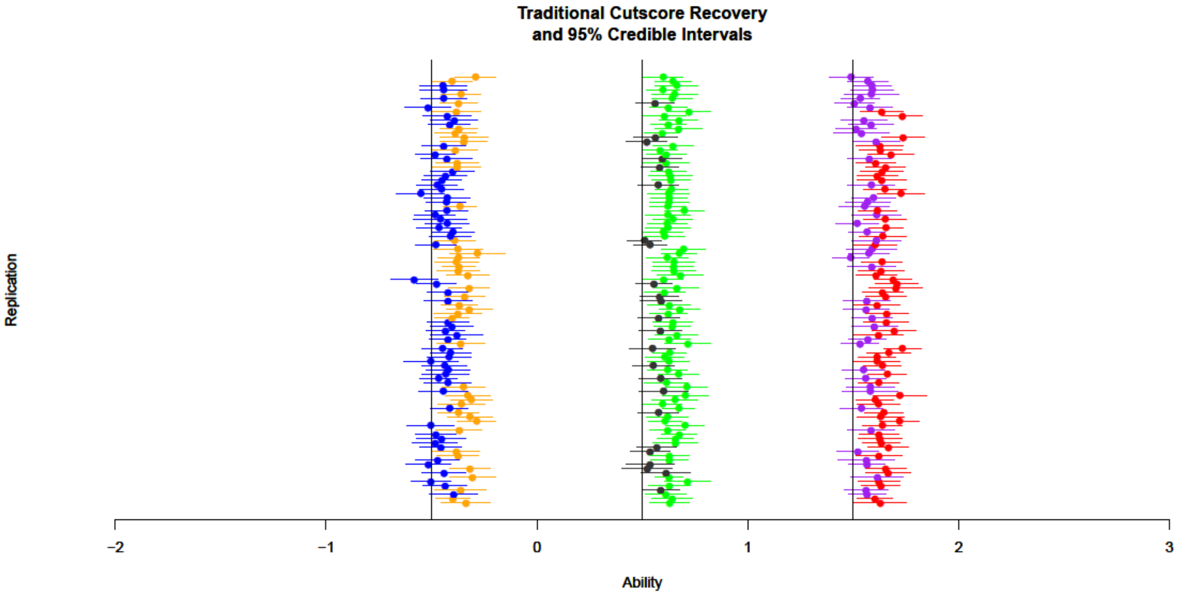

Figure A.143. $N_I = 16$, $N_G = 3$, $\sigma^2{}_B = 0.0$, $\sigma^2{}_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

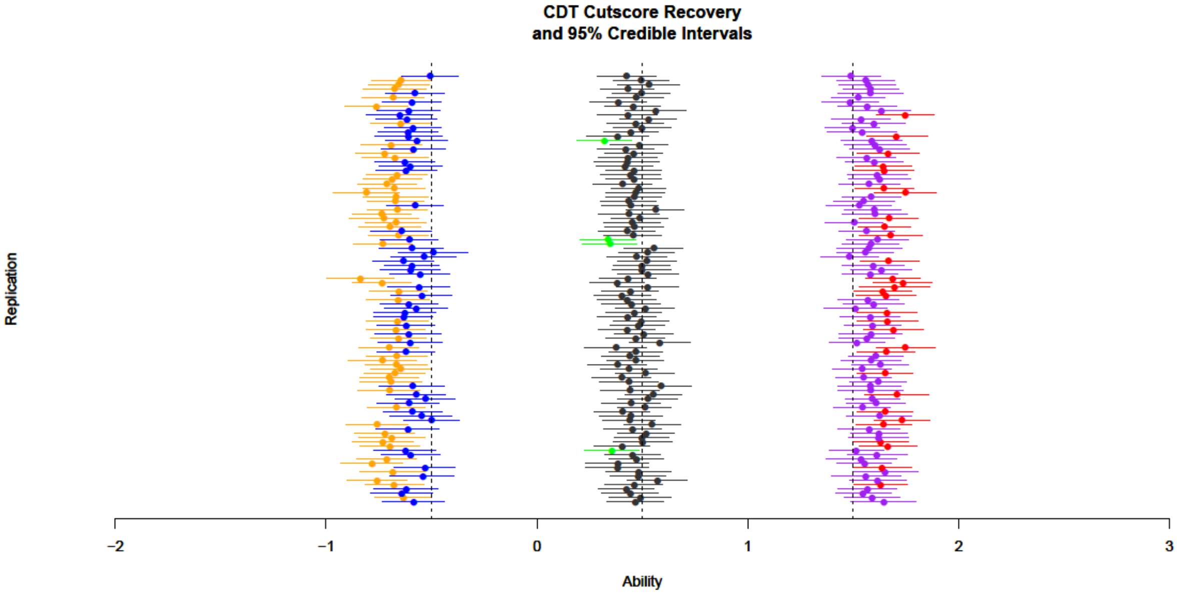

Figure A.144. $N_I = 16$, $N_G = 3$, $\sigma^2{}_B = 0.0$, $\sigma^2{}_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

