RATER DRIFT IN CONSTRUCTED RESPONSE SCORING VIA LATENT CLASS SIGNAL DETECTION THEORY AND ITEM RESPONSE THEORY

Yoon Soo Park

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2011

ABSTRACT

RATER DRIFT IN CONSTRUCTED RESPONSE SCORING VIA LATENT CLASS SIGNAL DETECTION THEORY AND ITEM RESPONSE THEORY

Yoon Soo Park

The use of constructed response (CR) items or performance tasks to assess test takers' ability has grown tremendously over the past decade. Examples of CR items in psychological and educational measurement range from essays, works of art, and admissions interviews. However, unlike multiple-choice (MC) items that have predetermined options, CR items require test takers to construct their own answer. As such, they require the judgment of multiple raters that are subject to differences in perception and prior knowledge of the material being evaluated. As with any scoring procedure, the scores assigned by raters must be comparable over time and over different test administrations and forms; in other words, scores must be reliable and valid for all test takers, regardless of when an individual takes the test.

This study examines how longitudinal patterns or changes in rater behavior affect model-based classification accuracy. Rater drift refers to changes in rater behavior across different test administrations. Prior research has found evidence of drift. Rater behavior in CR scoring is examined using two measurement models – latent class signal detection theory (SDT) and item response theory (IRT) models. Rater effects (e.g., leniency and strictness) are partly examined with simulations, where the ability of different models to capture changes in rater behavior is studied. Drift is also examined in two real-world

large scale tests: teacher certification test and high school writing test. These tests use the same set of raters for long periods of time, where each rater's scoring is examined on a monthly basis.

Results from the empirical analysis showed that rater models were effective to detect changes in rater behavior over testing administrations in real-world data. However, there were differences in rater discrimination between the latent class SDT and IRT models. Simulations were used to examine the effect of rater drift on classification accuracy and on differences between the latent class SDT and IRT models. Changes in rater severity had only a minimal effect on classification. Rater discrimination had a greater effect on classification accuracy. This study also found that IRT models detected changes in rater severity and in rater discrimination even when data were generated from the latent class SDT model. However, when data were non-normal, IRT models underestimated rater discrimination, which may lead to incorrect inferences on the precision of raters. These findings provide new and important insights into CR scoring and issues that emerge in practice, including methods to improve rater training.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES

ACKNOWLEDGMENTS

I am greatly indebted to my advisor, Professor Lawrence DeCarlo, for his patience and continuous guidance to help me think and develop the focus of this study. I am always inspired by his enthusiasm for research; he has taught me the importance of pursuing a deeper understanding of statistical models used for studying human behavior. I thank him for his invaluable words of advice.

I am also grateful to many mentors that have helped me to realize my career and motivation to do research. I thank Professor