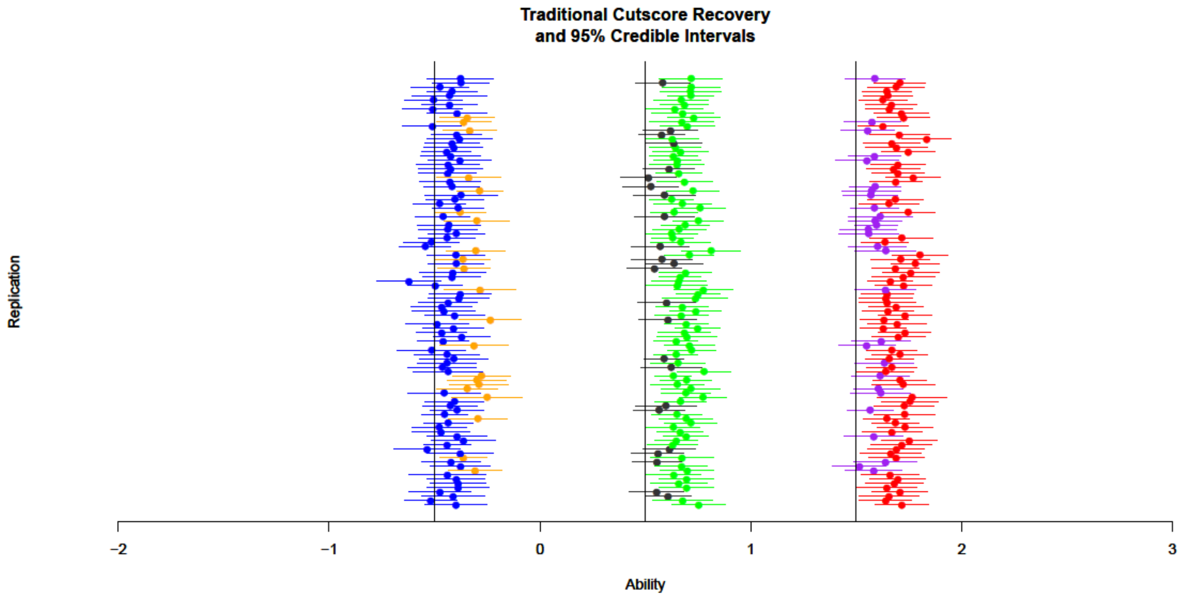

Figure A.145. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

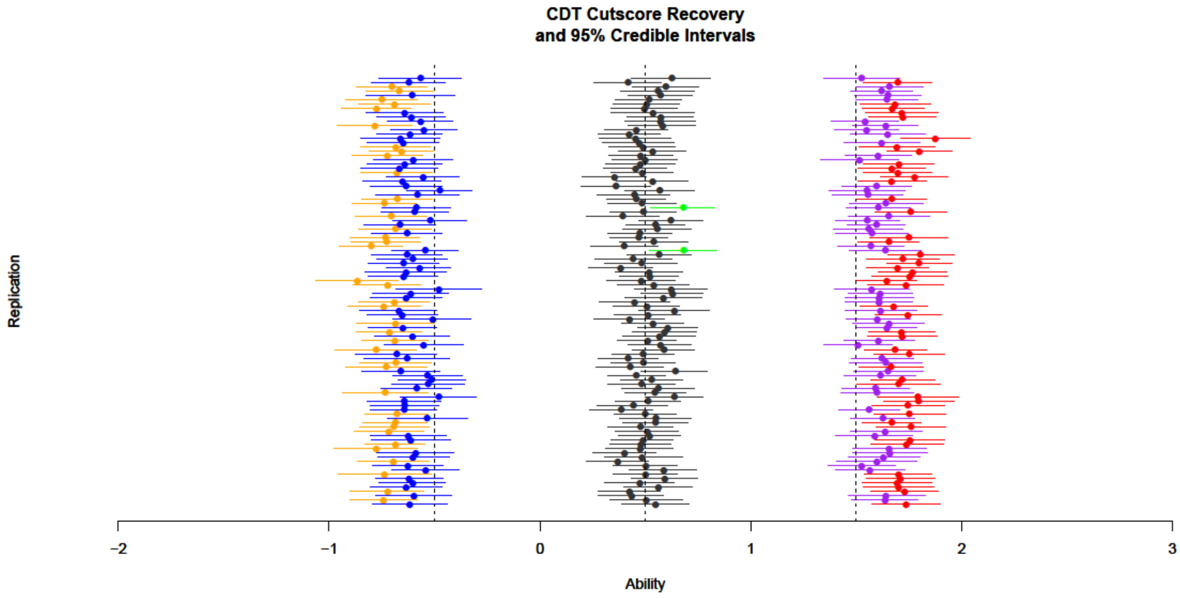

Figure A.146. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

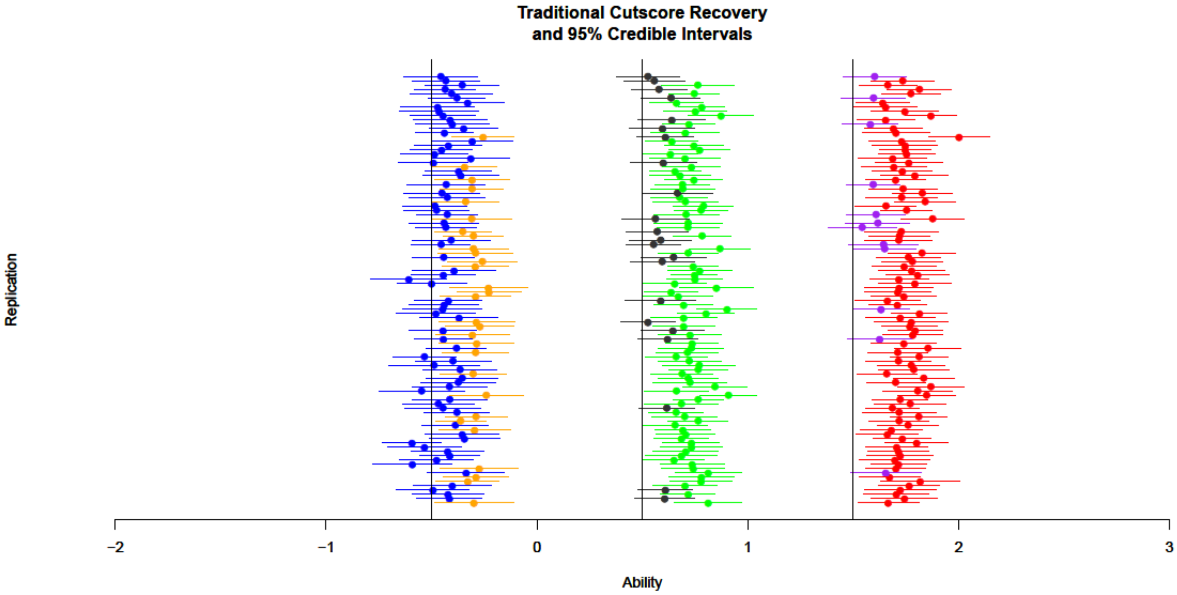

Figure A.147. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

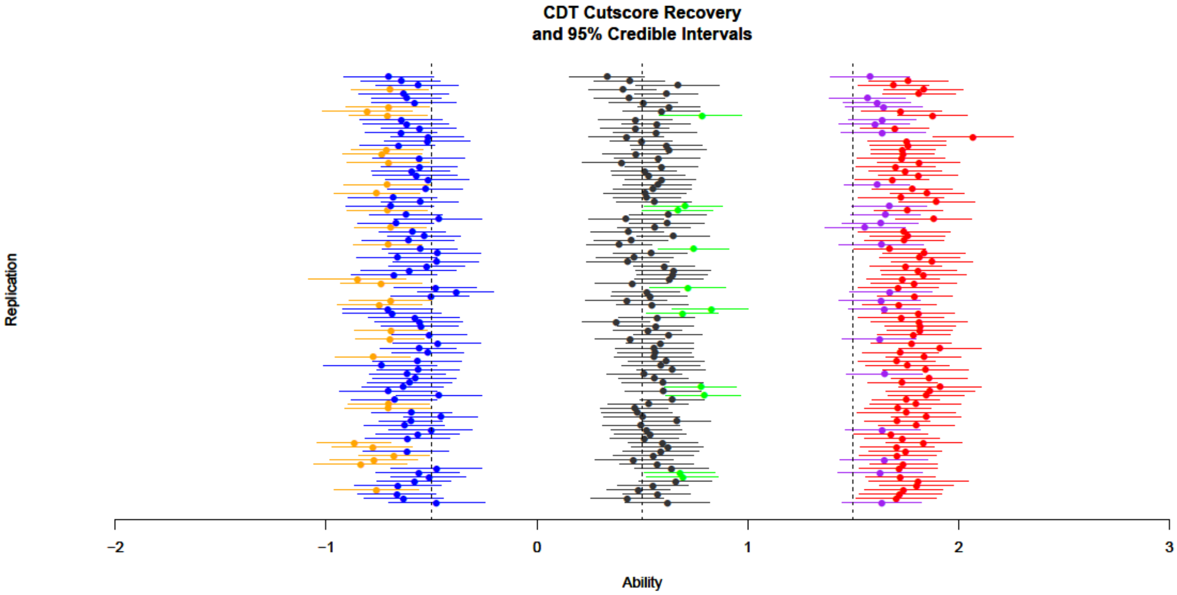

Figure A.148. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

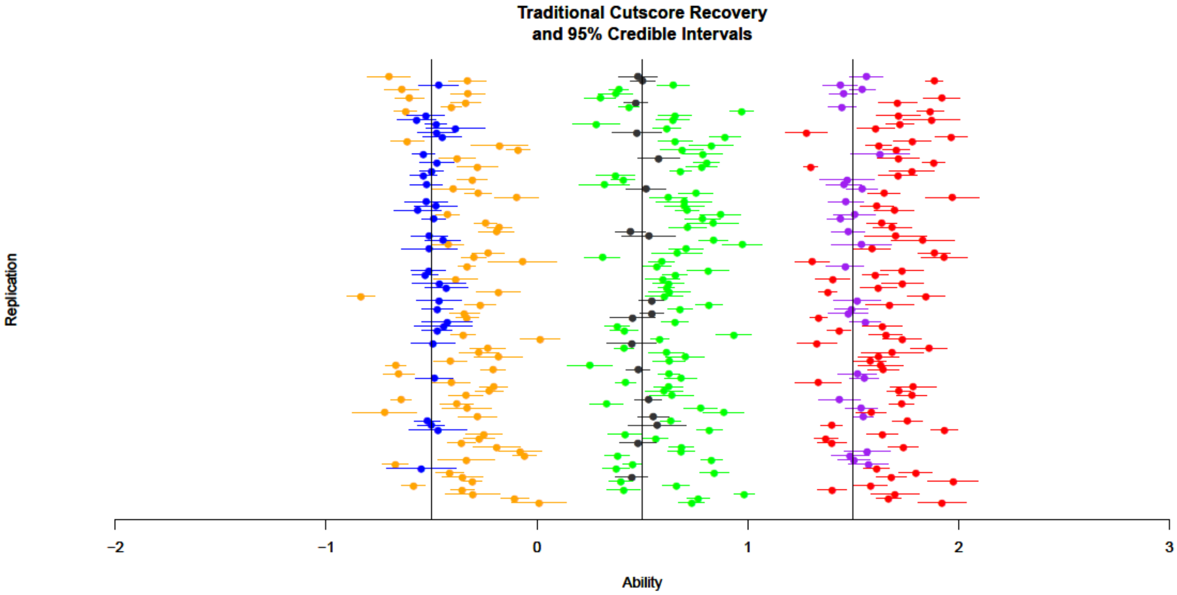

Figure A.149. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

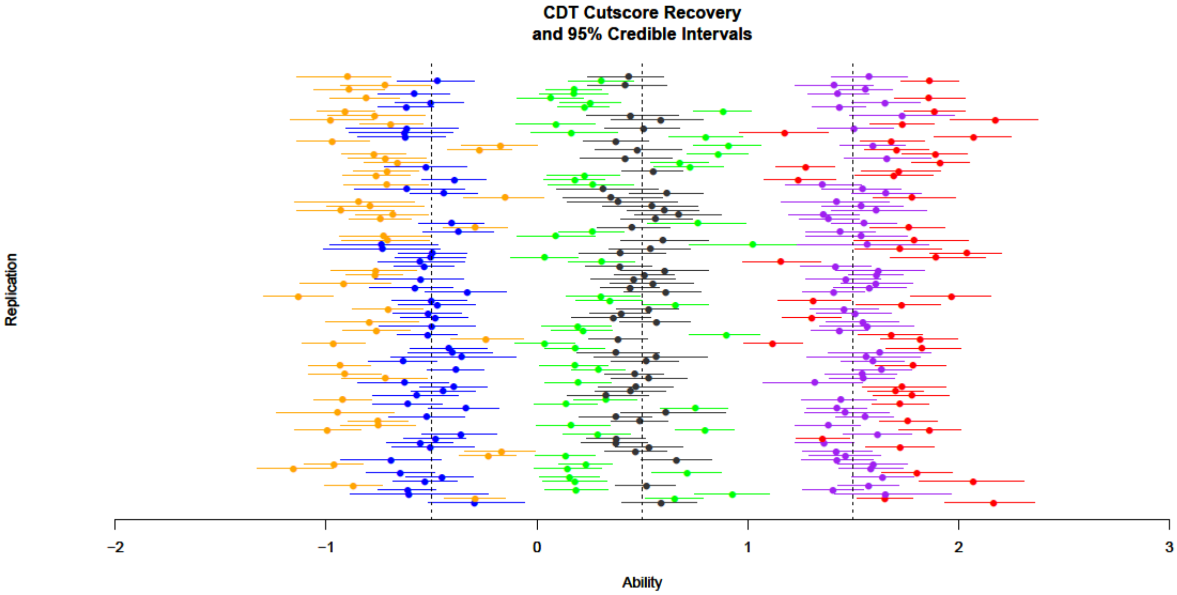

Figure A.150. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

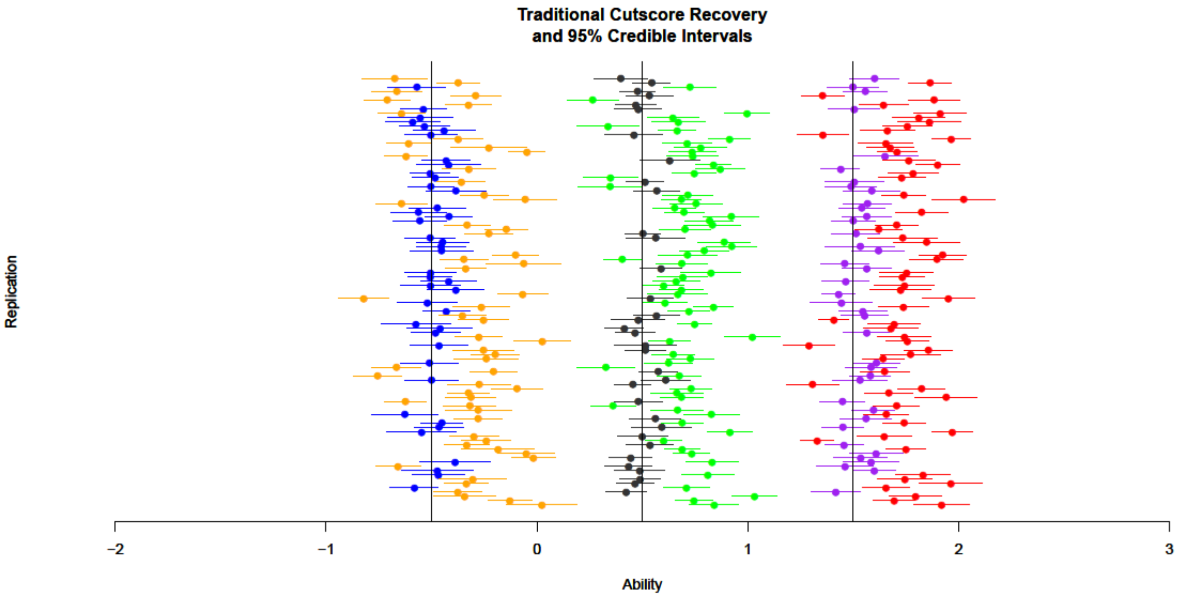

Figure A.151. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

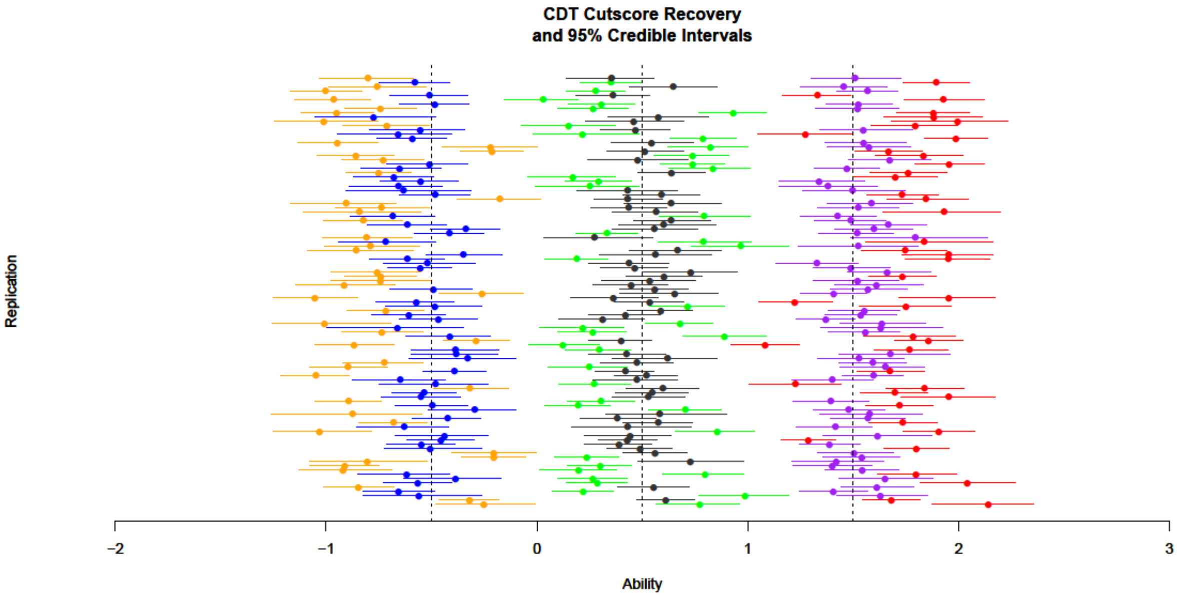

Figure A.152. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

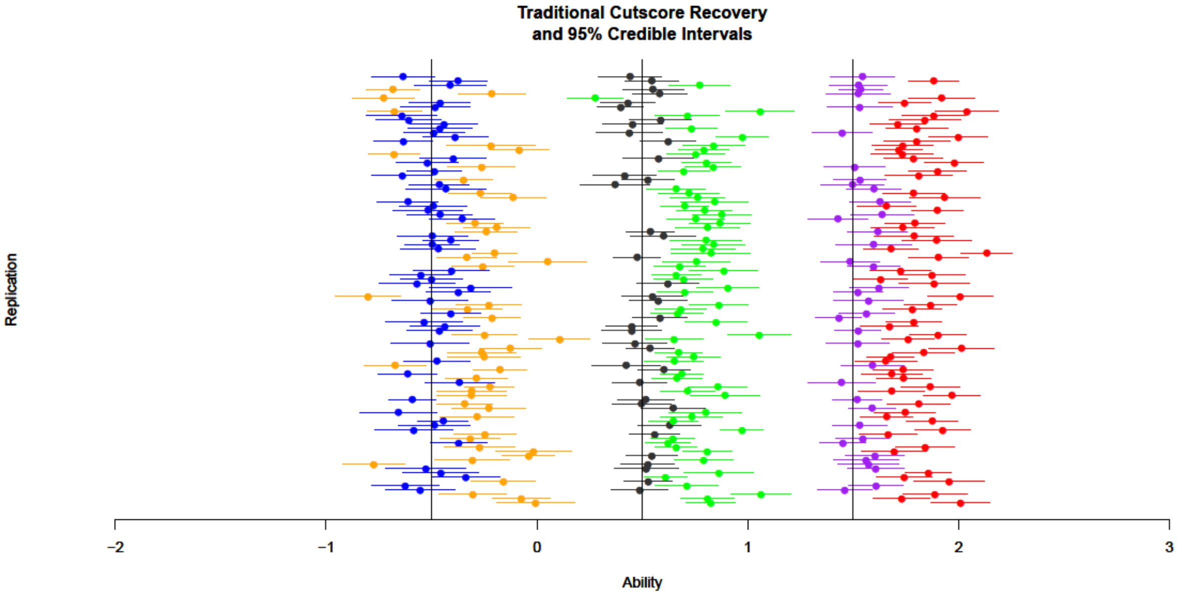

Figure A.153. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

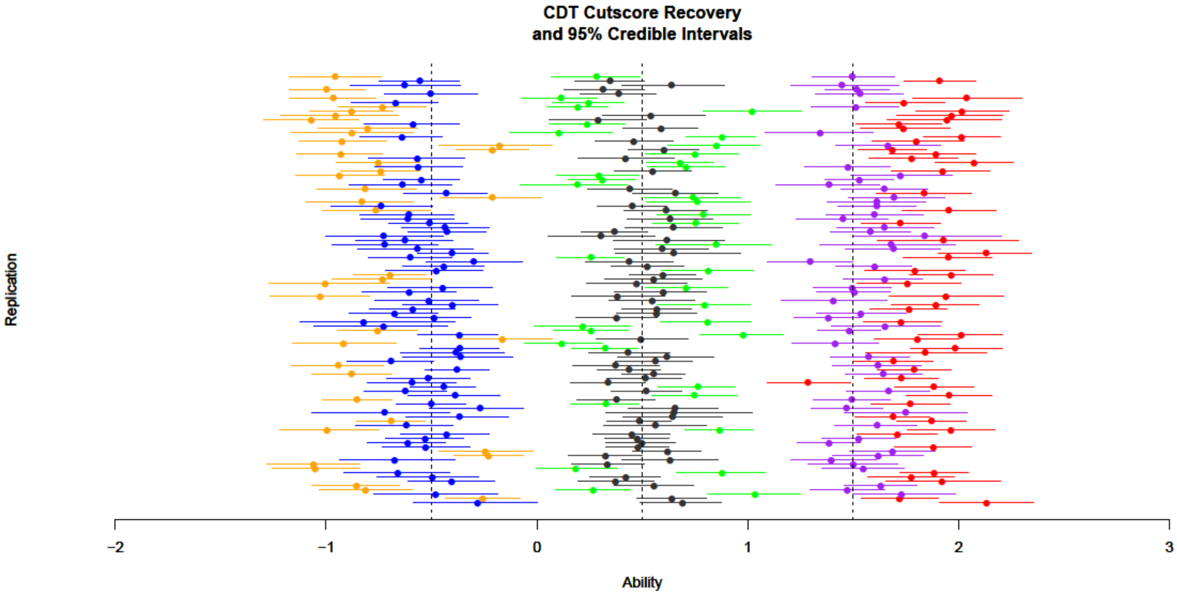

Figure A.154. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

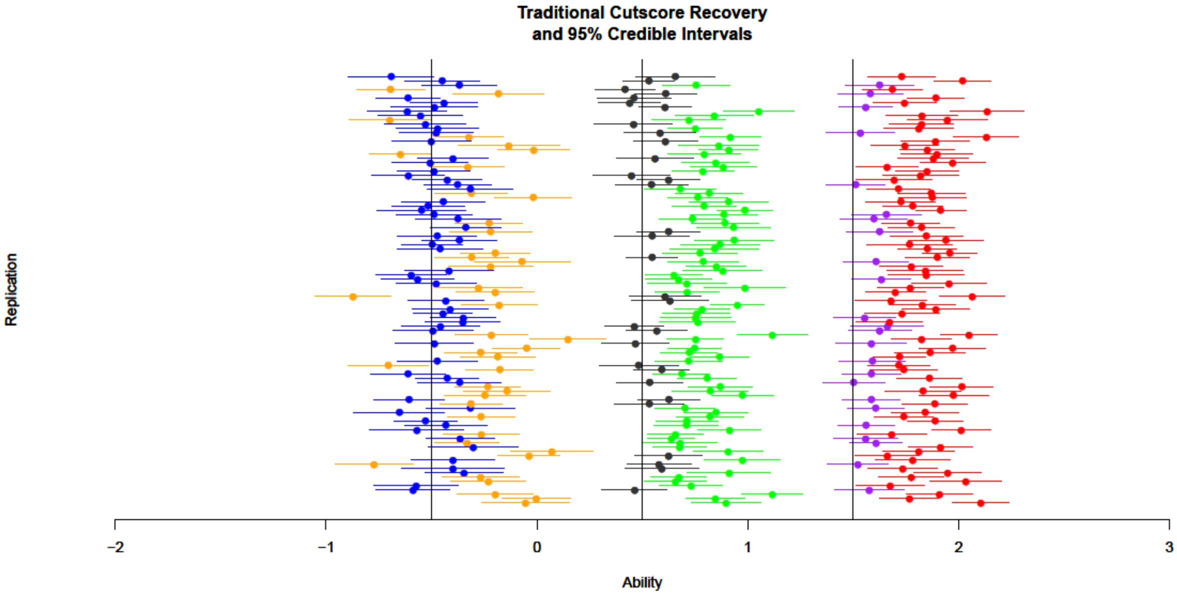

Figure A.155. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

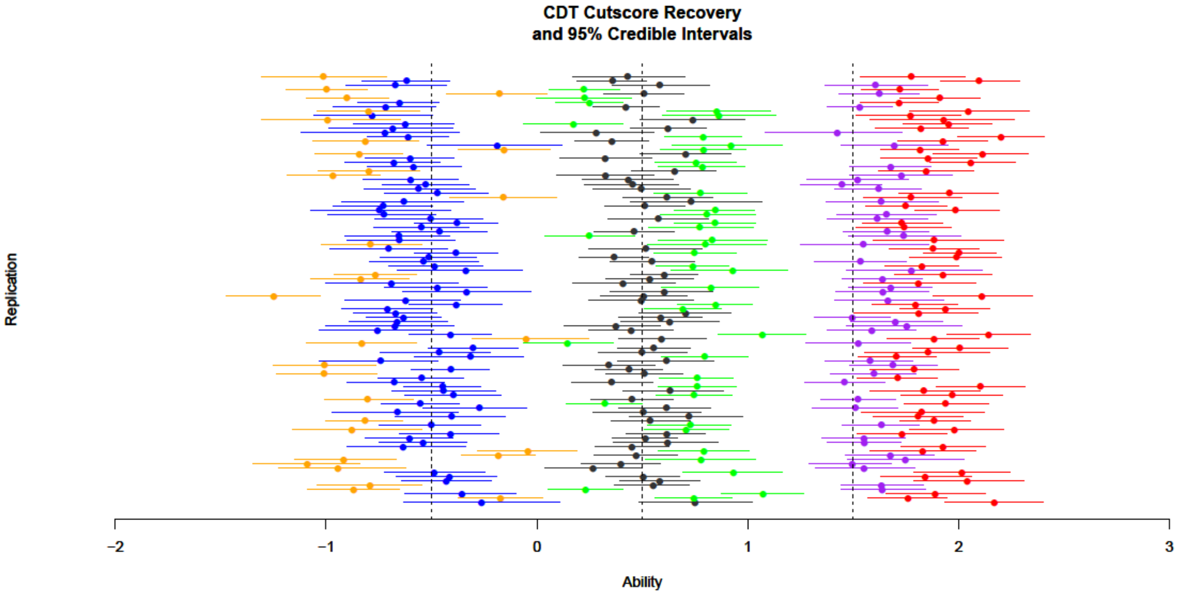

Figure A.156. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

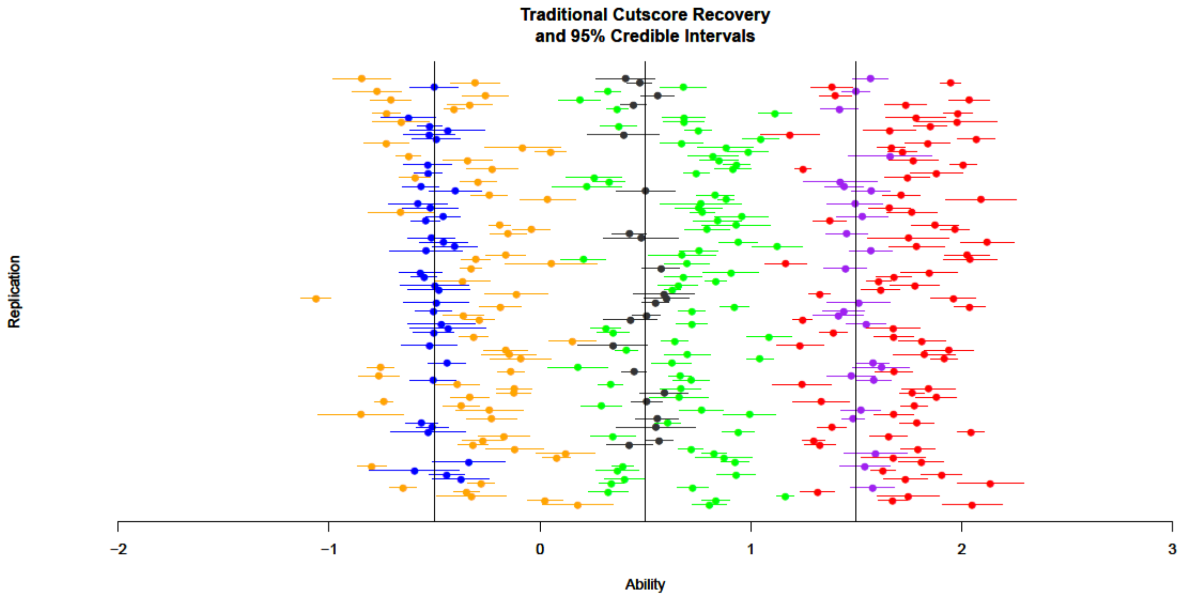

Figure A.157. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

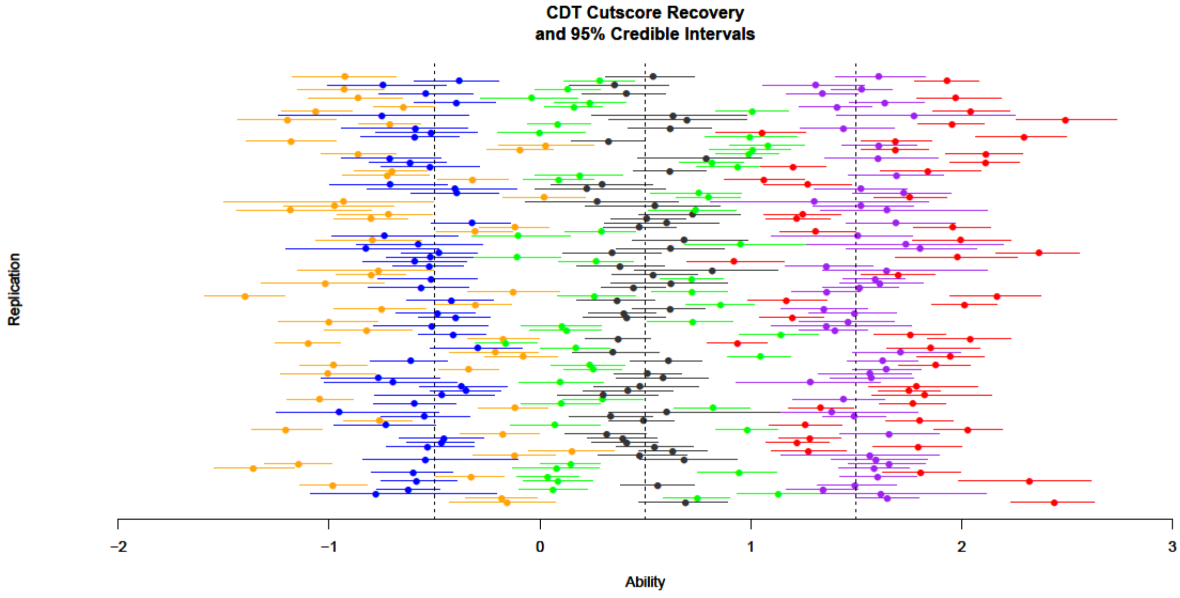

Figure A.158. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

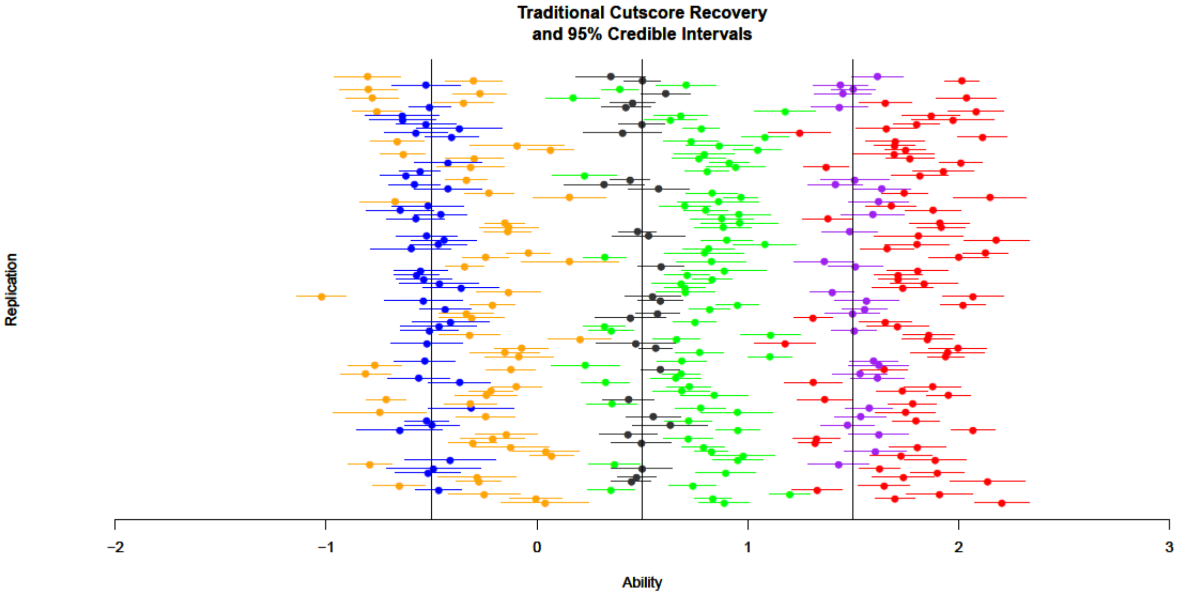

Figure A.159. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

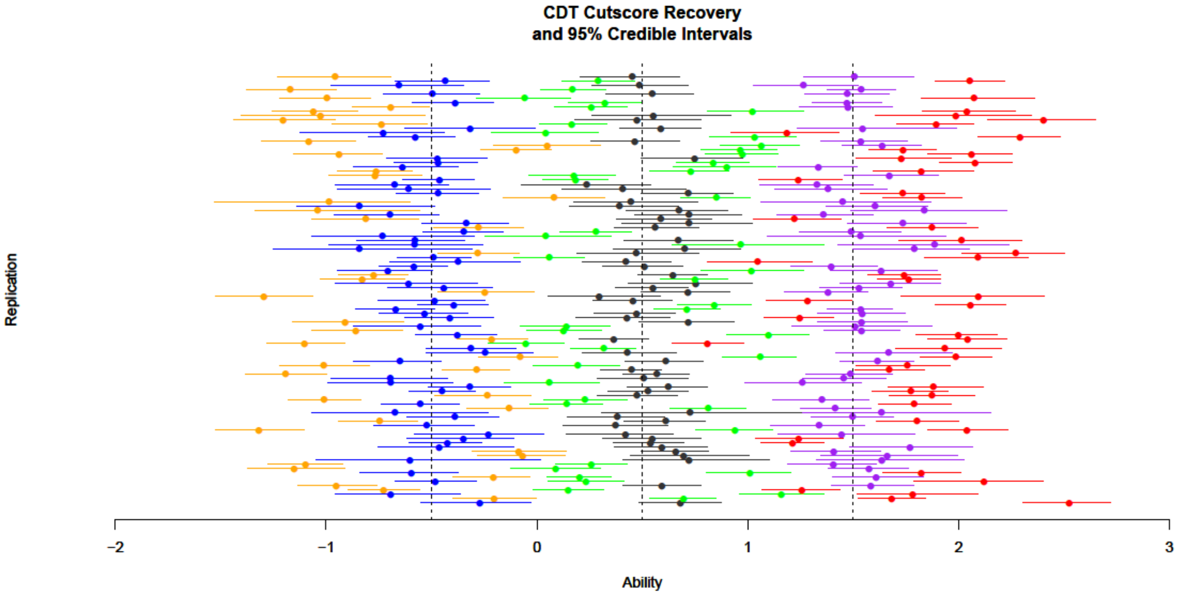

Figure A.160. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

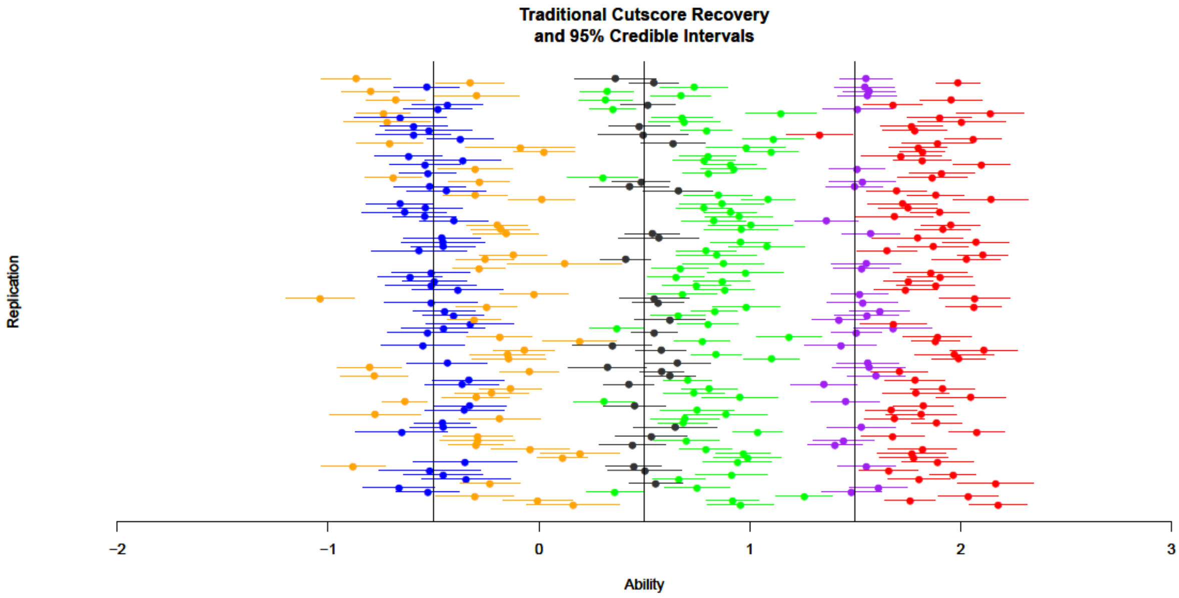

Figure A.161. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

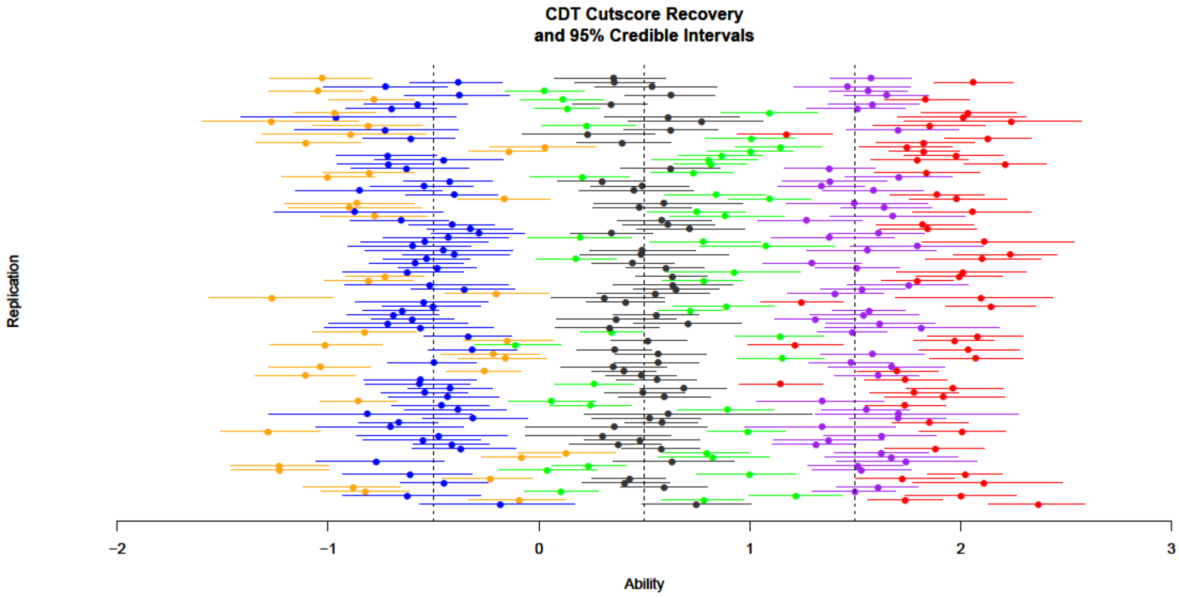

Figure A.162. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

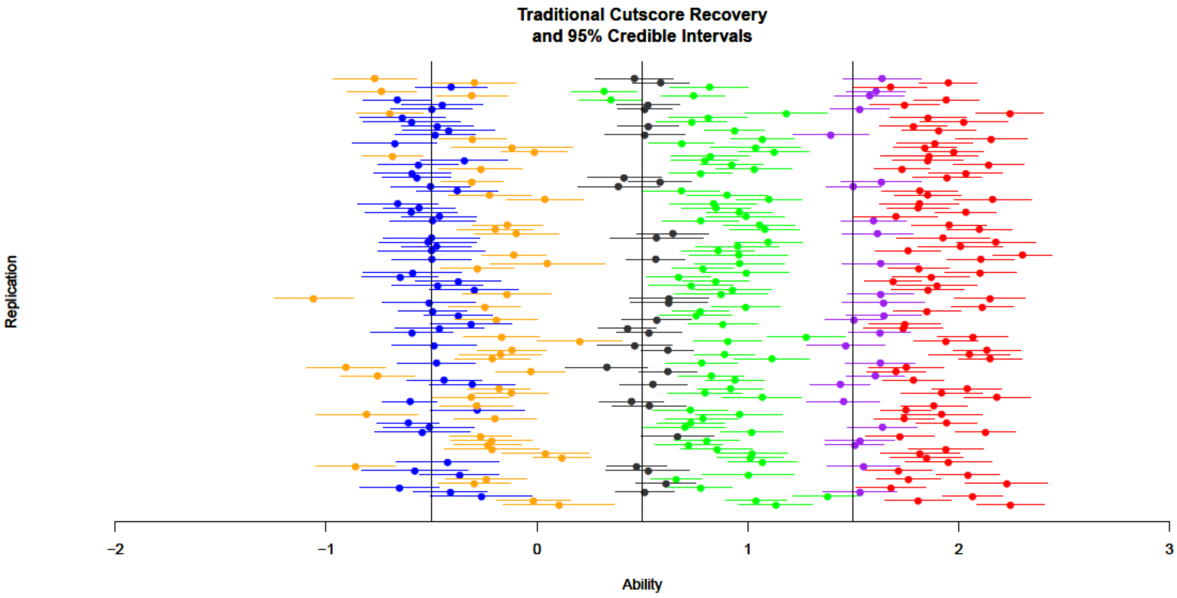

Figure A.163. N_I = 16, N_G = 3, σ^2_B = 0.2, σ^2_C = 0.3; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

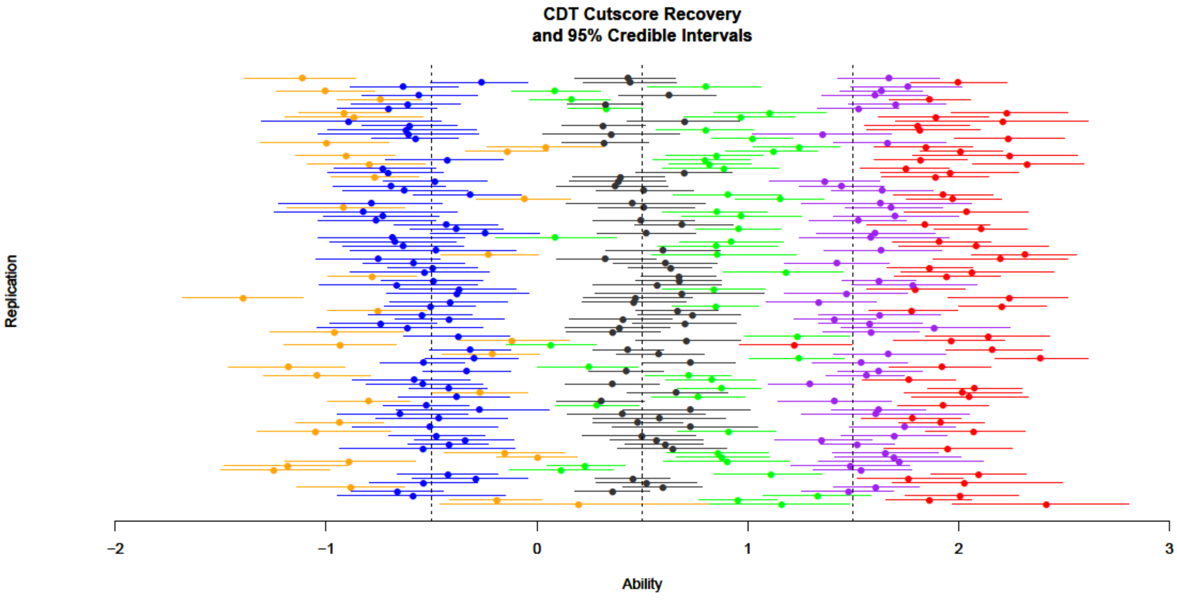

Figure A.164. N_I = 16, N_G = 3, σ^2_B = 0.2, σ^2_C = 0.3; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

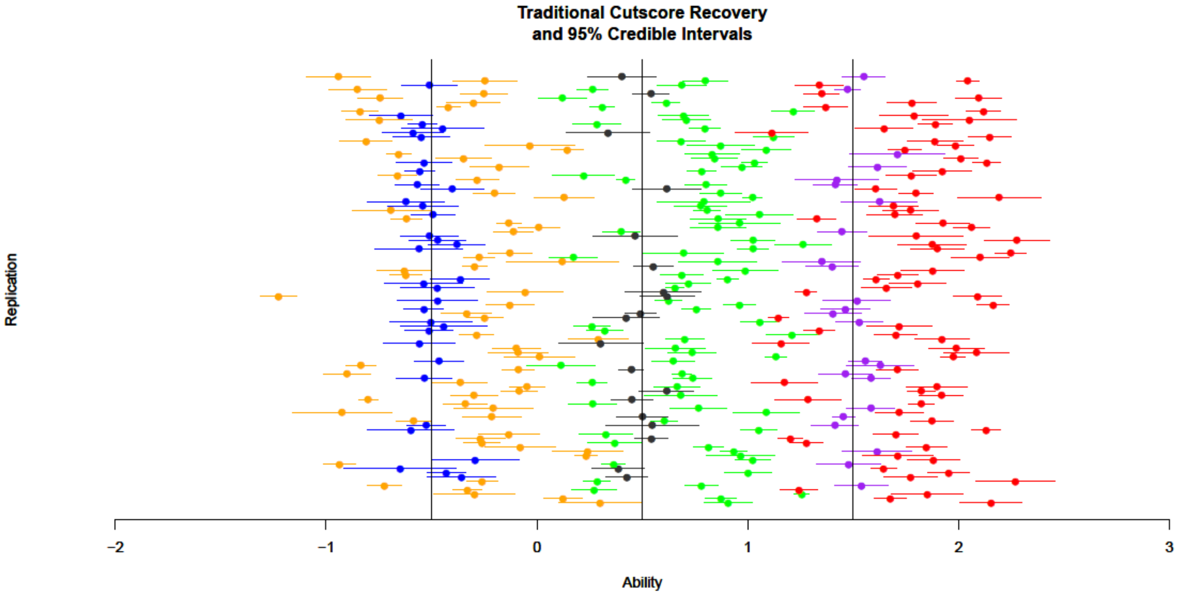

Figure A.165. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

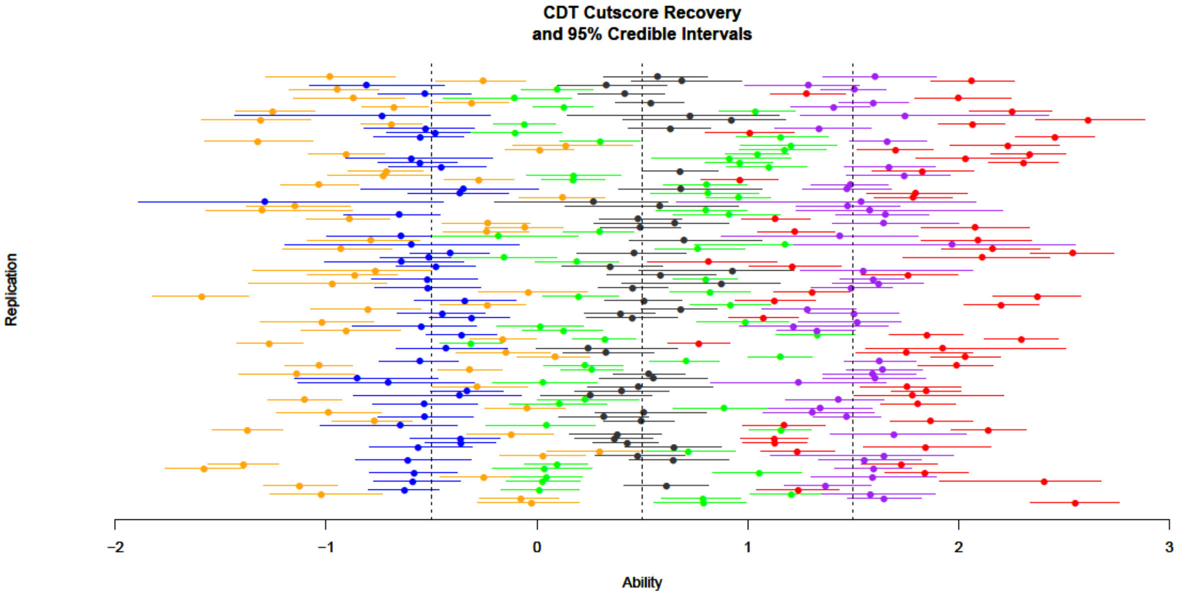

Figure A.166. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

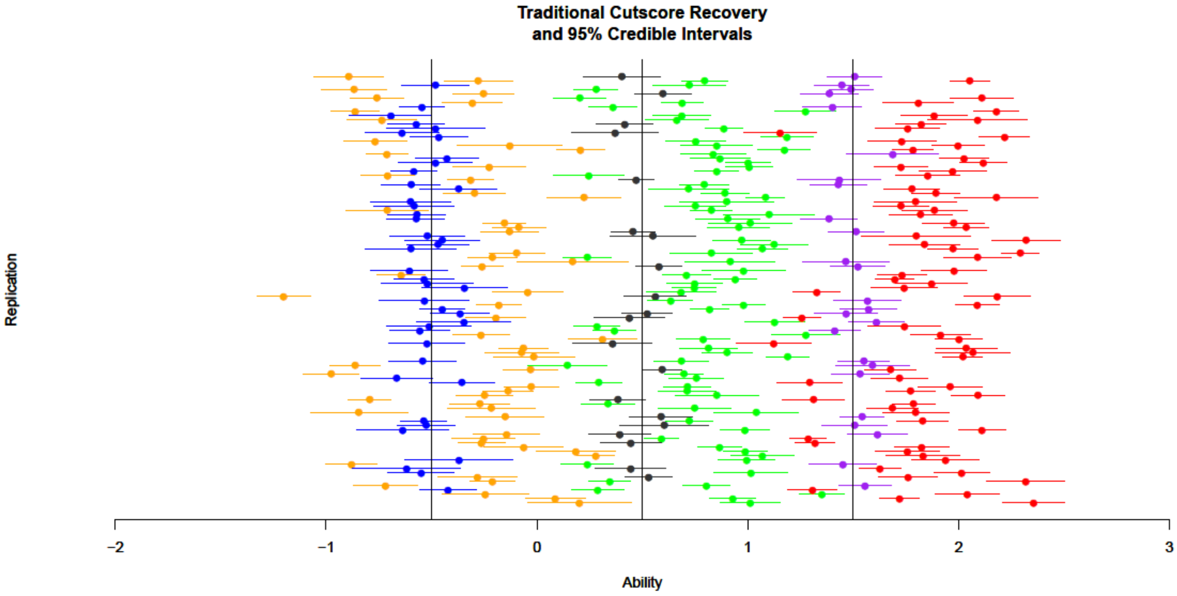

Figure A.167. N_I = 16, N_G = 3, $\sigma^2{}_B$ = 0.3, $\sigma^2{}_C$ = 0.1; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

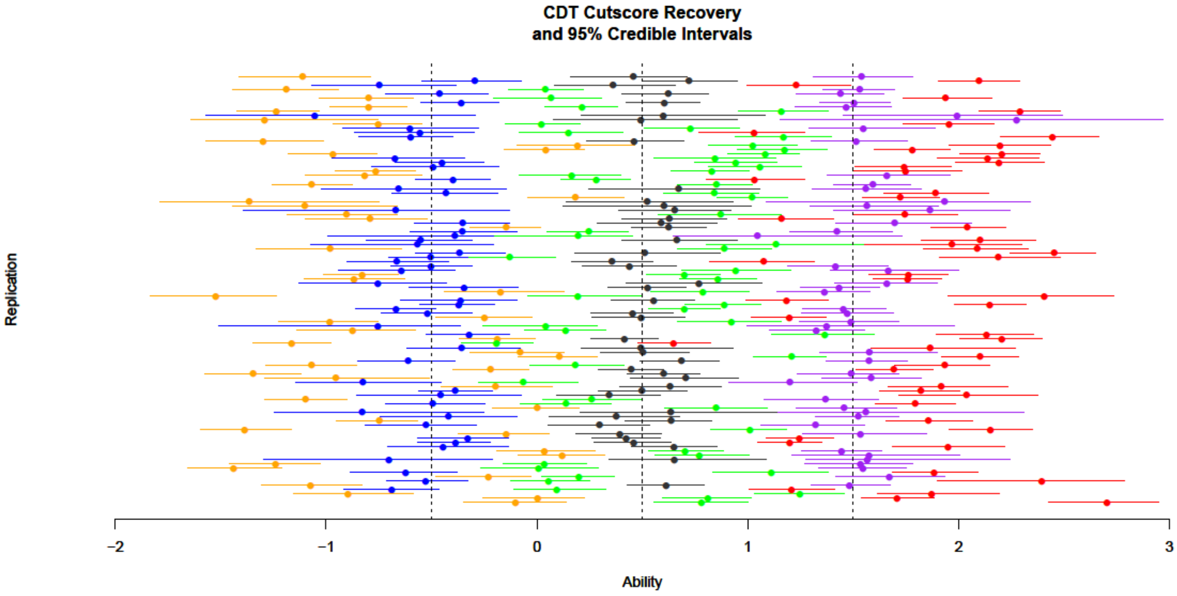

Figure A.168. N_I = 16, N_G = 3, $\sigma^2{}_B$ = 0.3, $\sigma^2{}_C$ = 0.1; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

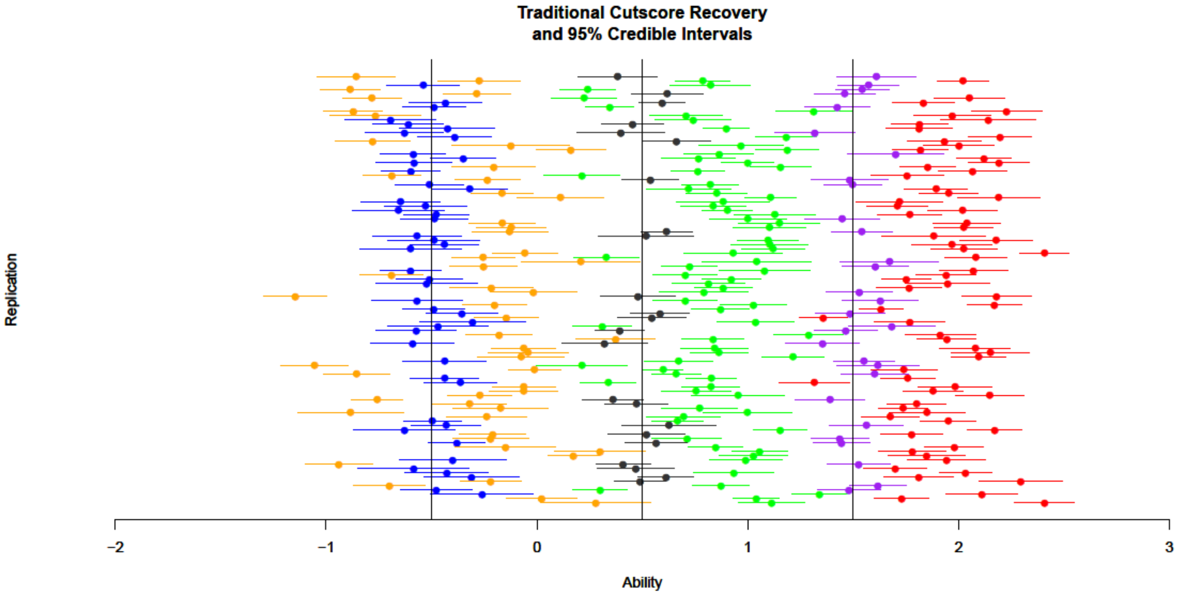

Figure A.169. N_I = 16, N_G = 3, σ^2_B = 0.3, σ^2_C = 0.2; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

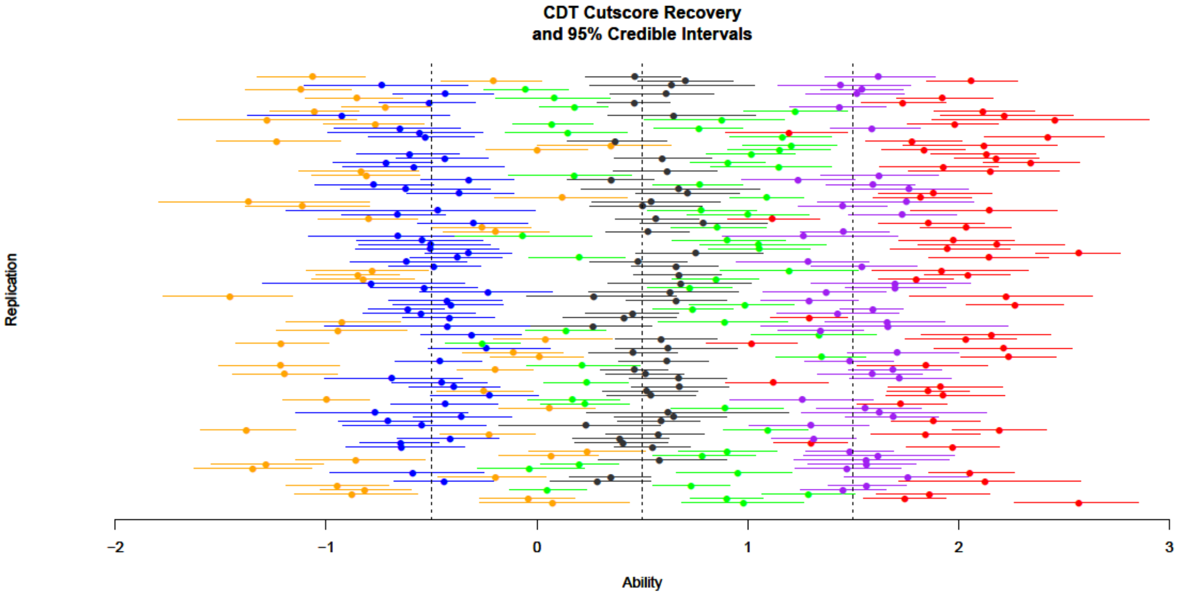

Figure A.170. N_I = 16, N_G = 3, σ^2_B = 0.3, σ^2_C = 0.2; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

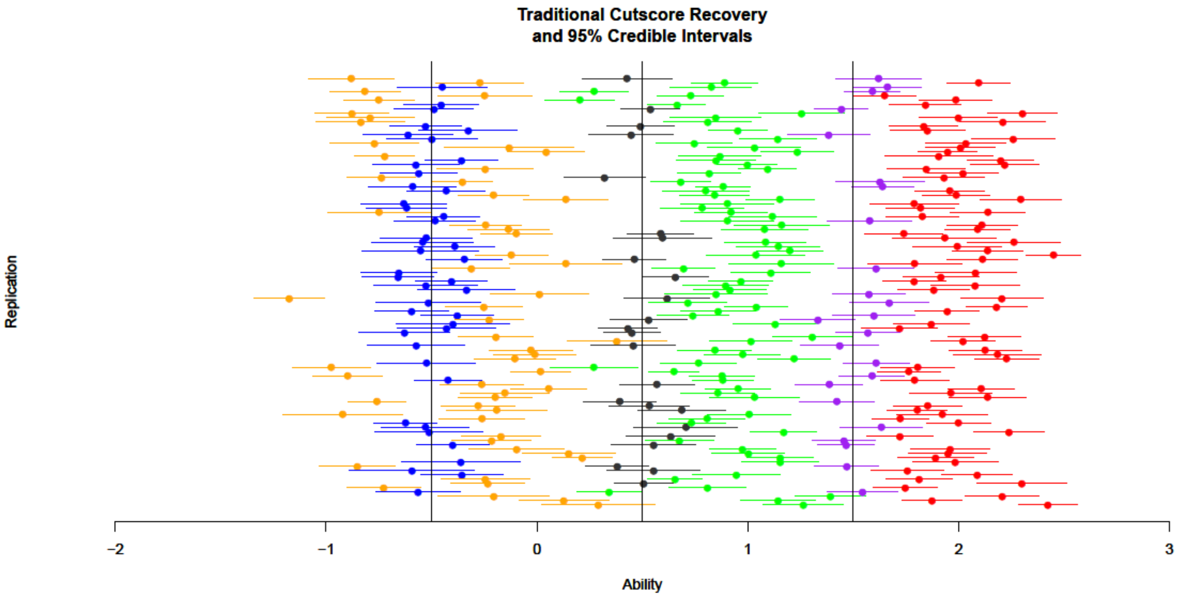

Figure A.171. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

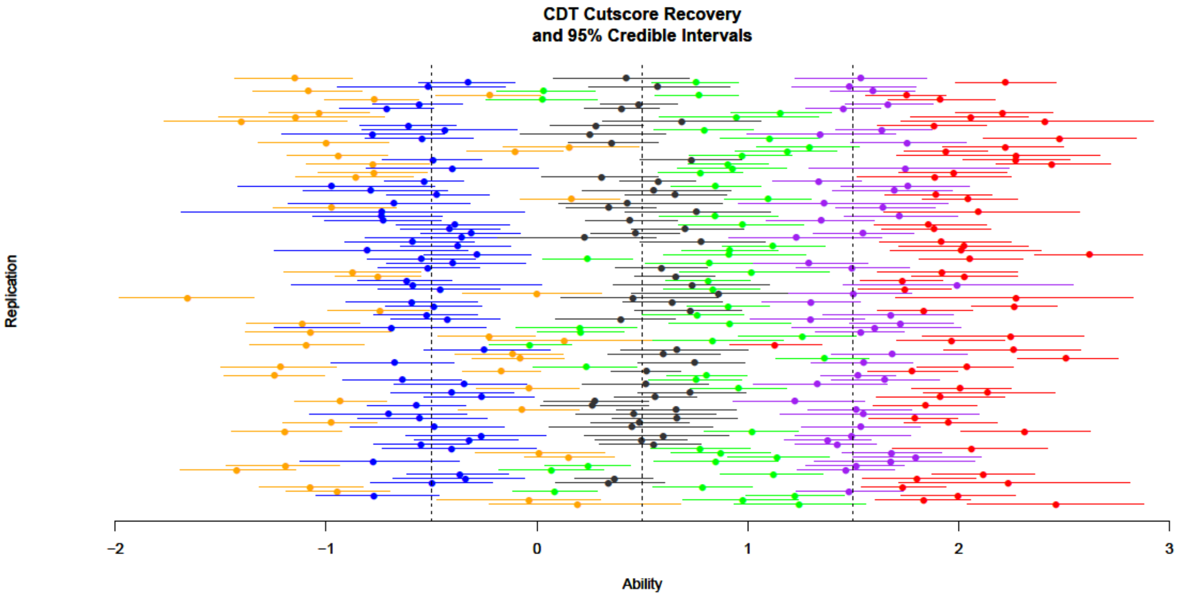

Figure A.172. $N_I = 16$, $N_G = 3$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

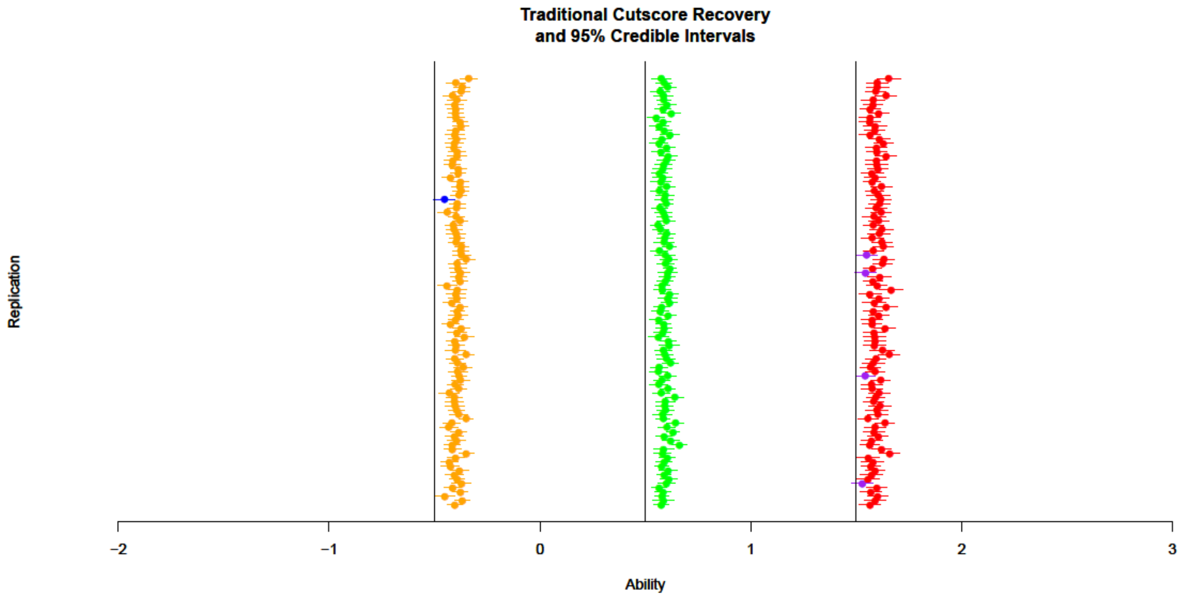

Figure A.173. N_I = 16, N_G = 4, σ^2_B = 0.0, σ^2_C = 0.0; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

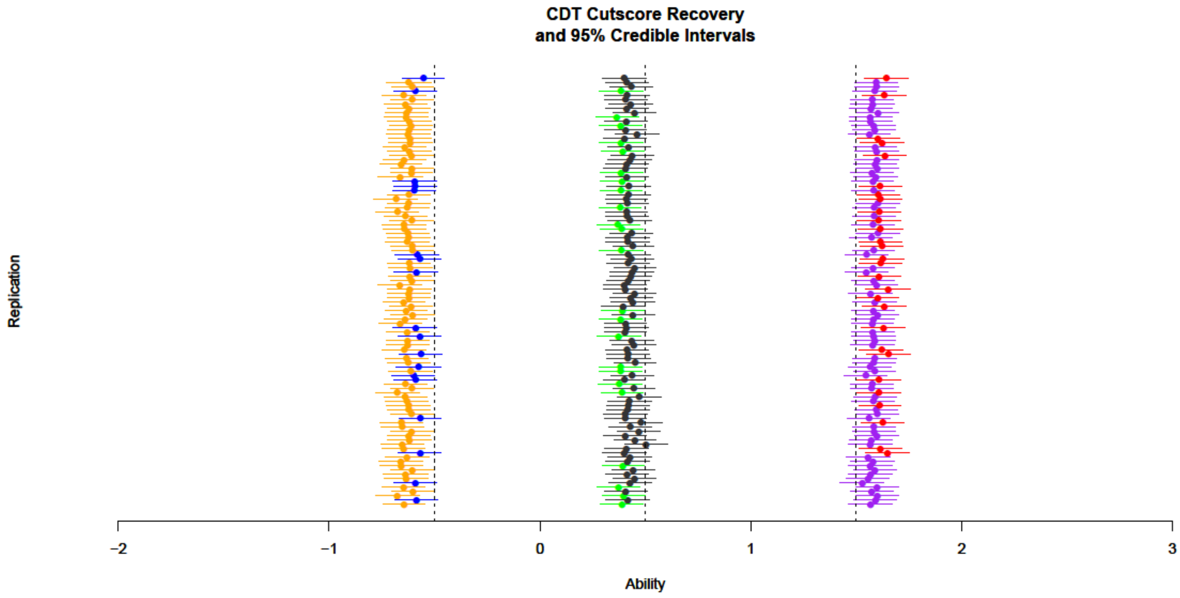

Figure A.174. N_I = 16, N_G = 4, σ^2_B = 0.0, σ^2_C = 0.0; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

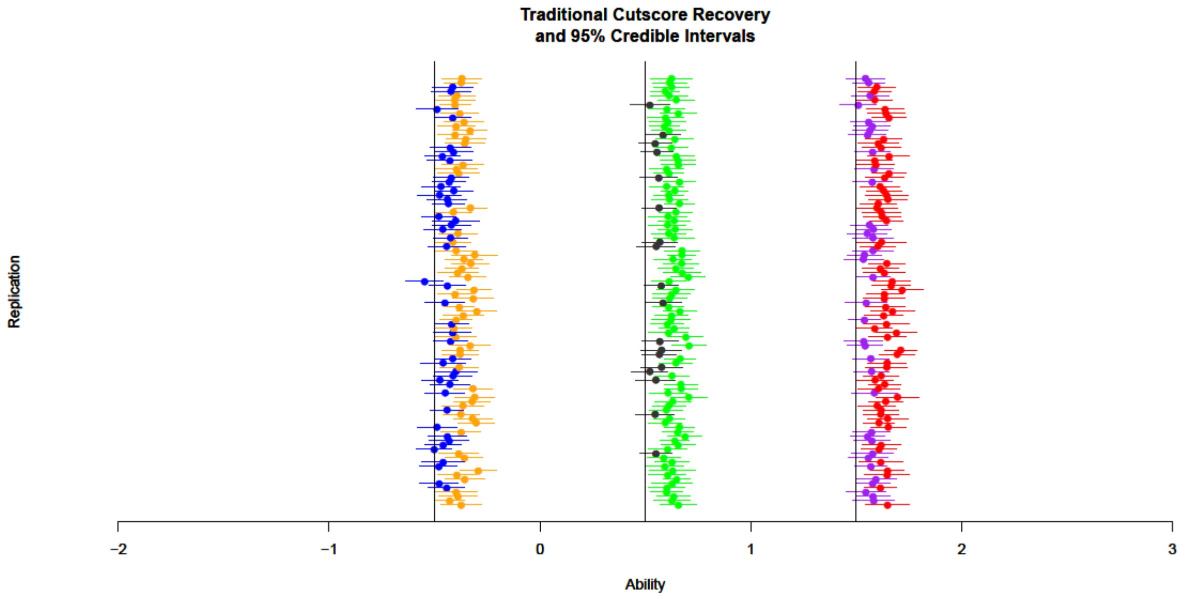

Figure A.175. N_I = 16, N_G = 4, σ^2_B = 0.0, σ^2_C = 0.1; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

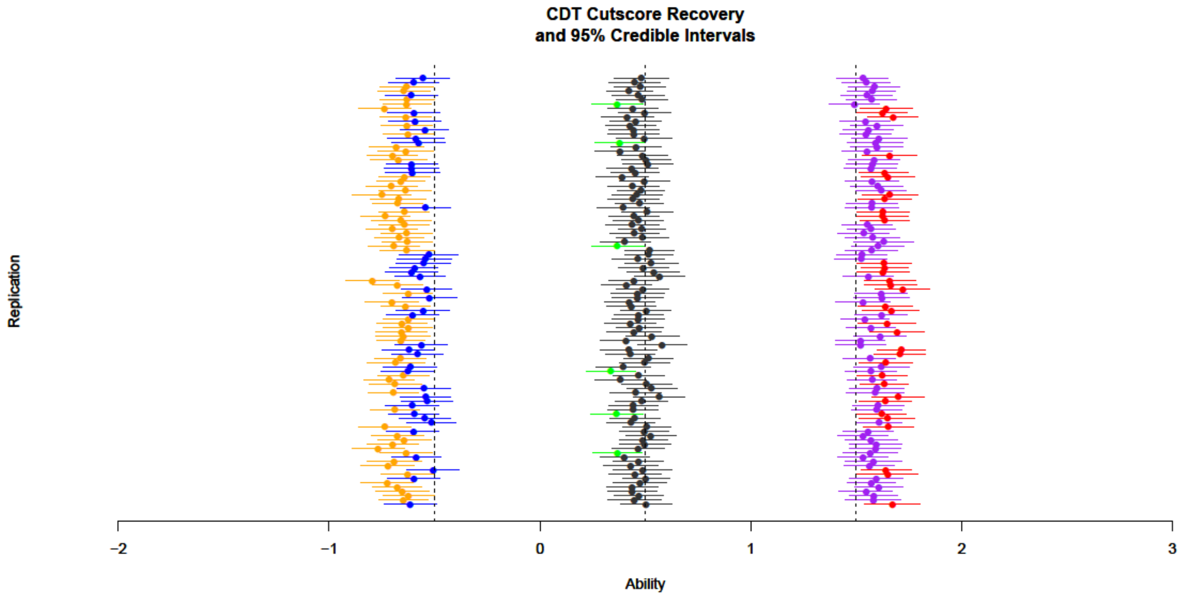

Figure A.176. N_I = 16, N_G = 4, σ^2_B = 0.0, σ^2_C = 0.1; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

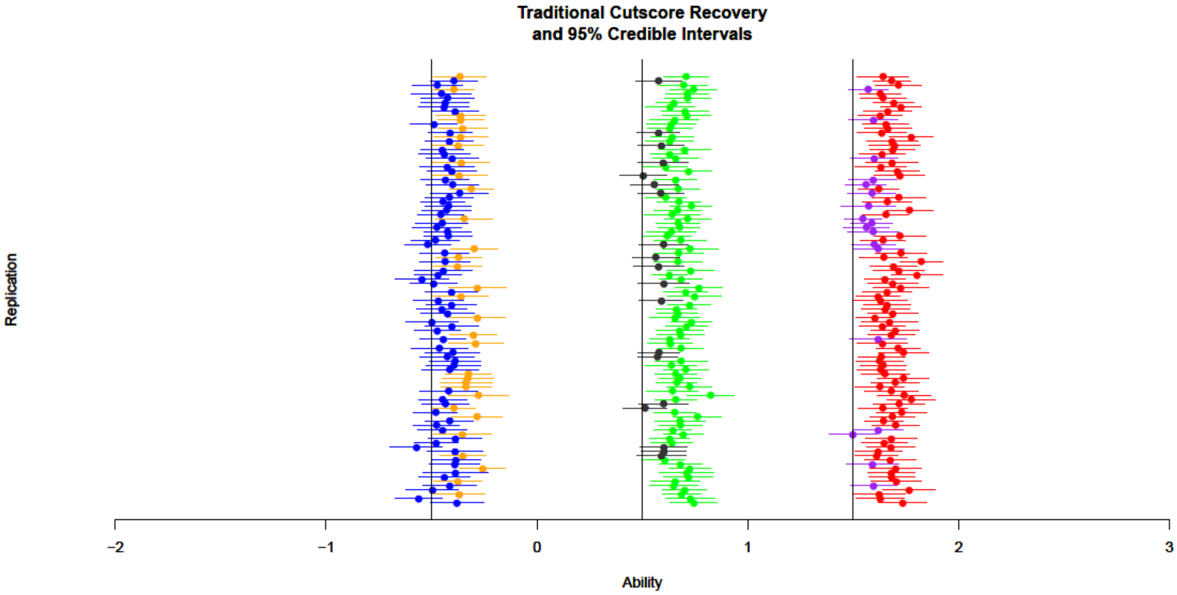

Figure A.177. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

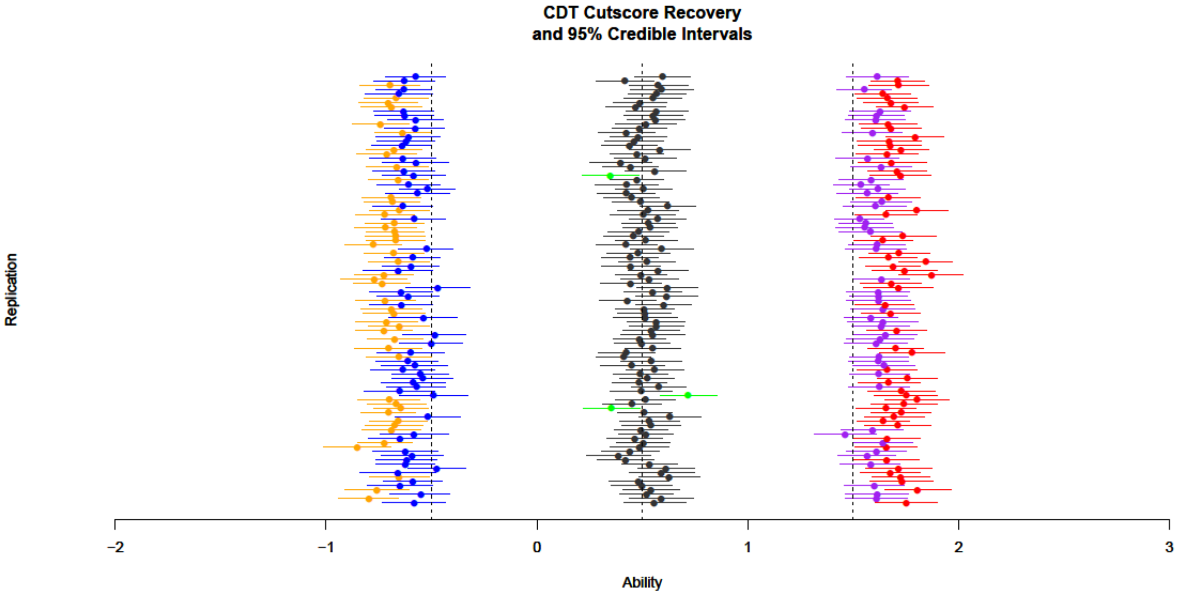

Figure A.178. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

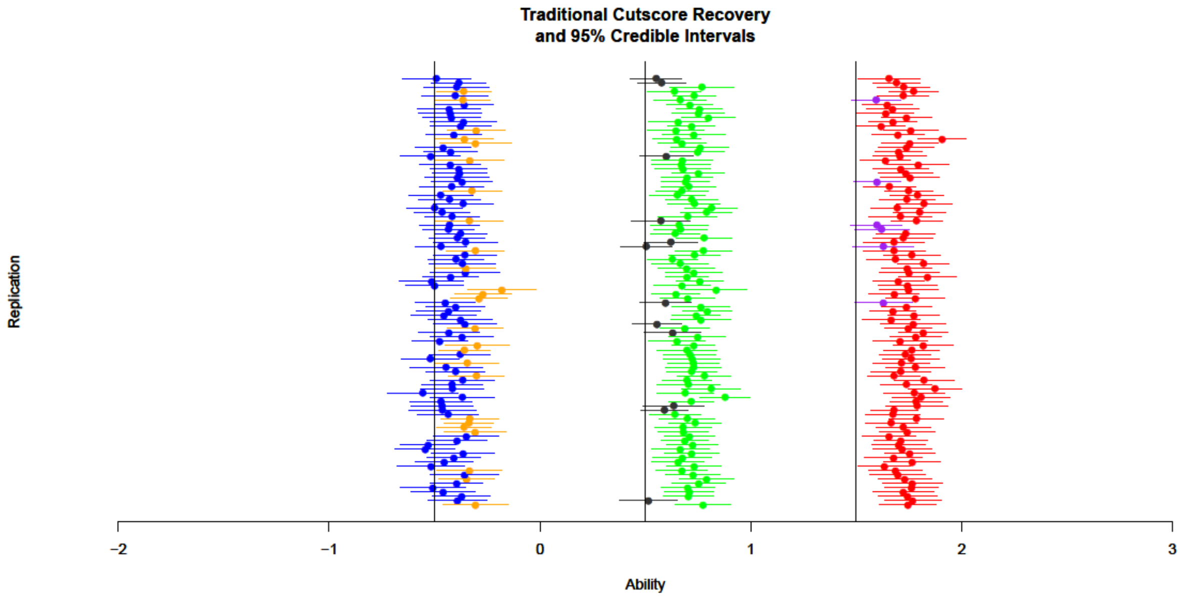

Figure A.179. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

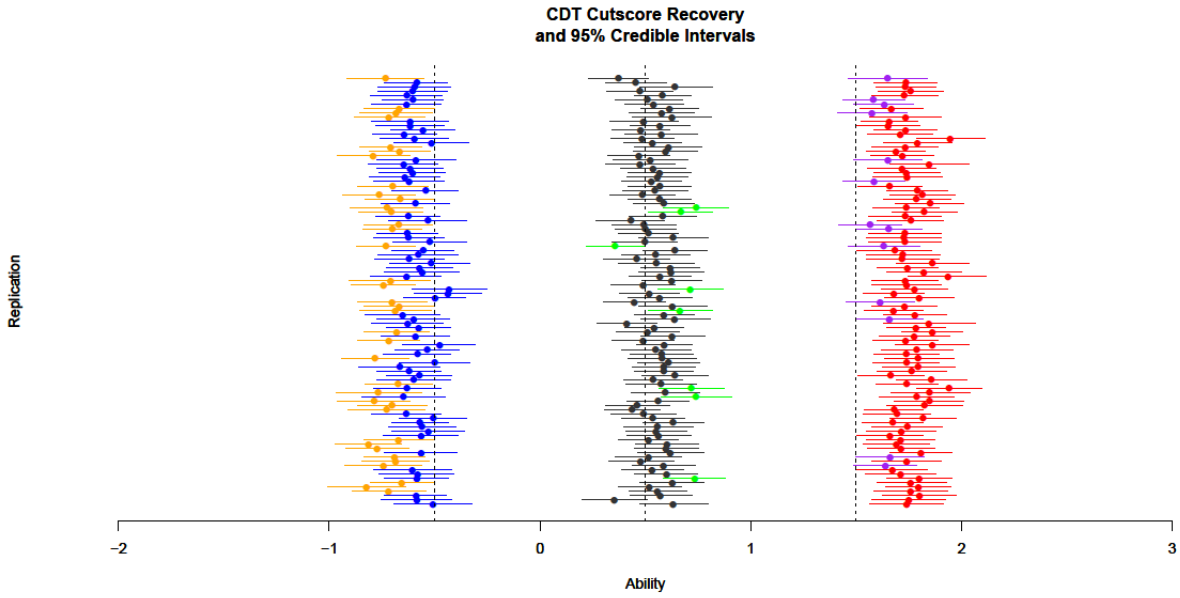

Figure A.180. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.0$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

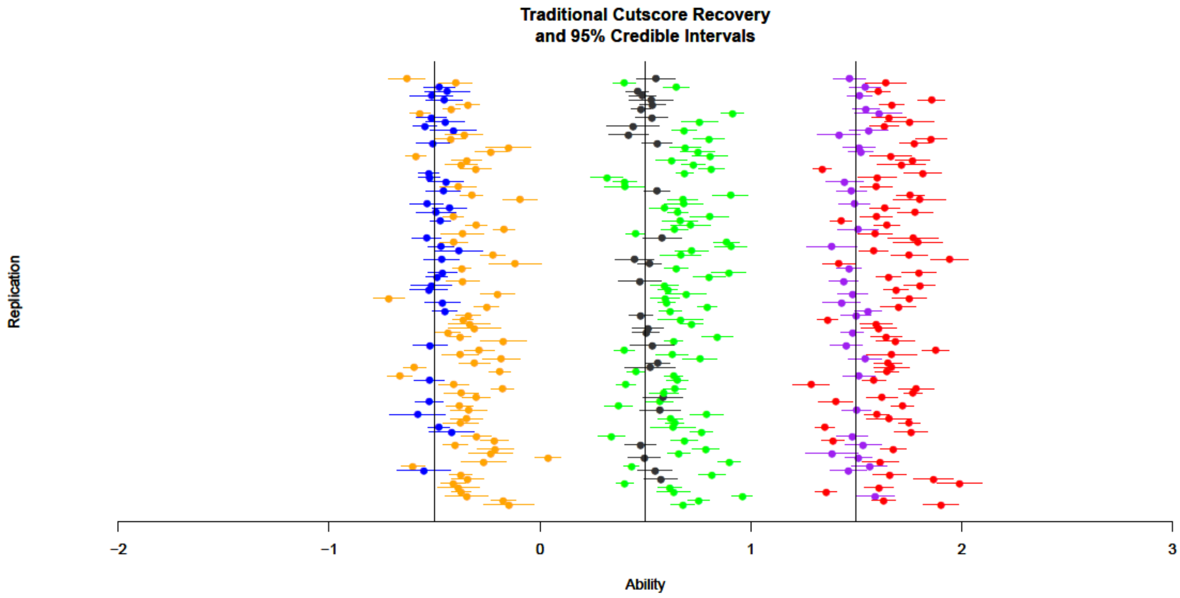

Figure A.181. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

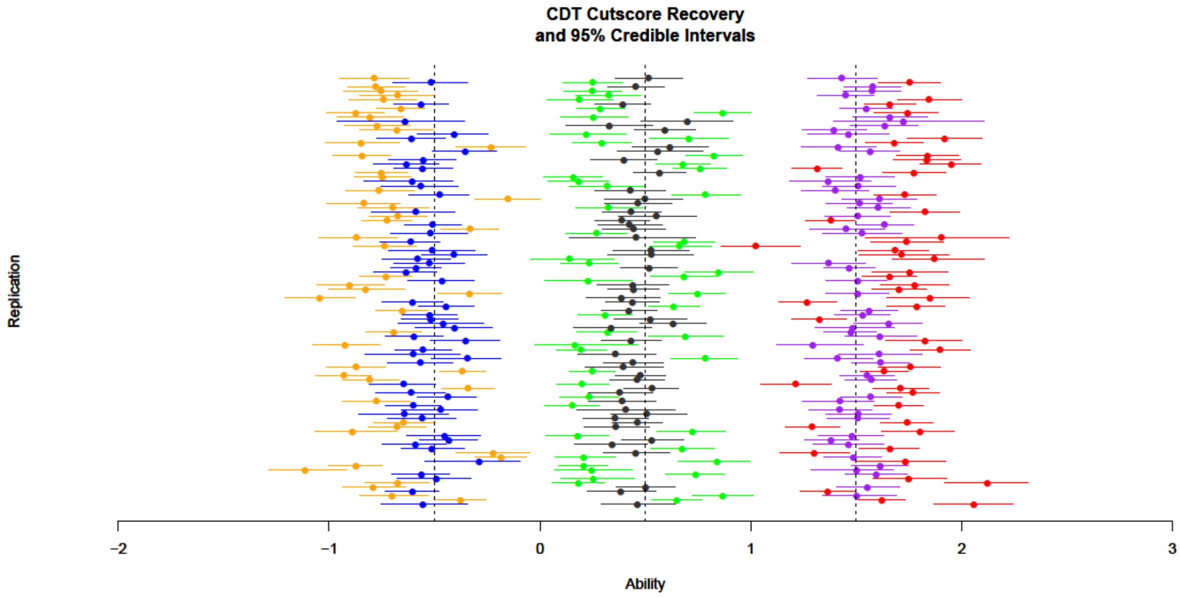

Figure A.182. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

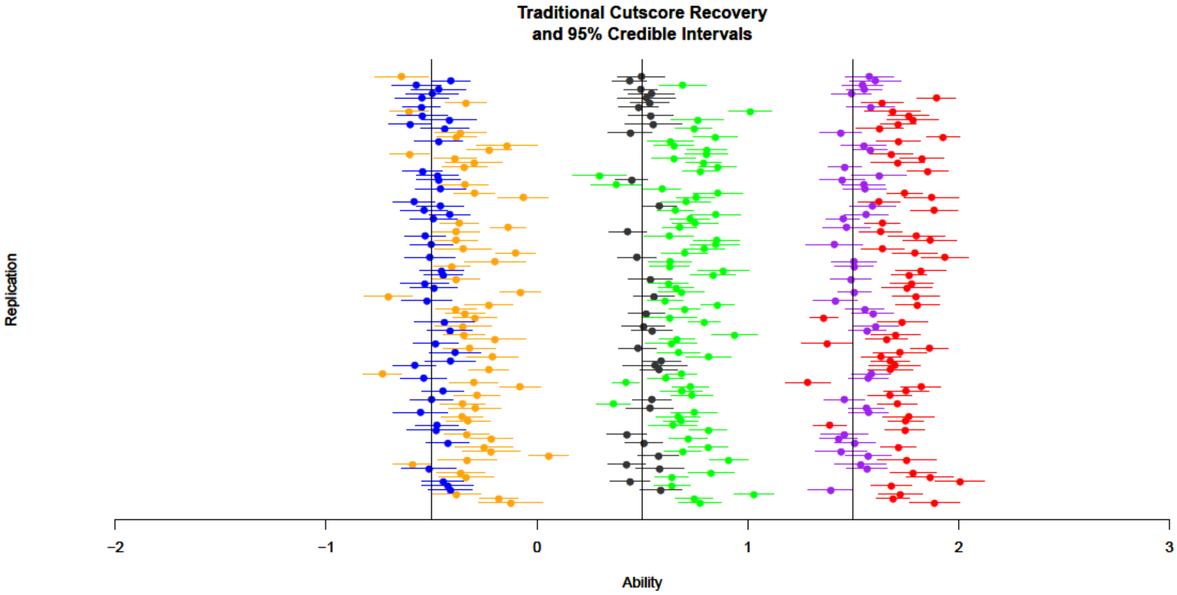

Figure A.183. N_I = 16, N_G = 4, σ^2_B = 0.1, σ^2_C = 0.1; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

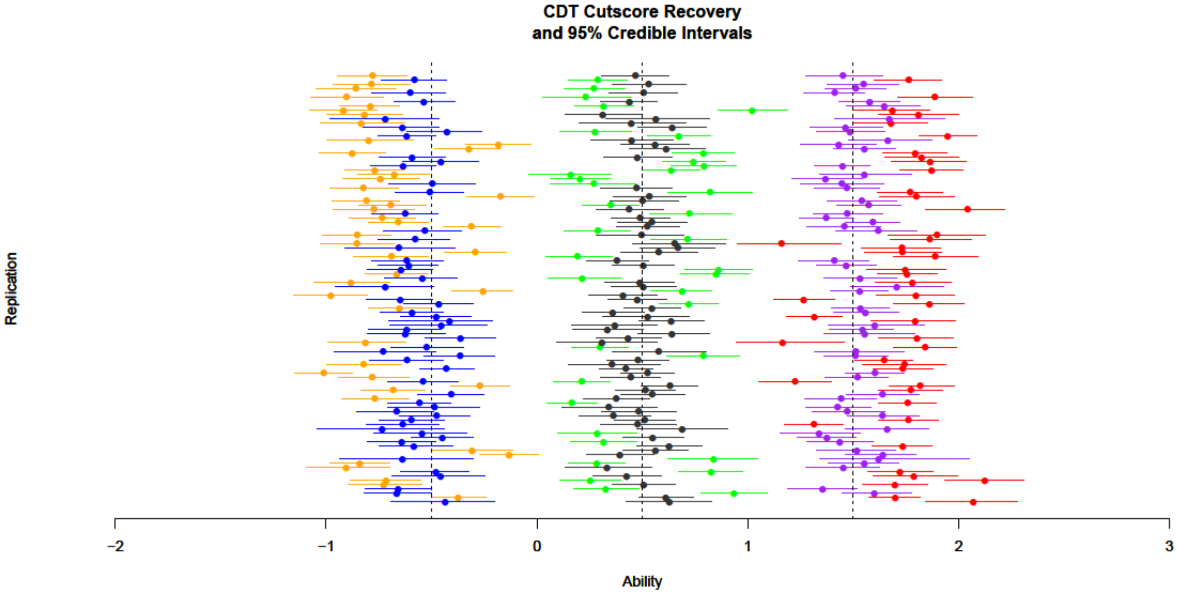

Figure A.184. N_I = 16, N_G = 4, σ^2_B = 0.1, σ^2_C = 0.1; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

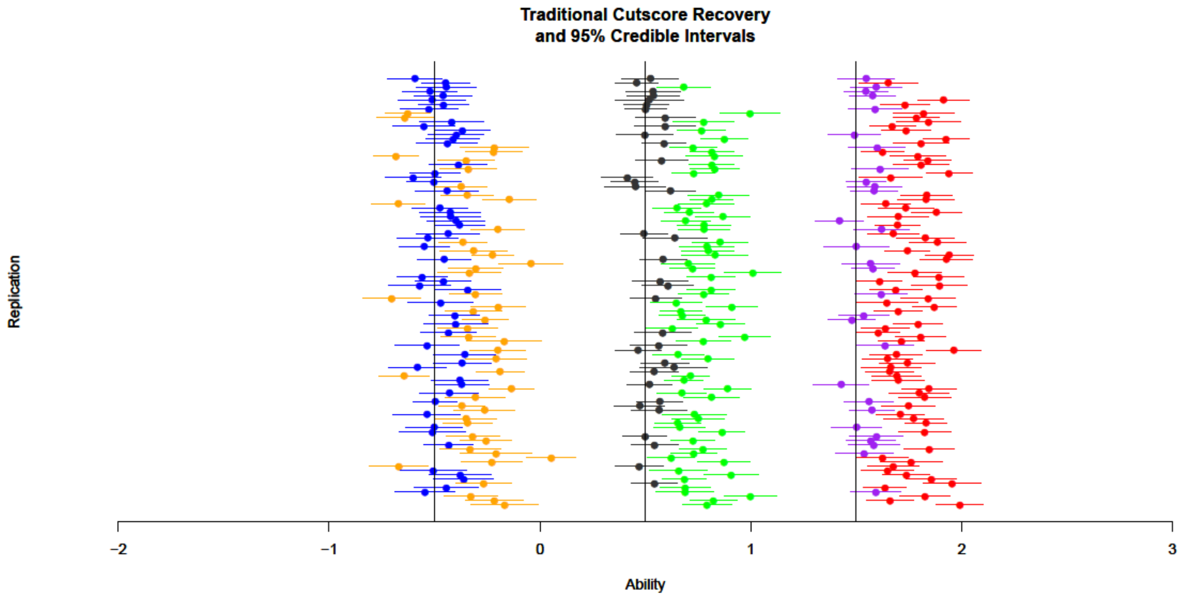

Figure A.185. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

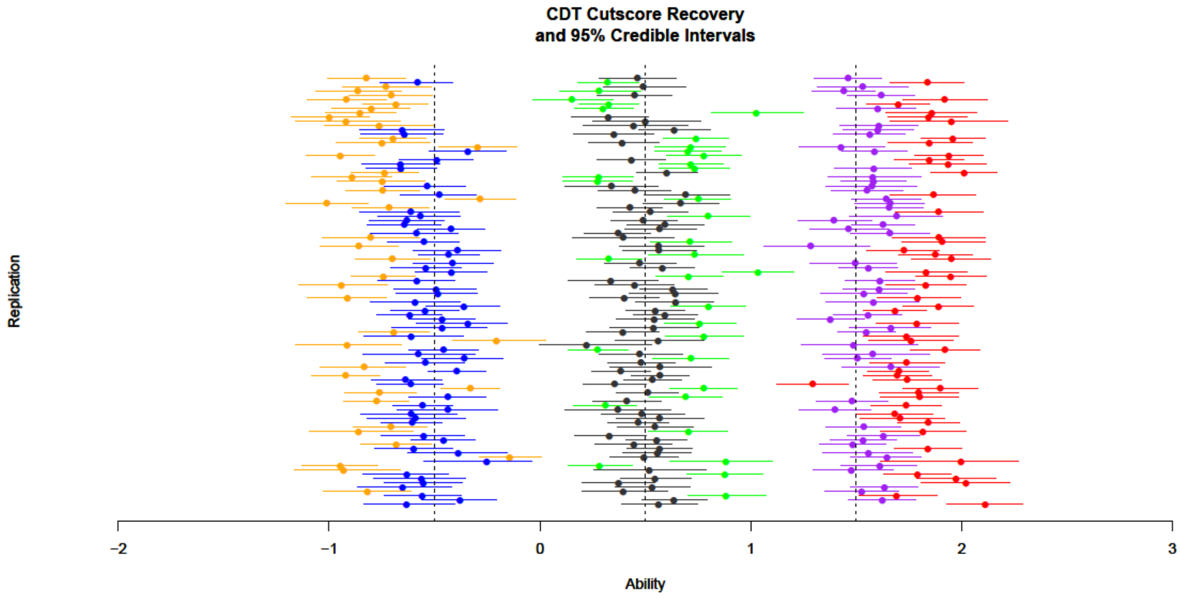

Figure A.186. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

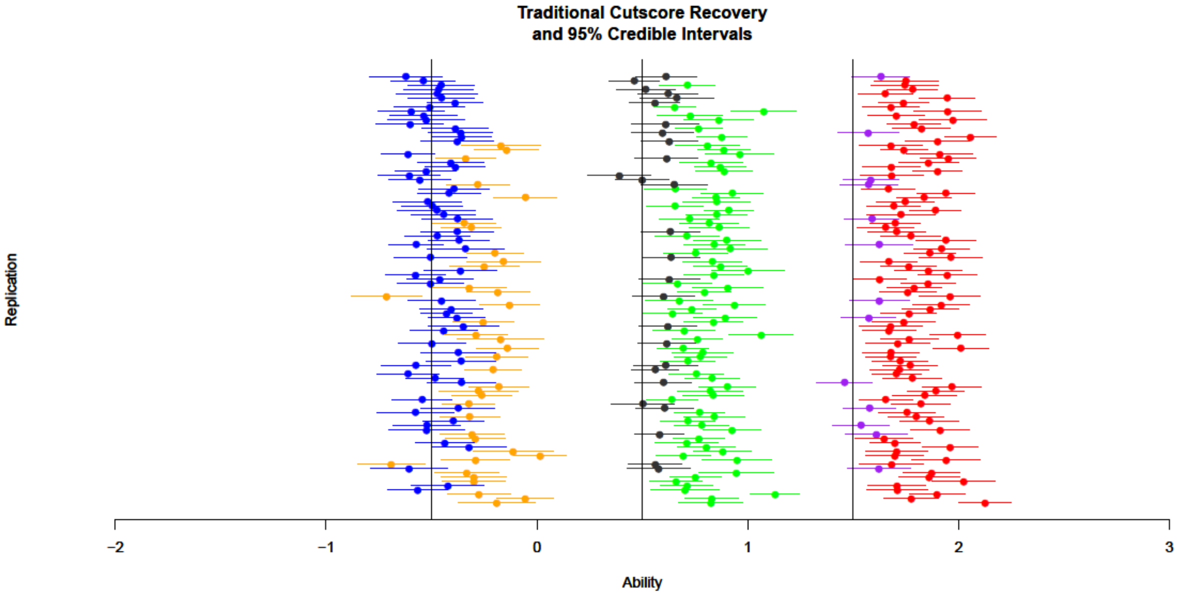

Figure A.187. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

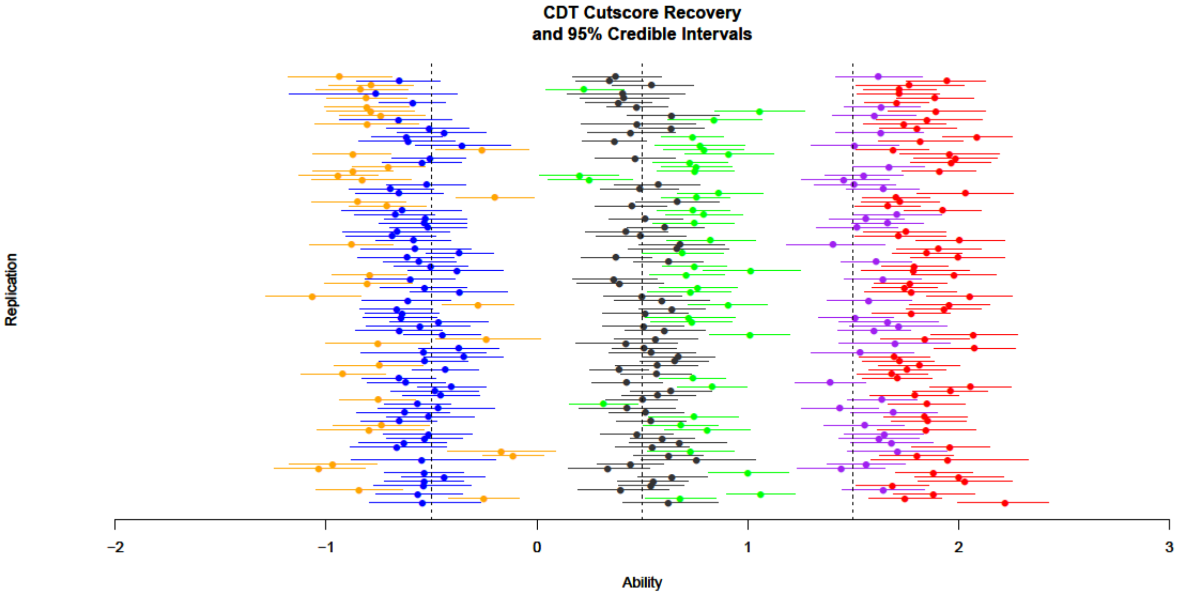

Figure A.188. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.1$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

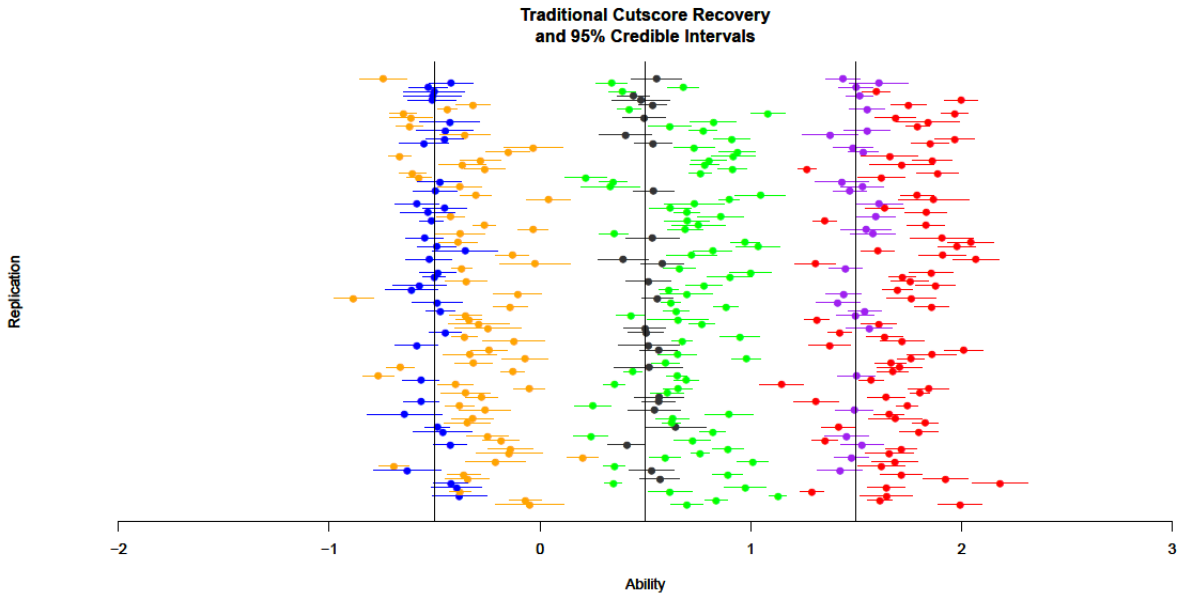

Figure A.189. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

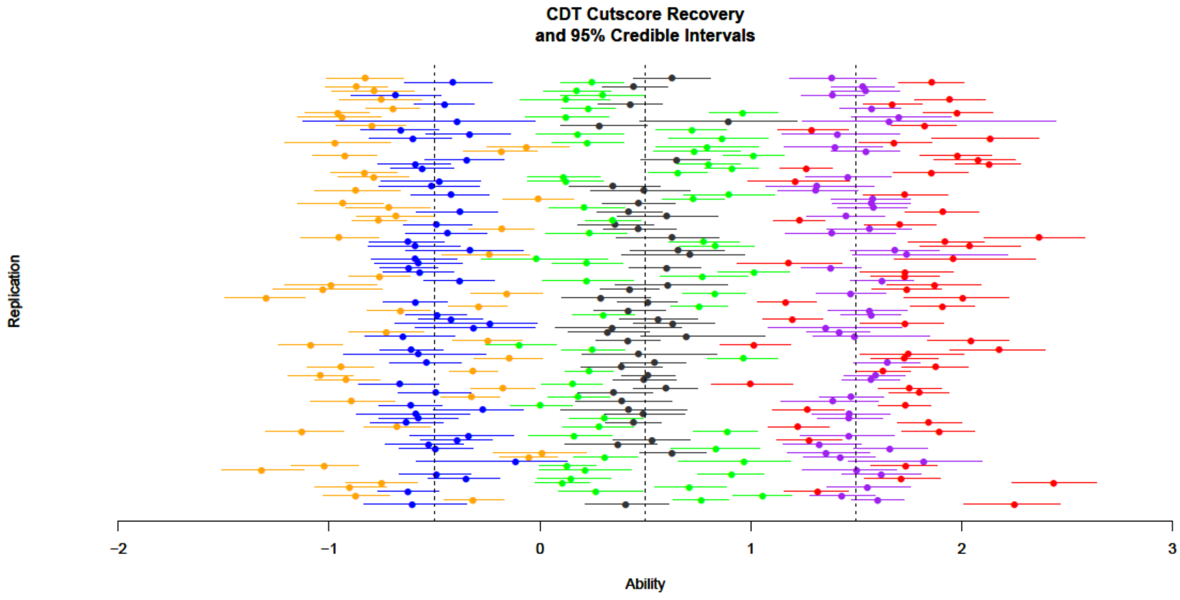

Figure A.190. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

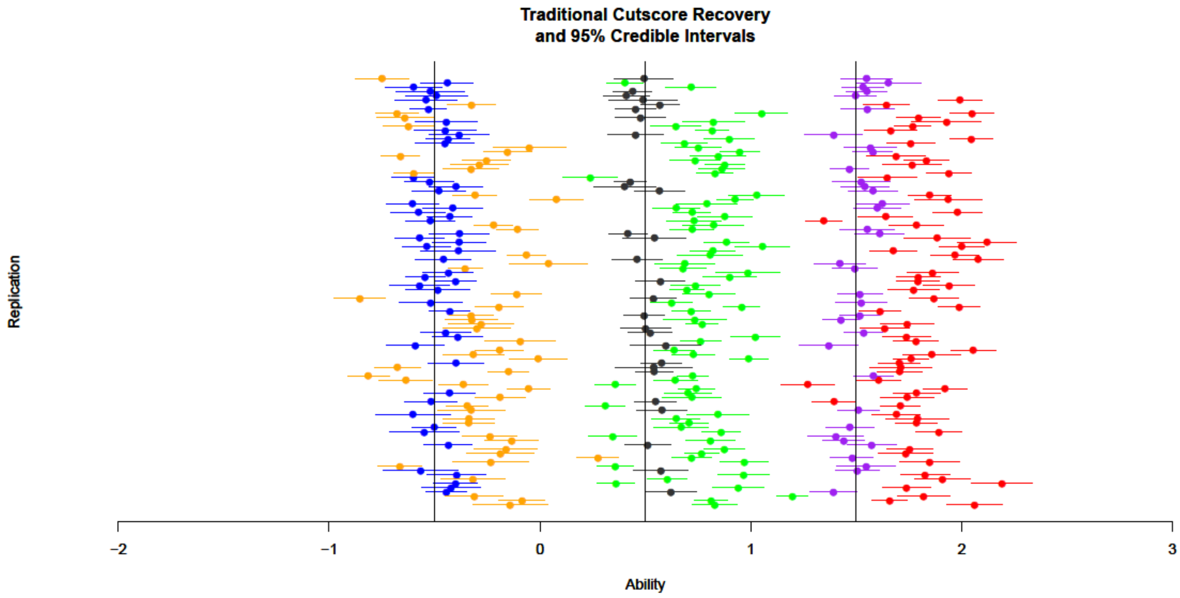

Figure A.191. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

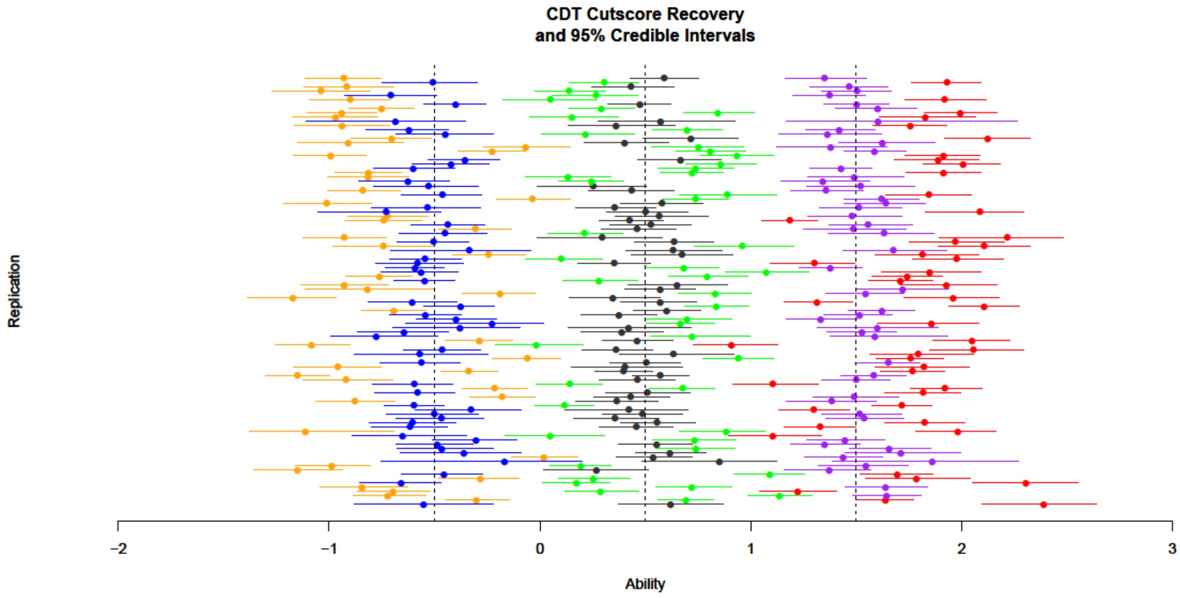

Figure A.192. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

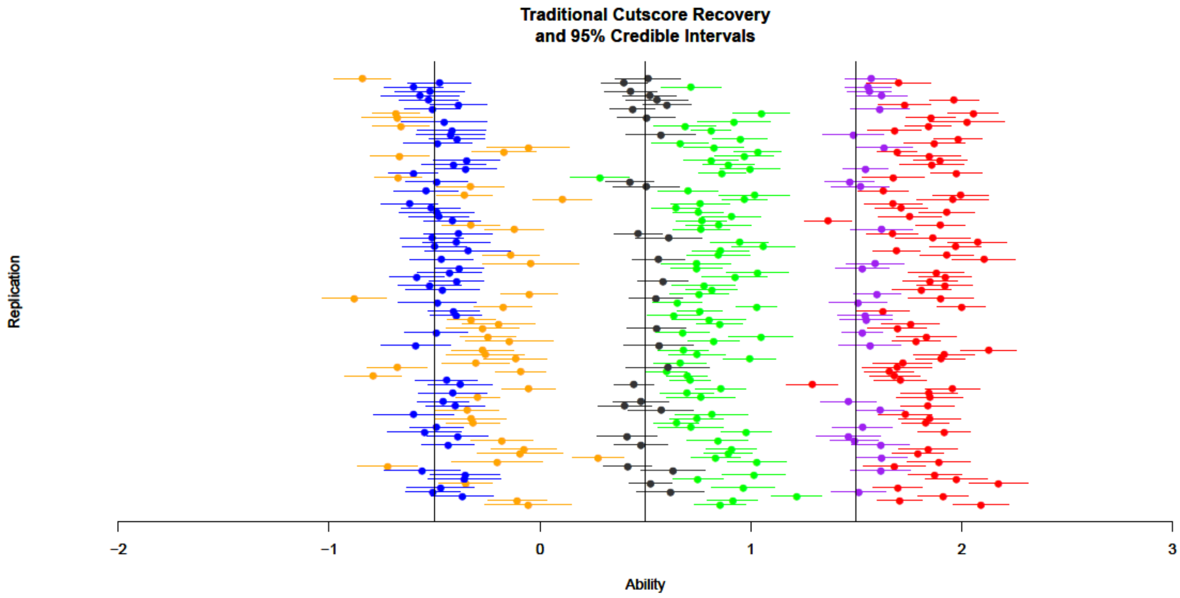

Figure A.193. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

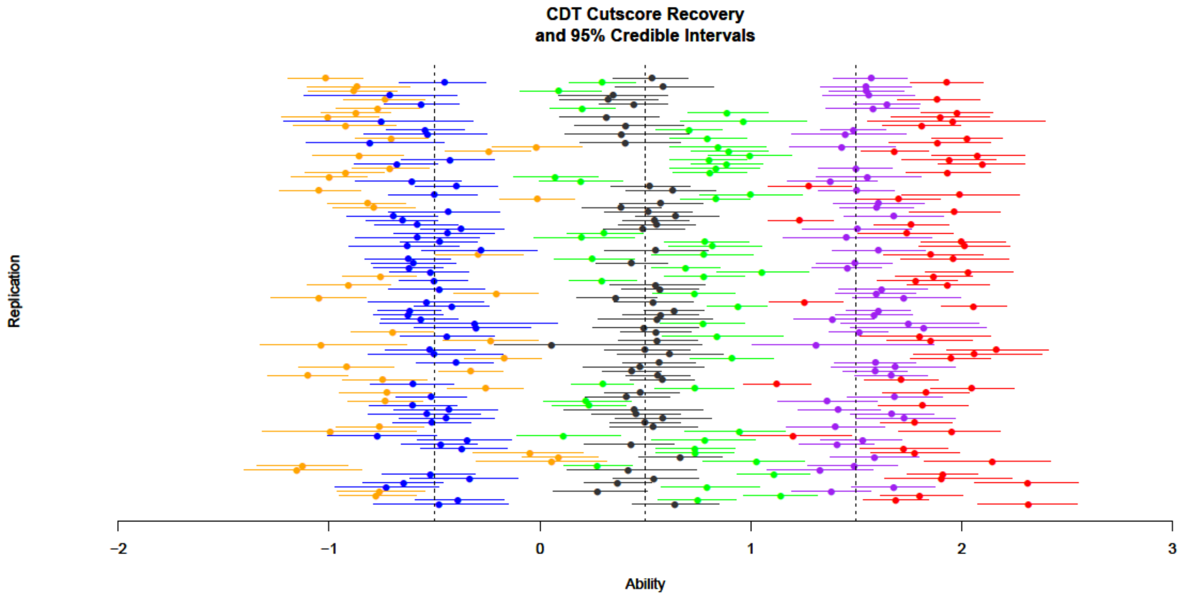

Figure A.194. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

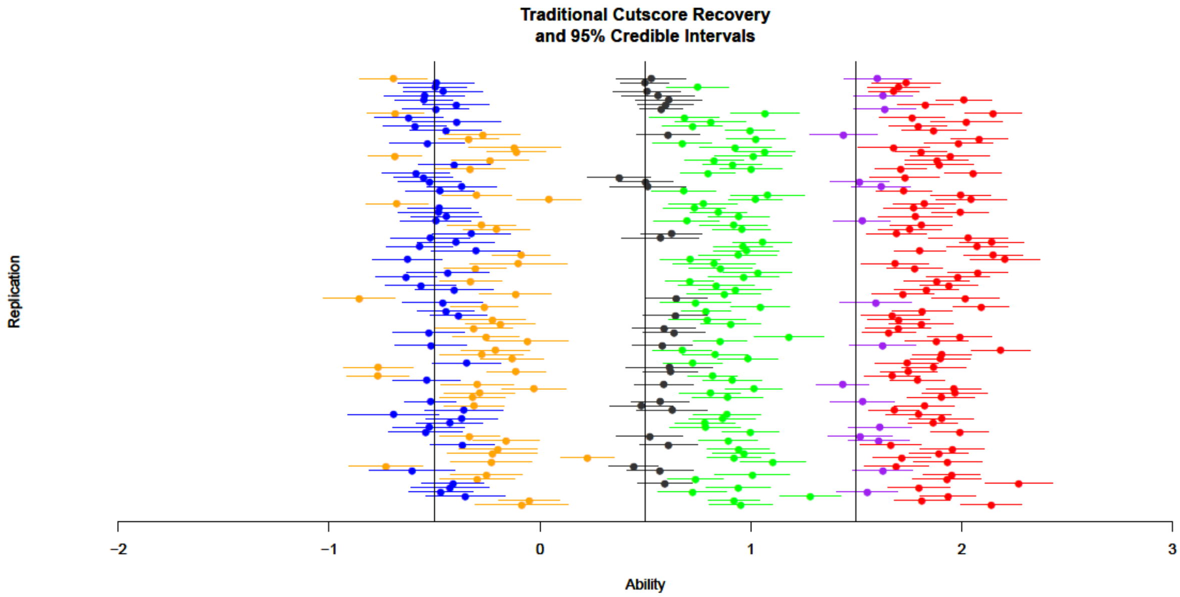

Figure A.195. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

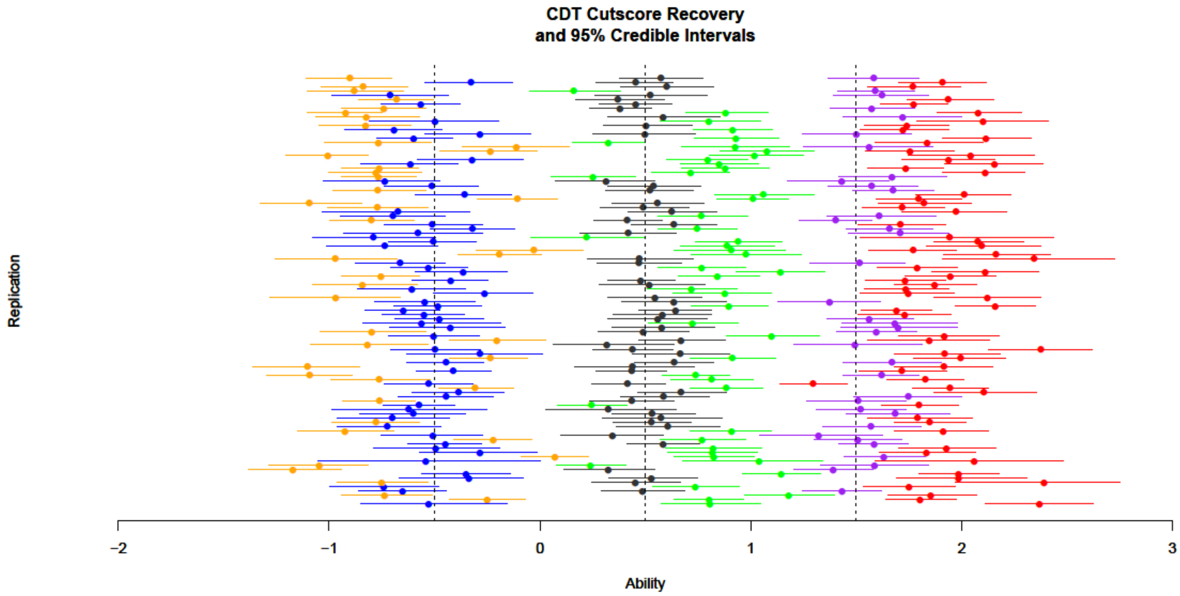

Figure A.196. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.2$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

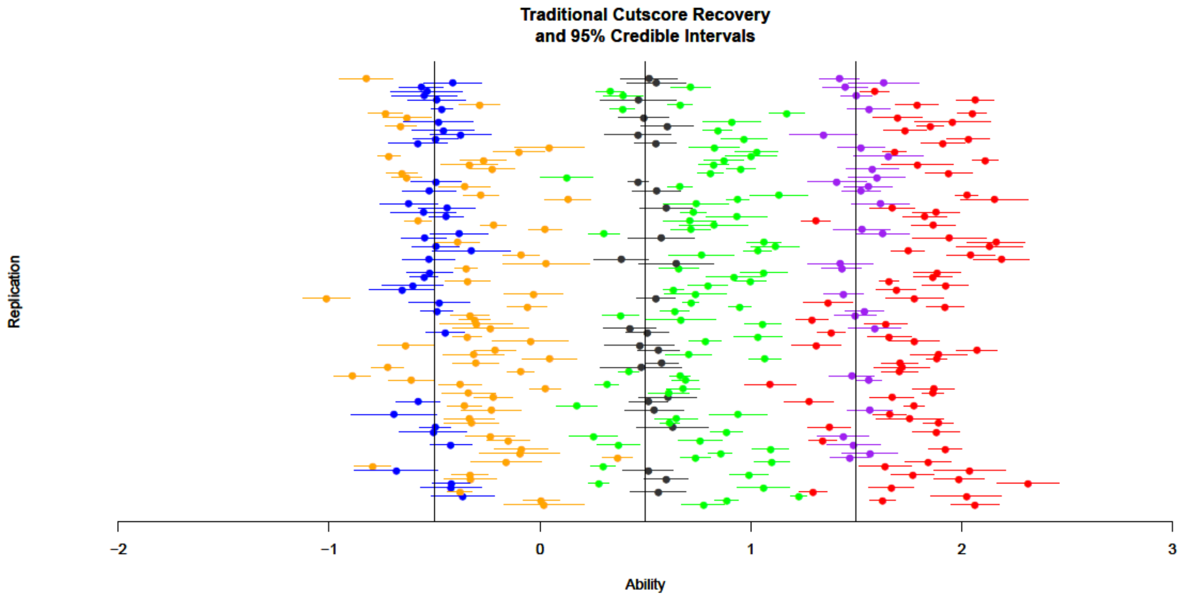

Figure A.197. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.0$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

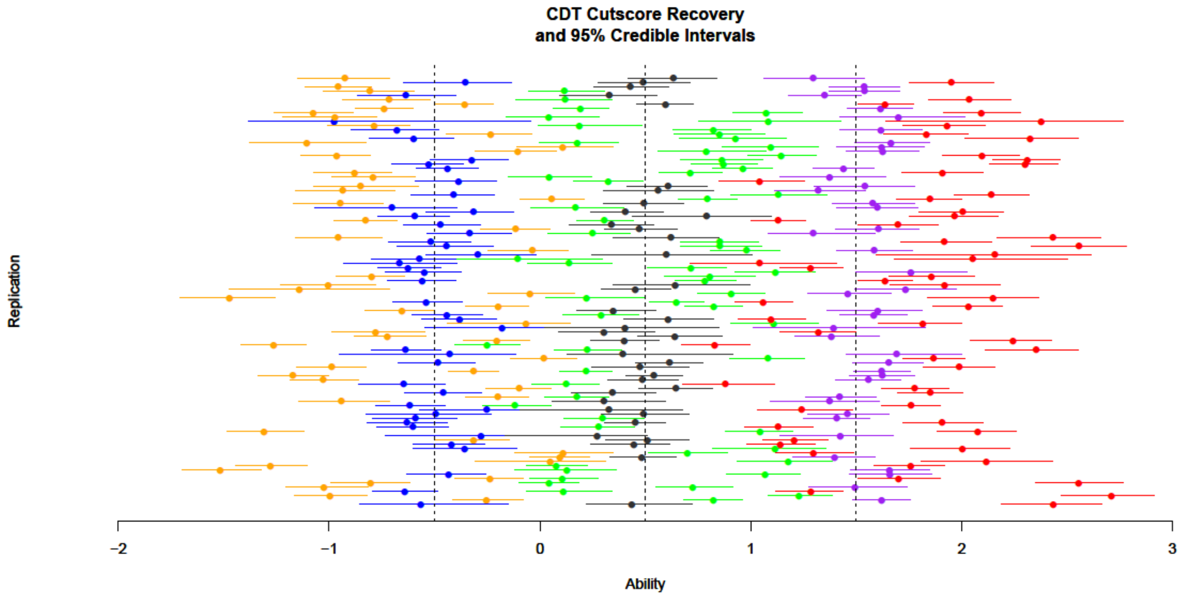

Figure A.198. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.0$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

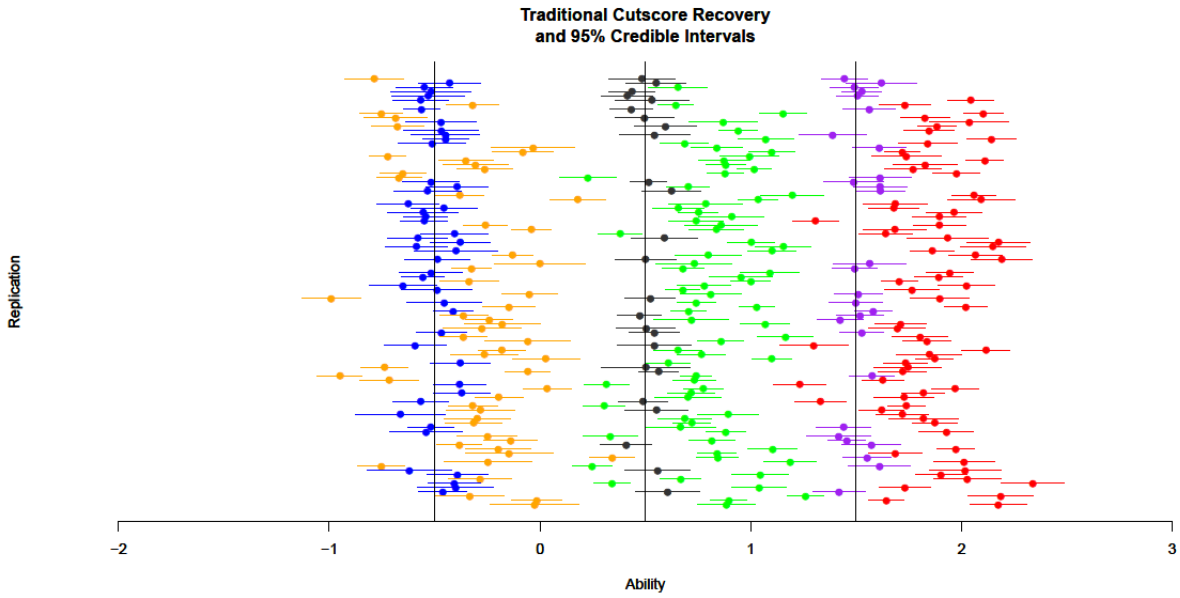

Figure A.199. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.1$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

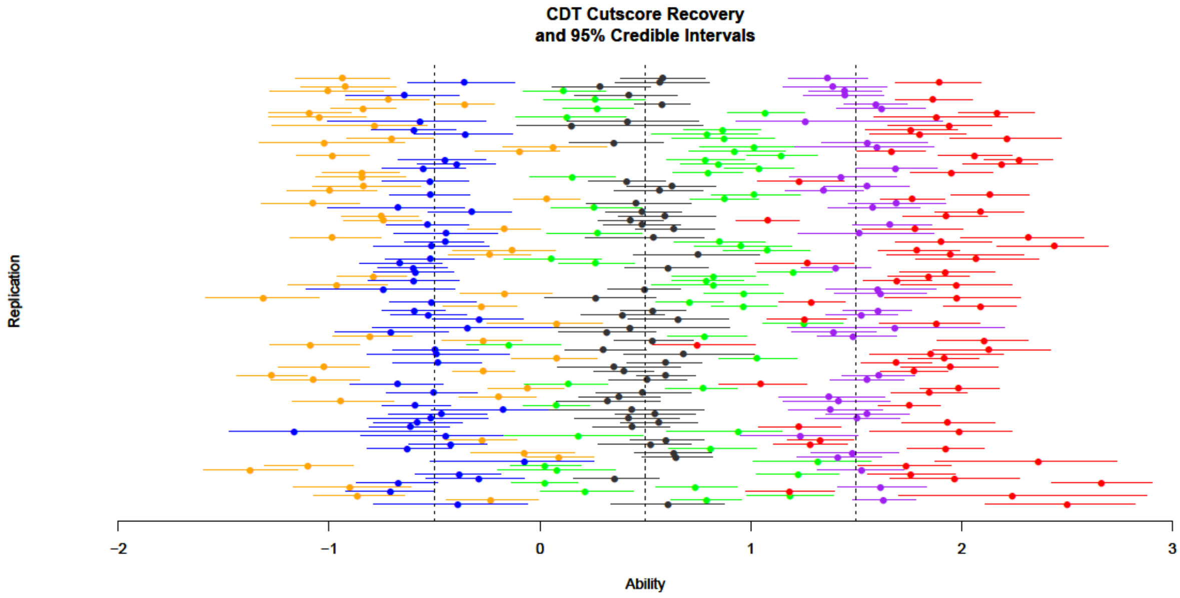

Figure A.200. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.1$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

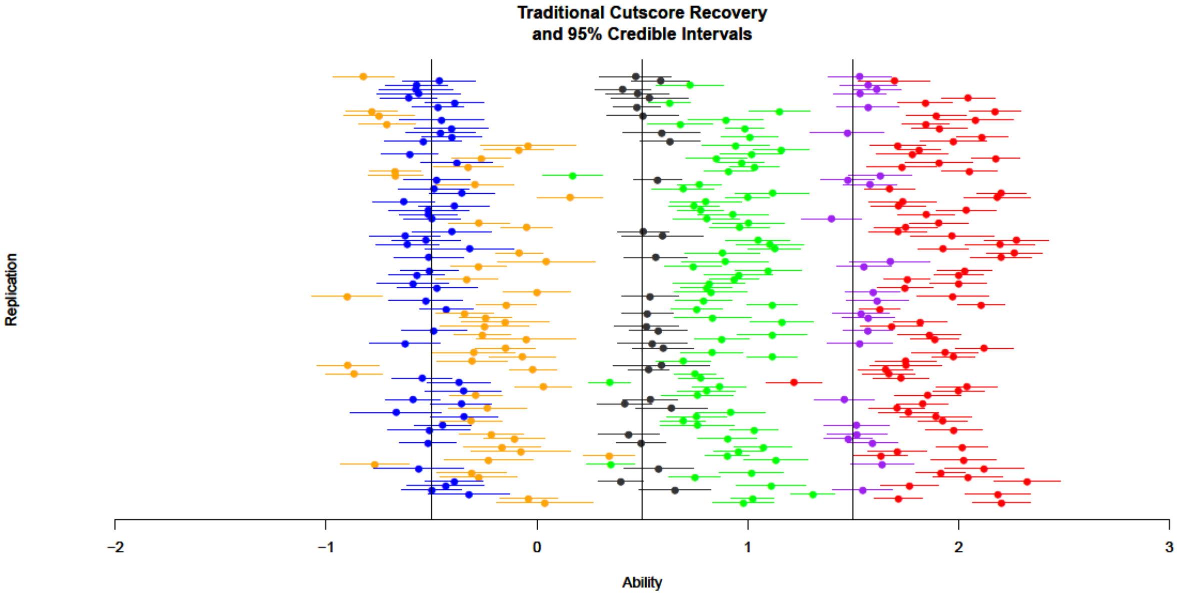

Figure A.201. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.2$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

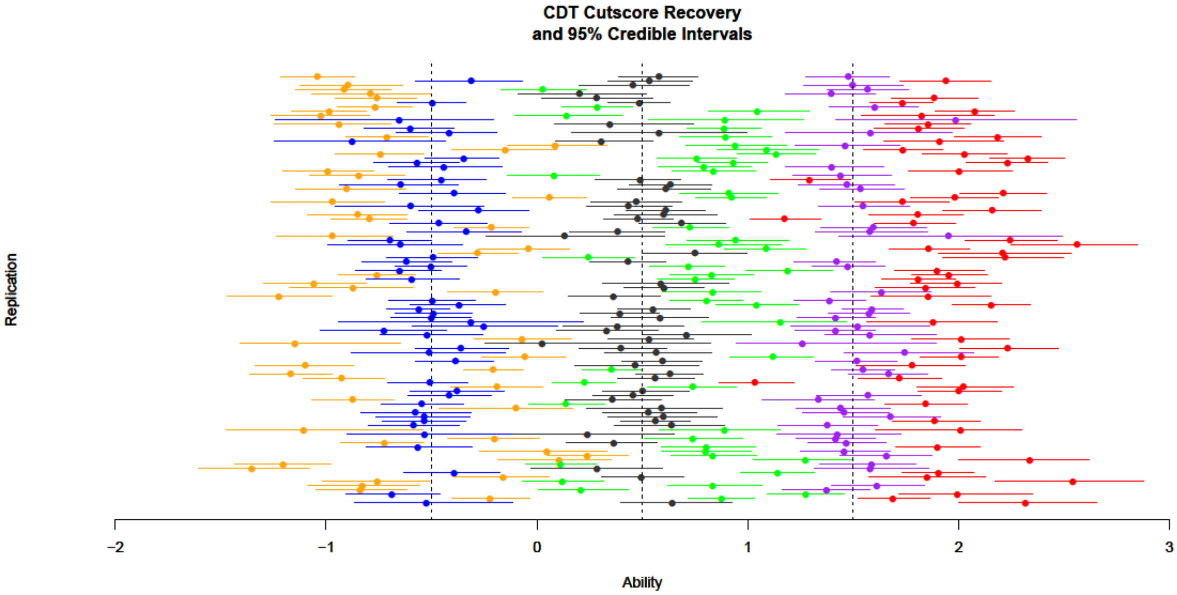

Figure A.202. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.2$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

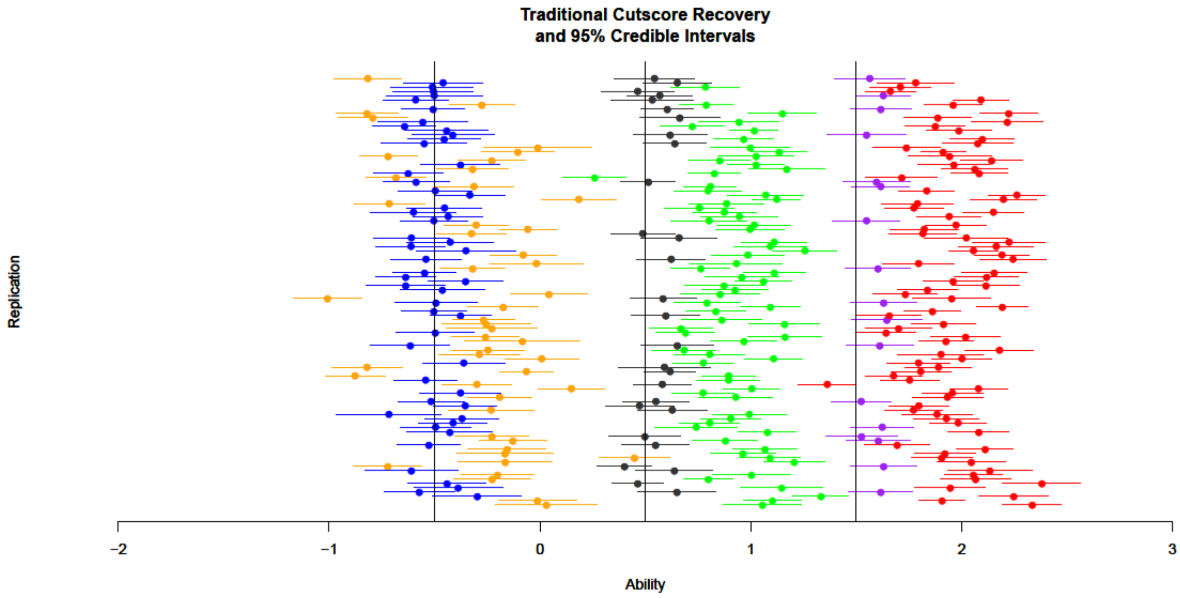

Figure A.203. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; Traditional cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

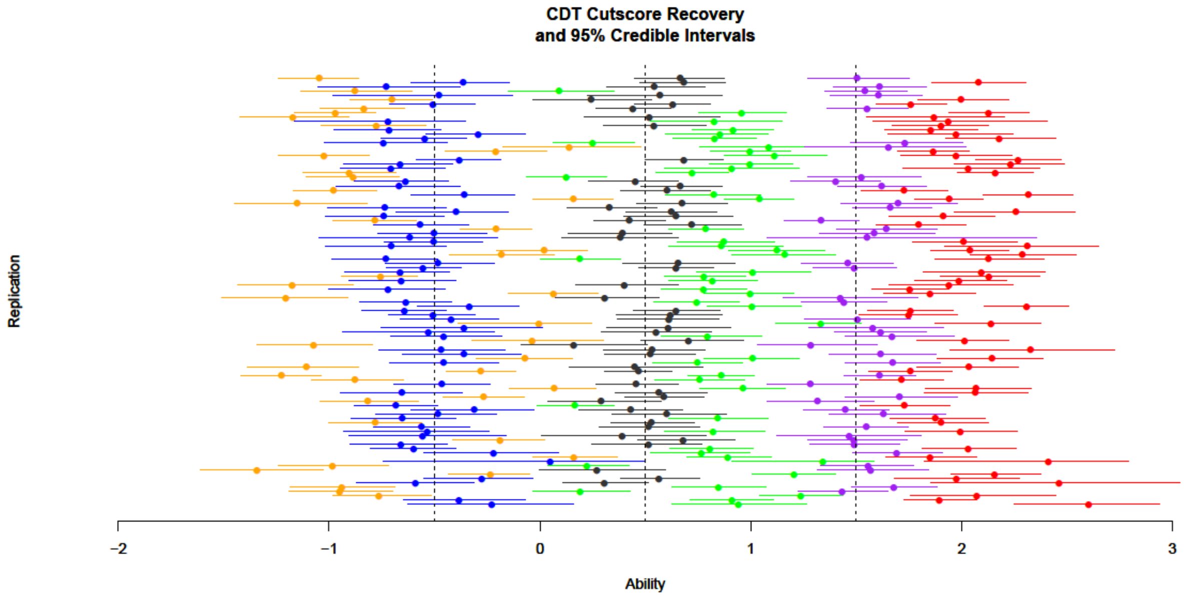

Figure A.204. $N_I = 16$, $N_G = 4$, $\sigma^2_B = 0.3$, $\sigma^2_C = 0.3$; CDT cutscore estimates and confidence intervals. Dotted lines indicate true cutscores. Blue, black, and purple represent CIs that include the true cutscore; orange, green and red indicated CIs that do not include the true cutscore.

Table A.4. Results for Cutscore 1, with 1 and 2 groups of 8 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
8	1	N/A	0.0	-0.386 (0.055)	-0.607 (0.071)	0.003 (0.001)	0.040 (0.002)	47%	100%
			0.1	-0.397 (0.125)	-0.630 (0.139)	0.018 (0.011)	0.056 (0.012)	86%	99%
			0.2	-0.387 (0.160)	-0.623 (0.180)	0.033 (0.019)	0.073 (0.024)	82%	100%
			0.3	-0.402 (0.184)	-0.648 (0.215)	0.046 (0.022)	0.094 (0.038)	89%	98%
		0.0	0.0	-0.394 (0.043)	-0.620 (0.053)	0.002 (0.001)	0.017 (0.001)	33%	99%
			0.1	-0.402 (0.095)	-0.638 (0.111)	0.009 (0.003)	0.023 (0.004)	82%	96%
			0.2	-0.401 (0.126)	-0.638 (0.144)	0.016 (0.005)	0.030 (0.006)	86%	96%
			0.3	-0.398 (0.154)	-0.632 (0.169)	0.022 (0.007)	0.036 (0.010)	85%	91%
		0.1	0.0	-0.401 (0.223)	-0.629 (0.260)	0.005 (0.004)	0.035 (0.028)	44%	72%
			0.1	-0.397 (0.238)	-0.625 (0.268)	0.012 (0.005)	0.039 (0.021)	56%	77%
			0.2	-0.398 (0.260)	-0.624 (0.299)	0.019 (0.007)	0.047 (0.026)	59%	79%
			0.3	-0.388 (0.261)	-0.611 (0.304)	0.025 (0.009)	0.052 (0.027)	68%	79%
		0.2	0.0	-0.395 (0.320)	-0.621 (0.358)	0.008 (0.007)	0.060 (0.065)	38%	69%
			0.1	-0.391 (0.322)	-0.601 (0.351)	0.016 (0.009)	0.063 (0.056)	49%	73%
			0.2	-0.401 (0.338)	-0.610 (0.402)	0.022 (0.010)	0.069 (0.061)	51%	73%
			0.3	-0.409 (0.342)	-0.618 (0.392)	0.028 (0.011)	0.071 (0.058)	65%	79%
		0.3	0.0	-0.399 (0.388)	-0.626 (0.430)	0.011 (0.010)	0.089 (0.091)	35%	68%
			0.1	-0.402 (0.400)	-0.604 (0.442)	0.019 (0.012)	0.092 (0.099)	43%	67%
			0.2	-0.404 (0.406)	-0.603 (0.451)	0.025 (0.013)	0.087 (0.094)	52%	69%
			0.3	-0.415 (0.408)	-0.605 (0.478)	0.032 (0.014)	0.097 (0.092)	57%	75%

Table A.5. Results for Cutscore 1, with 3 groups of 8 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
8	3	0.0	0.0	-0.398 (0.035)	-0.623 (0.043)	0.001 (<0.001)	0.010 (<0.001)	18%	95%
			0.1	-0.402 (0.078)	-0.634 (0.090)	0.006 (0.002)	0.014 (0.002)	74%	89%
			0.2	-0.408 (0.102)	-0.642 (0.118)	0.011 (0.003)	0.018 (0.003)	84%	82%
			0.3	-0.392 (0.121)	-0.614 (0.138)	0.015 (0.004)	0.022 (0.005)	88%	89%
		0.1	0.0	-0.391 (0.177)	-0.618 (0.214)	0.004 (0.003)	0.020 (0.008)	42%	70%
			0.1	-0.393 (0.184)	-0.624 (0.219)	0.009 (0.004)	0.023 (0.008)	56%	73%
			0.2	-0.392 (0.202)	-0.609 (0.231)	0.013 (0.004)	0.028 (0.011)	64%	73%
			0.3	-0.384 (0.210)	-0.595 (0.251)	0.017 (0.005)	0.032 (0.011)	70%	75%
		0.2	0.0	-0.387 (0.251)	-0.616 (0.316)	0.007 (0.005)	0.028 (0.015)	46%	65%
			0.1	-0.386 (0.258)	-0.601 (0.312)	0.012 (0.006)	0.032 (0.017)	53%	71%
			0.2	-0.393 (0.268)	-0.602 (0.324)	0.016 (0.007)	0.036 (0.018)	57%	68%
			0.3	-0.391 (0.281)	-0.585 (0.347)	0.020 (0.007)	0.038 (0.014)	61%	67%
		0.3	0.0	-0.389 (0.303)	-0.616 (0.388)	0.010 (0.007)	0.034 (0.022)	44%	63%
			0.1	-0.398 (0.311)	-0.599 (0.377)	0.015 (0.008)	0.040 (0.026)	53%	65%
			0.2	-0.392 (0.313)	-0.590 (0.375)	0.019 (0.008)	0.043 (0.026)	60%	64%
			0.3	-0.394 (0.339)	-0.584 (0.406)	0.024 (0.009)	0.048 (0.027)	60%	66%

Table A.6. Results for Cutscore 1, with 4 groups of 8 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ_{bias}^2	$\sigma_{consistency}^2$	Trad.	CDT	Trad.	CDT	Trad.	CDT
8	4	0.0	0.0	-0.398 (0.031)	-0.633 (0.038)	0.001 (<0.001)	0.007 (<0.001)	9%	77%
			0.1	-0.403 (0.071)	-0.645 (0.083)	0.004 (0.001)	0.010 (0.001)	69%	70%
			0.2	-0.412 (0.083)	-0.657 (0.096)	0.008 (0.002)	0.013 (0.002)	82%	78%
			0.3	-0.390 (0.105)	-0.631 (0.125)	0.011 (0.003)	0.016 (0.003)	76%	83%
		0.1	0.0	-0.391 (0.143)	-0.629 (0.180)	0.003 (0.002)	0.014 (0.004)	48%	68%
			0.1	-0.396 (0.148)	-0.632 (0.183)	0.007 (0.002)	0.017 (0.005)	69%	70%
			0.2	-0.391 (0.160)	-0.610 (0.195)	0.010 (0.003)	0.019 (0.005)	71%	75%
			0.3	-0.388 (0.167)	-0.605 (0.206)	0.013 (0.003)	0.023 (0.007)	70%	77%
		0.2	0.0	-0.387 (0.201)	-0.619 (0.261)	0.006 (0.003)	0.022 (0.013)	49%	61%
			0.1	-0.381 (0.203)	-0.600 (0.252)	0.009 (0.003)	0.024 (0.013)	61%	68%
			0.2	-0.388 (0.213)	-0.580 (0.263)	0.013 (0.004)	0.026 (0.010)	64%	70%
			0.3	-0.385 (0.215)	-0.578 (0.286)	0.016 (0.004)	0.030 (0.018)	68%	66%
		0.3	0.0	-0.389 (0.243)	-0.618 (0.334)	0.008 (0.005)	0.027 (0.020)	51%	59%
			0.1	-0.393 (0.250)	-0.599 (0.321)	0.011 (0.005)	0.030 (0.019)	59%	63%
			0.2	-0.388 (0.254)	-0.577 (0.318)	0.014 (0.005)	0.030 (0.014)	60%	66%
			0.3	-0.398 (0.269)	-0.597 (0.314)	0.018 (0.006)	0.036 (0.022)	68%	71%

Table A.7. Results for Cutscore 1, with 1 and 2 groups of 16 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ_{bias}^2	$\sigma_{consistency}^2$	Trad.	CDT	Trad.	CDT	Trad.	CDT
16	1	N/A	0.0	-0.394 (0.043)	-0.621 (0.053)	0.002 (0.001)	0.012 (<0.001)	33%	97%
			0.1	-0.402 (0.095)	-0.637 (0.108)	0.009 (0.003)	0.019 (0.003)	82%	89%
			0.2	-0.401 (0.126)	-0.640 (0.142)	0.016 (0.005)	0.026 (0.006)	86%	92%
			0.3	-0.398 (0.154)	-0.642 (0.173)	0.022 (0.007)	0.034 (0.010)	85%	88%
	2	0.0	0.0	-0.398 (0.031)	-0.625 (0.037)	0.001 (<0.001)	0.010 (0.035)	9%	69%
			0.1	-0.403 (0.071)	-0.638 (0.081)	0.004 (0.001)	0.008 (0.001)	69%	69%
			0.2	-0.412 (0.083)	-0.649 (0.095)	0.008 (0.002)	0.012 (0.002)	82%	78%
			0.3	-0.390 (0.105)	-0.627 (0.125)	0.011 (0.003)	0.015 (0.003)	76%	83%
		0.1	0.0	-0.399 (0.221)	-0.611 (0.255)	0.002 (0.002)	0.040 (0.158)	32%	56%
			0.1	-0.399 (0.227)	-0.614 (0.260)	0.006 (0.002)	0.020 (0.019)	48%	63%
			0.2	-0.397 (0.234)	-0.620 (0.266)	0.009 (0.003)	0.019 (0.011)	53%	63%
			0.3	-0.381 (0.240)	-0.594 (0.283)	0.013 (0.003)	0.024 (0.012)	59%	67%
		2	0.0	-0.402 (0.315)	-0.603 (0.361)	0.004 (0.004)	0.052 (0.097)	27%	52%
			0.1	-0.400 (0.305)	-0.601 (0.357)	0.007 (0.004)	0.032 (0.034)	38%	57%
			0.2	-0.403 (0.309)	-0.611 (0.359)	0.011 (0.004)	0.032 (0.030)	44%	61%
			0.3	-0.395 (0.307)	-0.590 (0.364)	0.013 (0.005)	0.031 (0.022)	52%	64%
		0.3	0.0	-0.403 (0.386)	-0.591 (0.439)	0.005 (0.005)	0.079 (0.174)	28%	49%
			0.1	-0.408 (0.381)	-0.608 (0.456)	0.008 (0.006)	0.057 (0.080)	34%	51%
			0.2	-0.400 (0.376)	-0.604 (0.444)	0.012 (0.005)	0.042 (0.060)	44%	55%
			0.3	-0.409 (0.395)	-0.579 (0.491)	0.015 (0.006)	0.037 (0.027)	45%	58%

Table A.8. Results for Cutscore 1, with 3 groups of 16 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ_{bias}^2	$\sigma_{consistency}^2$	Trad.	CDT	Trad.	CDT	Trad.	CDT
16	3	0.0	0.0	-0.397 (0.025)	-0.626 (0.031)	0.001 (<0.001)	0.004 (0.000)	5%	40%
			0.1	-0.408 (0.058)	-0.644 (0.066)	0.003 (0.001)	0.006 (0.001)	57%	49%
			0.2	-0.410 (0.067)	-0.647 (0.075)	0.005 (0.001)	0.008 (0.001)	79%	62%
			0.3	-0.392 (0.083)	-0.624 (0.098)	0.007 (0.002)	0.010 (0.002)	71%	73%
		0.1	0.0	-0.392 (0.169)	-0.623 (0.220)	0.002 (0.001)	0.009 (0.004)	34%	51%
			0.1	-0.394 (0.183)	-0.628 (0.223)	0.004 (0.001)	0.011 (0.005)	42%	51%
			0.2	-0.394 (0.187)	-0.623 (0.228)	0.006 (0.002)	0.013 (0.005)	53%	61%
			0.3	-0.382 (0.193)	-0.603 (0.239)	0.009 (0.002)	0.016 (0.006)	58%	69%
		0.2	0.0	-0.392 (0.243)	-0.611 (0.324)	0.003 (0.003)	0.014 (0.011)	37%	48%
			0.1	-0.393 (0.247)	-0.602 (0.315)	0.006 (0.003)	0.017 (0.012)	43%	55%
			0.2	-0.395 (0.251)	-0.606 (0.301)	0.008 (0.003)	0.018 (0.011)	49%	62%
			0.3	-0.389 (0.240)	-0.589 (0.292)	0.010 (0.003)	0.021 (0.014)	52%	65%
		0.3	0.0	-0.391 (0.300)	-0.616 (0.405)	0.005 (0.004)	0.020 (0.028)	34%	42%
			0.1	-0.396 (0.303)	-0.619 (0.395)	0.007 (0.004)	0.022 (0.021)	41%	49%
			0.2	-0.389 (0.302)	-0.587 (0.385)	0.009 (0.004)	0.022 (0.015)	46%	52%
			0.3	-0.398 (0.302)	-0.594 (0.383)	0.011 (0.004)	0.026 (0.027)	43%	56%

Table A.9. Results for Cutscore 1, with 4 groups of 16 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ_{bias}^2	$\sigma_{consistency}^2$	Trad.	CDT	Trad.	CDT	Trad.	CDT
16	4	0.0	0.0	-0.395 (0.021)	-0.621 (0.026)	<0.001 (<0.001)	0.003 (<0.001)	1%	18%
			0.1	-0.401 (0.051)	-0.633 (0.059)	0.002 (<0.001)	0.004 (<0.001)	44%	39%
			0.2	-0.408 (0.062)	-0.641 (0.071)	0.004 (0.001)	0.005 (0.001)	69%	54%
			0.3	-0.398 (0.067)	-0.632 (0.082)	0.005 (0.001)	0.007 (0.001)	75%	63%
		0.1	0.0	-0.389 (0.135)	-0.610 (0.187)	0.002 (0.001)	0.007 (0.003)	34%	51%
			0.1	-0.393 (0.147)	-0.617 (0.185)	0.003 (0.001)	0.008 (0.004)	45%	54%
			0.2	-0.394 (0.145)	-0.623 (0.192)	0.005 (0.001)	0.010 (0.003)	53%	57%
			0.3	-0.388 (0.150)	-0.609 (0.189)	0.006 (0.001)	0.012 (0.005)	63%	65%
		0.2	0.0	-0.386 (0.195)	-0.591 (0.286)	0.003 (0.002)	0.010 (0.009)	37%	49%
			0.1	-0.387 (0.201)	-0.604 (0.274)	0.004 (0.002)	0.012 (0.009)	48%	51%
			0.2	-0.393 (0.202)	-0.593 (0.266)	0.006 (0.002)	0.014 (0.009)	56%	56%
			0.3	-0.388 (0.196)	-0.595 (0.253)	0.007 (0.002)	0.015 (0.010)	52%	56%
		0.3	0.0	-0.383 (0.241)	-0.586 (0.365)	0.004 (0.002)	0.013 (0.015)	38%	44%
			0.1	-0.393 (0.235)	-0.586 (0.336)	0.005 (0.002)	0.015 (0.010)	45%	49%
			0.2	-0.387 (0.238)	-0.578 (0.332)	0.007 (0.003)	0.017 (0.016)	51%	48%
			0.3	-0.390 (0.249)	-0.574 (0.334)	0.008 (0.003)	0.018 (0.012)	52%	57%

Table A.10. Results for Cutscore 2, with 1 and 2 groups of 8 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
8	1	N/A	0.0	0.582 (0.056)	0.396 (0.073)	0.003 (0.002)	0.037 (0.002)	66%	100%
			0.1	0.623 (0.110)	0.460 (0.126)	0.015 (0.009)	0.050 (0.010)	84%	100%
			0.2	0.649 (0.132)	0.498 (0.150)	0.024 (0.014)	0.061 (0.016)	86%	100%
			0.3	0.704 (0.156)	0.565 (0.174)	0.037 (0.020)	0.078 (0.025)	82%	98%
		0.0	0.0	0.583 (0.043)	0.405 (0.055)	0.002 (0.001)	0.016 (0.001)	44%	100%
			0.1	0.624 (0.076)	0.462 (0.091)	0.007 (0.003)	0.021 (0.003)	65%	100%
			0.2	0.666 (0.115)	0.516 (0.129)	0.012 (0.004)	0.025 (0.005)	67%	98%
			0.3	0.717 (0.136)	0.583 (0.148)	0.017 (0.006)	0.030 (0.007)	63%	94%
		0.1	0.0	0.611 (0.208)	0.449 (0.248)	0.004 (0.003)	0.030 (0.020)	40%	77%
			0.1	0.646 (0.219)	0.494 (0.256)	0.009 (0.004)	0.033 (0.015)	50%	81%
			0.2	0.687 (0.222)	0.553 (0.262)	0.014 (0.005)	0.036 (0.015)	55%	83%
			0.3	0.730 (0.239)	0.599 (0.272)	0.020 (0.008)	0.040 (0.014)	58%	85%
		0.2	0.0	0.648 (0.288)	0.495 (0.337)	0.006 (0.006)	0.050 (0.053)	32%	68%
			0.1	0.688 (0.297)	0.549 (0.345)	0.011 (0.006)	0.047 (0.030)	48%	75%
			0.2	0.732 (0.297)	0.599 (0.349)	0.016 (0.008)	0.047 (0.029)	46%	77%
			0.3	0.761 (0.304)	0.634 (0.337)	0.022 (0.009)	0.052 (0.029)	52%	77%
		0.3	0.0	0.682 (0.349)	0.541 (0.405)	0.009 (0.010)	0.083 (0.107)	33%	68%
			0.1	0.731 (0.357)	0.612 (0.421)	0.014 (0.008)	0.065 (0.054)	38%	68%
			0.2	0.771 (0.352)	0.656 (0.419)	0.018 (0.009)	0.061 (0.051)	41%	69%
			0.3	0.808 (0.357)	0.687 (0.419)	0.025 (0.011)	0.068 (0.045)	47%	72%

Table A.11. Results for Cutscore 2, with 3 groups of 8 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
8	3	0.0	0.0	0.587 (0.034)	0.406 (0.043)	0.001 (<0.001)	0.009 (<0.001)	31%	100%
			0.1	0.622 (0.067)	0.456 (0.078)	0.005 (0.001)	0.012 (0.001)	59%	99%
			0.2	0.660 (0.091)	0.506 (0.103)	0.008 (0.003)	0.016 (0.003)	61%	97%
			0.3	0.707 (0.112)	0.567 (0.122)	0.011 (0.003)	0.018 (0.004)	50%	95%
		0.1	0.0	0.607 (0.165)	0.432 (0.199)	0.003 (0.002)	0.018 (0.006)	38%	67%
			0.1	0.650 (0.176)	0.488 (0.208)	0.007 (0.003)	0.021 (0.006)	45%	81%
			0.2	0.690 (0.172)	0.539 (0.198)	0.010 (0.003)	0.023 (0.008)	48%	89%
			0.3	0.733 (0.189)	0.596 (0.215)	0.014 (0.004)	0.027 (0.009)	47%	86%
		0.2	0.0	0.655 (0.235)	0.488 (0.295)	0.005 (0.004)	0.025 (0.013)	31%	63%
			0.1	0.688 (0.243)	0.532 (0.289)	0.009 (0.005)	0.028 (0.013)	41%	71%
			0.2	0.729 (0.241)	0.580 (0.266)	0.012 (0.006)	0.031 (0.015)	46%	79%
			0.3	0.776 (0.235)	0.627 (0.267)	0.016 (0.006)	0.032 (0.012)	42%	78%
		0.3	0.0	0.698 (0.284)	0.539 (0.360)	0.008 (0.006)	0.032 (0.024)	28%	61%
			0.1	0.742 (0.283)	0.589 (0.342)	0.011 (0.006)	0.032 (0.016)	32%	65%
			0.2	0.774 (0.276)	0.624 (0.319)	0.014 (0.007)	0.035 (0.024)	37%	68%
			0.3	0.824 (0.282)	0.678 (0.327)	0.017 (0.007)	0.038 (0.020)	38%	70%

Table A.12. Results for Cutscore 2, with 4 groups of 8 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
8	4	0.0	0.0	0.589 (0.030)	0.413 (0.037)	0.001 (<0.001)	0.007 (<0.001)	15%	96%
			0.1	0.622 (0.054)	0.460 (0.062)	0.004 (0.001)	0.009 (0.001)	48%	100%
			0.2	0.665 (0.074)	0.513 (0.091)	0.006 (0.002)	0.012 (0.002)	44%	99%
			0.3	0.708 (0.095)	0.566 (0.113)	0.009 (0.002)	0.014 (0.002)	37%	93%
		0.1	0.0	0.621 (0.145)	0.454 (0.186)	0.003 (0.002)	0.014 (0.004)	36%	72%
			0.1	0.650 (0.147)	0.494 (0.170)	0.005 (0.002)	0.016 (0.006)	48%	83%
			0.2	0.694 (0.149)	0.561 (0.187)	0.008 (0.003)	0.017 (0.004)	46%	84%
			0.3	0.731 (0.154)	0.602 (0.179)	0.010 (0.003)	0.019 (0.005)	42%	85%
		0.2	0.0	0.658 (0.211)	0.496 (0.264)	0.004 (0.003)	0.019 (0.008)	33%	65%
			0.1	0.690 (0.203)	0.530 (0.245)	0.007 (0.003)	0.021 (0.009)	42%	70%
			0.2	0.745 (0.205)	0.594 (0.234)	0.009 (0.003)	0.021 (0.007)	39%	70%
			0.3	0.789 (0.194)	0.648 (0.232)	0.012 (0.004)	0.024 (0.010)	41%	73%
		0.3	0.0	0.704 (0.251)	0.543 (0.326)	0.006 (0.004)	0.025 (0.015)	31%	60%
			0.1	0.747 (0.251)	0.590 (0.319)	0.008 (0.004)	0.025 (0.012)	35%	64%
			0.2	0.786 (0.244)	0.638 (0.297)	0.011 (0.004)	0.025 (0.010)	36%	63%
			0.3	0.842 (0.240)	0.697 (0.303)	0.013 (0.005)	0.027 (0.012)	31%	64%

Table A.13. Results for Cutscore 2, with 1 and 2 groups of 16 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
16	1	N/A	0.0	0.583 (0.043)	0.403 (0.055)	0.002 (0.001)	0.012 (0.001)	44%	98%
			0.1	0.624 (0.076)	0.464 (0.090)	0.007 (0.003)	0.017 (0.003)	65%	100%
			0.2	0.666 (0.115)	0.521 (0.134)	0.012 (0.004)	0.023 (0.005)	67%	99%
			0.3	0.717 (0.136)	0.586 (0.153)	0.017 (0.006)	0.030 (0.007)	63%	91%
	2	0.0	0.0	0.589 (0.030)	0.414 (0.038)	0.001 (<0.001)	0.008 (0.023)	15%	94%
			0.1	0.622 (0.054)	0.463 (0.062)	0.004 (0.001)	0.008 (0.001)	48%	99%
			0.2	0.665 (0.074)	0.518 (0.089)	0.006 (0.002)	0.010 (0.002)	44%	98%
			0.3	0.708 (0.095)	0.570 (0.111)	0.009 (0.002)	0.013 (0.002)	37%	92%
		0.1	0.0	0.618 (0.201)	0.460 (0.247)	0.002 (0.002)	0.039 (0.207)	30%	53%
			0.1	0.659 (0.198)	0.524 (0.247)	0.005 (0.002)	0.015 (0.009)	45%	63%
			0.2	0.686 (0.203)	0.557 (0.254)	0.007 (0.002)	0.015 (0.007)	44%	63%
			0.3	0.747 (0.203)	0.626 (0.254)	0.010 (0.003)	0.018 (0.007)	41%	64%
		2	0.0	0.650 (0.286)	0.499 (0.348)	0.003 (0.003)	0.044 (0.084)	25%	48%
			0.1	0.688 (0.285)	0.561 (0.348)	0.006 (0.003)	0.027 (0.037)	34%	57%
			0.2	0.725 (0.280)	0.586 (0.358)	0.008 (0.003)	0.024 (0.023)	42%	55%
			0.3	0.781 (0.278)	0.657 (0.344)	0.011 (0.004)	0.025 (0.020)	35%	57%
		0.3	0.0	0.686 (0.347)	0.541 (0.418)	0.004 (0.004)	0.076 (0.212)	25%	51%
			0.1	0.726 (0.347)	0.588 (0.433)	0.007 (0.004)	0.039 (0.056)	36%	50%
			0.2	0.771 (0.341)	0.623 (0.434)	0.009 (0.005)	0.035 (0.055)	36%	50%
			0.3	0.818 (0.335)	0.702 (0.411)	0.012 (0.005)	0.032 (0.035)	37%	45%

Table A.14. Results for Cutscore 2, with 3 groups of 16 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
16	3	0.0	0.0	0.590 (0.026)	0.414 (0.033)	0.001 (<0.001)	0.003 (<0.001)	9%	82%
			0.1	0.622 (0.045)	0.461 (0.054)	0.002 (<0.001)	0.005 (<0.001)	24%	96%
			0.2	0.661 (0.059)	0.510 (0.071)	0.004 (0.001)	0.006 (0.001)	24%	98%
			0.3	0.701 (0.079)	0.559 (0.097)	0.006 (0.001)	0.008 (0.001)	22%	89%
		0.1	0.0	0.607 (0.172)	0.434 (0.222)	0.002 (0.001)	0.008 (0.003)	18%	51%
			0.1	0.640 (0.165)	0.492 (0.207)	0.003 (0.001)	0.009 (0.004)	34%	57%
			0.2	0.675 (0.162)	0.524 (0.209)	0.005 (0.001)	0.011 (0.004)	38%	62%
			0.3	0.733 (0.159)	0.588 (0.204)	0.007 (0.002)	0.012 (0.004)	31%	60%
		0.2	0.0	0.646 (0.241)	0.478 (0.312)	0.003 (0.002)	0.011 (0.007)	22%	49%
			0.1	0.686 (0.233)	0.530 (0.288)	0.004 (0.002)	0.013 (0.010)	28%	54%
			0.2	0.725 (0.230)	0.566 (0.284)	0.006 (0.002)	0.015 (0.009)	30%	58%
			0.3	0.787 (0.227)	0.649 (0.289)	0.008 (0.002)	0.015 (0.006)	29%	55%
		0.3	0.0	0.689 (0.283)	0.530 (0.376)	0.004 (0.003)	0.015 (0.013)	19%	43%
			0.1	0.732 (0.281)	0.573 (0.344)	0.005 (0.003)	0.017 (0.016)	20%	44%
			0.2	0.771 (0.281)	0.618 (0.344)	0.007 (0.003)	0.017 (0.009)	23%	49%
			0.3	0.823 (0.269)	0.670 (0.316)	0.009 (0.003)	0.018 (0.008)	24%	48%

Table A.15. Results for Cutscore 2, with 4 groups of 16 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
16	4	0.0	0.0	0.591 (0.019)	0.414 (0.025)	<0.001 (<0.001)	0.003 (<0.001)	0%	77%
			0.1	0.619 (0.039)	0.459 (0.047)	0.002 (<0.001)	0.004 (<0.001)	18%	94%
			0.2	0.660 (0.056)	0.510 (0.065)	0.003 (0.001)	0.005 (0.001)	19%	97%
			0.3	0.698 (0.066)	0.556 (0.077)	0.004 (0.001)	0.006 (0.001)	12%	92%
		0.1	0.0	0.622 (0.145)	0.448 (0.191)	0.001 (0.001)	0.007 (0.003)	27%	50%
			0.1	0.653 (0.148)	0.497 (0.184)	0.003 (0.001)	0.007 (0.003)	32%	63%
			0.2	0.690 (0.141)	0.535 (0.177)	0.004 (0.001)	0.009 (0.003)	36%	66%
			0.3	0.752 (0.141)	0.606 (0.184)	0.005 (0.001)	0.010 (0.003)	27%	62%
		0.2	0.0	0.658 (0.207)	0.483 (0.273)	0.002 (0.001)	0.010 (0.006)	25%	43%
			0.1	0.691 (0.195)	0.520 (0.251)	0.003 (0.001)	0.010 (0.005)	28%	52%
			0.2	0.738 (0.193)	0.575 (0.247)	0.005 (0.001)	0.012 (0.006)	29%	52%
			0.3	0.796 (0.188)	0.651 (0.238)	0.006 (0.002)	0.012 (0.004)	29%	52%
		0.3	0.0	0.700 (0.247)	0.538 (0.345)	0.003 (0.002)	0.012 (0.008)	27%	38%
			0.1	0.745 (0.241)	0.582 (0.319)	0.004 (0.002)	0.013 (0.007)	25%	49%
			0.2	0.787 (0.235)	0.611 (0.298)	0.005 (0.002)	0.014 (0.010)	29%	52%
			0.3	0.839 (0.224)	0.672 (0.280)	0.007 (0.002)	0.014 (0.006)	28%	50%

Table A.16. Results for Cutscore 3, with 1 and 2 groups of 8 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
8	1	N/A	0.0	1.595 (0.059)	1.596 (0.054)	0005 (0.002)	0.036 (0.002)	81%	100%
			0.1	1.590 (0.120)	1.593 (0.134)	0.017 (0.009)	0.049 (0.011)	86%	100%
			0.2	1.649 (0.171)	1.661 (0.196)	0.028 (0.014)	0.065 (0.028)	77%	95%
			0.3	1.728 (0.188)	1.763 (0.229)	0.037 (0.020)	0.087 (0.056)	71%	93%
		0.0	0.0	1.601 (0.053)	1.597 (0.048)	0.003 (0.001)	0.016 (0.001)	47%	100%
			0.1	1.611 (0.085)	1.609 (0.095)	0.009 (0.003)	0.021 (0.003)	80%	96%
			0.2	1.658 (0.116)	1.656 (0.135)	0.014 (0.005)	0.027 (0.007)	70%	92%
			0.3	1.738 (0.118)	1.742 (0.144)	0.019 (0.006)	0.033 (0.010)	57%	79%
		0.1	0.0	1.641 (0.220)	1.642 (0.240)	0.005 (0.004)	0.035 (0.035)	40%	79%
			0.1	1.677 (0.216)	1.677 (0.256)	0.011 (0.004)	0.037 (0.021)	46%	79%
			0.2	1.737 (0.211)	1.736 (0.264)	0.016 (0.006)	0.043 (0.023)	51%	75%
			0.3	1.804 (0.209)	1.812 (0.254)	0.021 (0.007)	0.047 (0.020)	43%	64%
		0.2	0.0	1.685 (0.304)	1.693 (0.350)	0.007 (0.006)	0.059 (0.060)	32%	71%
			0.1	1.717 (0.302)	1.729 (0.349)	0.012 (0.006)	0.053 (0.043)	40%	66%
			0.2	1.793 (0.284)	1.789 (0.339)	0.018 (0.007)	0.059 (0.050)	42%	68%
			0.3	1.846 (0.286)	1.855 (0.357)	0.022 (0.007)	0.062 (0.036)	41%	61%
		0.3	0.0	1.743 (0.366)	1.744 (0.409)	0.010 (0.009)	0.094 (0.104)	30%	64%
			0.1	1.782 (0.362)	1.784 (0.429)	0.015 (0.009)	0.080 (0.076)	36%	62%
			0.2	1.840 (0.347)	1.833 (0.406)	0.018 (0.008)	0.073 (0.074)	39%	65%
			0.3	1.900 (0.336)	1.903 (0.434)	0.024 (0.009)	0.093 (0.099)	35%	61%

Table A.17. Results for Cutscore 3, with 3 groups of 8 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
8	3	0.0	0.0	1.601 (0.046)	1.600 (0.041)	0.002 (<0.001)	0.010 (<0.001)	32%	100%
			0.1	1.612 (0.064)	1.607 (0.072)	0.006 (0.002)	0.013 (0.002)	73%	93%
			0.2	1.663 (0.086)	1.661 (0.105)	0.009 (0.003)	0.017 (0.004)	64%	84%
			0.3	1.734 (0.101)	1.748 (0.128)	0.012 (0.003)	0.020 (0.005)	40%	60%
		0.1	0.0	1.623 (0.172)	1.617 (0.201)	0.004 (0.003)	0.019 (0.008)	39%	78%
			0.1	1.668 (0.171)	1.652 (0.214)	0.008 (0.003)	0.022 (0.008)	51%	76%
			0.2	1.719 (0.176)	1.703 (0.220)	0.011 (0.004)	0.028 (0.013)	51%	72%
			0.3	1.797 (0.173)	1.785 (0.222)	0.014 (0.004)	0.031 (0.012)	37%	64%
		0.2	0.0	1.672 (0.245)	1.657 (0.315)	0.006 (0.004)	0.026 (0.013)	33%	65%
			0.1	1.712 (0.248)	1.682 (0.308)	0.010 (0.004)	0.032 (0.031)	42%	66%
			0.2	1.782 (0.243)	1.755 (0.284)	0.013 (0.005)	0.034 (0.016)	40%	64%
			0.3	1.853 (0.235)	1.816 (0.281)	0.016 (0.004)	0.037 (0.015)	34%	58%
		0.3	0.0	1.721 (0.294)	1.700 (0.383)	0.008 (0.006)	0.035 (0.024)	32%	59%
			0.1	1.762 (0.298)	1.722 (0.369)	0.011 (0.006)	0.039 (0.029)	35%	59%
			0.2	1.834 (0.294)	1.799 (0.354)	0.013 (0.005)	0.040 (0.021)	35%	61%
			0.3	1.897 (0.276)	1.844 (0.354)	0.017 (0.006)	0.045 (0.024)	29%	57%

Table A.18. Results for Cutscore 3, with 4 groups of 8 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
8	4	0.0	0.0	1.598 (0.038)	1.595 (0.033)	0.001 (<0.001)	0.007 (<0.001)	27%	100%
			0.1	1.610 (0.059)	1.605 (0.066)	0.004 (0.001)	0.010 (0.001)	59%	96%
			0.2	1.669 (0.072)	1.666 (0.086)	0.007 (0.002)	0.013 (0.002)	45%	74%
			0.3	1.735 (0.087)	1.739 (0.107)	0.009 (0.002)	0.015 (0.003)	30%	51%
		0.1	0.0	1.612 (0.159)	1.606 (0.192)	0.003 (0.002)	0.014 (0.005)	42%	72%
			0.1	1.652 (0.154)	1.633 (0.191)	0.006 (0.002)	0.018 (0.007)	46%	72%
			0.2	1.710 (0.146)	1.692 (0.185)	0.008 (0.002)	0.019 (0.007)	41%	68%
			0.3	1.791 (0.146)	1.781 (0.186)	0.010 (0.003)	0.022 (0.008)	26%	50%
		0.2	0.0	1.655 (0.214)	1.641 (0.287)	0.005 (0.003)	0.022 (0.012)	39%	60%
			0.1	1.705 (0.211)	1.670 (0.259)	0.007 (0.003)	0.024 (0.013)	39%	64%
			0.2	1.773 (0.192)	1.742 (0.229)	0.009 (0.003)	0.024 (0.010)	36%	61%
			0.3	1.833 (0.191)	1.799 (0.250)	0.012 (0.003)	0.028 (0.012)	28%	51%
		0.3	0.0	1.709 (0.256)	1.696 (0.366)	0.007 (0.004)	0.026 (0.015)	35%	54%
			0.1	1.759 (0.246)	1.730 (0.328)	0.009 (0.004)	0.029 (0.016)	38%	56%
			0.2	1.824 (0.240)	1.777 (0.296)	0.011 (0.004)	0.031 (0.016)	29%	63%
			0.3	1.886 (0.238)	1.832 (0.316)	0.013 (0.004)	0.033 (0.015)	27%	54%

Table A.19. Results for Cutscore 3, with 1 and 2 groups of 16 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ_{bias}^2	$\sigma_{consistency}^2$	Trad.	CDT	Trad.	CDT	Trad.	CDT
16	1	N/A	0.0	1.601 (0.053)	1.599 (0.048)	0.003 (0.001)	0.012 (<0.001)	47%	99%
			0.1	1.611 (0.085)	1.612 (0.097)	0.009 (0.003)	0.019 (0.003)	80%	94%
			0.2	1.658 (0.116)	1.667 (0.133)	0.014 (0.005)	0.026 (0.007)	70%	85%
			0.3	1.738 (0.118)	1.768 (0.140)	0.019 (0.006)	0.035 (0.013)	57%	79%
	2	0.0	0.0	1.598 (0.038)	1.596 (0.033)	0.001 (<0.001)	0.006 (0.001)	27%	92%
			0.1	1.610 (0.059)	1.606 (0.065)	0.004 (0.001)	0.008 (0.001)	59%	90%
			0.2	1.669 (0.072)	1.674 (0.086)	0.007 (0.002)	0.011 (0.002)	45%	68%
			0.3	1.735 (0.087)	1.755 (0.107)	0.009 (0.002)	0.014 (0.003)	30%	44%
		0.1	0.0	1.633 (0.215)	1.630 (0.238)	0.002 (0.002)	0.023 (0.040)	24%	58%
			0.1	1.663 (0.207)	1.666 (0.240)	0.005 (0.002)	0.017 (0.011)	36%	58%
			0.2	1.720 (0.202)	1.720 (0.234)	0.008 (0.002)	0.019 (0.010)	37%	56%
			0.3	1.792 (0.188)	1.811 (0.231)	0.010 (0.002)	0.022 (0.010)	27%	46%
		2	0.0	1.676 (0.310)	1.668 (0.367)	0.004 (0.003)	0.044 (0.057)	18%	52%
			0.1	1.717 (0.290)	1.712 (0.325)	0.006 (0.003)	0.033 (0.039)	28%	55%
			0.2	1.770 (0.274)	1.752 (0.312)	0.008 (0.003)	0.027 (0.022)	31%	55%
			0.3	1.841 (0.272)	1.830 (0.336)	0.010 (0.003)	0.031 (0.021)	30%	47%
		0.3	0.0	1.738 (0.370)	1.725 (0.450)	0.005 (0.004)	0.069 (0.113)	19%	46%
			0.1	1.780 (0.354)	1.745 (0.409)	0.007 (0.004)	0.046 (0.060)	23%	54%
			0.2	1.829 (0.342)	1.794 (0.407)	0.009 (0.003)	0.041 (0.058)	24%	47%
			0.3	1.905 (0.326)	1.890 (0.420)	0.011 (0.004)	0.041 (0.042)	27%	44%

Table A.20. Results for Cutscore 3, with 3 groups of 16 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ_{bias}^2	$\sigma_{consistency}^2$	Trad.	CDT	Trad.	CDT	Trad.	CDT
16	3	0.0	0.0	1.601 (0.031)	1.597 (0.028)	0.001 (<0.001)	0.003 (<0.001)	8%	75%
			0.1	1.612 (0.054)	1.605 (0.060)	0.003 (0.001)	0.005 (0.001)	44%	70%
			0.2	1.664 (0.062)	1.661 (0.075)	0.004 (0.001)	0.007 (0.001)	28%	52%
			0.3	1.731 (0.074)	1.744 (0.088)	0.006 (0.001)	0.009 (0.002)	12%	22%
		0.1	0.0	1.623 (0.171)	1.603 (0.212)	0.002 (0.001)	0.008 (0.004)	28%	57%
			0.1	1.656 (0.168)	1.627 (0.206)	0.004 (0.001)	0.010 (0.004)	39%	57%
			0.2	1.717 (0.166)	1.703 (0.199)	0.005 (0.001)	0.012 (0.005)	37%	51%
			0.3	1.783 (0.152)	1.776 (0.183)	0.006 (0.001)	0.013 (0.005)	24%	42%
		0.2	0.0	1.669 (0.240)	1.628 (0.320)	0.003 (0.002)	0.014 (0.013)	25%	50%
			0.1	1.710 (0.238)	1.657 (0.306)	0.005 (0.002)	0.015 (0.010)	28%	51%
			0.2	1.763 (0.214)	1.720 (0.274)	0.006 (0.002)	0.016 (0.009)	31%	50%
			0.3	1.839 (0.216)	1.799 (0.270)	0.007 (0.002)	0.019 (0.009)	25%	47%
		0.3	0.0	1.714 (0.289)	1.668 (0.398)	0.004 (0.003)	0.020 (0.030)	24%	45%
			0.1	1.762 (0.284)	1.700 (0.375)	0.005 (0.003)	0.021 (0.027)	25%	46%
			0.2	1.821 (0.262)	1.763 (0.343)	0.007 (0.002)	0.020 (0.014)	29%	45%
			0.3	1.886 (0.258)	1.809 (0.334)	0.008 (0.002)	0.024 (0.017)	24%	46%

Table A.21. Results for Cutscore 3, with 4 groups of 16 panelists.

				Mean Cutscore (SD)		Mean $\hat{\sigma}_E^2$ (SD)		Accuracy Rate	
N_I	N_G	σ^2_{bias}	$\sigma^2_{consistency}$	Trad.	CDT	Trad.	CDT	Trad.	CDT
16	4	0.0	0.0	1.594 (0.027)	1.592 (0.024)	0.001 (<0.001)	0.003 (<0.001)	4%	72%
			0.1	1.607 (0.043)	1.600 (0.048)	0.002 (<0.001)	0.004 (<0.001)	37%	66%
			0.2	1.662 (0.057)	1.660 (0.071)	0.003 (0.001)	0.005 (0.001)	18%	45%
			0.3	1.727 (0.060)	1.741 (0.077)	0.004 (0.001)	0.007 (0.001)	6%	13%
		0.1	0.0	1.611 (0.148)	1.595 (0.189)	0.002 (0.001)	0.007 (0.005)	32%	53%
			0.1	1.645 (0.151)	1.618 (0.193)	0.003 (0.001)	0.008 (0.004)	41%	54%
			0.2	1.709 (0.132)	1.691 (0.177)	0.004 (0.001)	0.009 (0.003)	29%	51%
			0.3	1.777 (0.133)	1.761 (0.174)	0.005 (0.001)	0.010 (0.004)	13%	36%
		0.2	0.0	1.654 (0.205)	1.624 (0.286)	0.002 (0.001)	0.011 (0.011)	28%	47%
			0.1	1.703 (0.199)	1.652 (0.275)	0.004 (0.001)	0.012 (0.010)	35%	51%
			0.2	1.755 (0.185)	1.706 (0.252)	0.005 (0.001)	0.013 (0.007)	28%	48%
			0.3	1.828 (0.181)	1.795 (0.238)	0.006 (0.001)	0.015 (0.009)	16%	36%
		0.3	0.0	1.708 (0.249)	1.699 (0.392)	0.003 (0.002)	0.013 (0.015)	29%	41%
			0.1	1.763 (0.239)	1.715 (0.345)	0.004 (0.002)	0.015 (0.015)	28%	38%
			0.2	1.820 (0.234)	1.754 (0.309)	0.005 (0.002)	0.016 (0.012)	25%	46%
			0.3	1.898 (0.219)	1.826 (0.295)	0.006 (0.002)	0.018 (0.015)	17%	42%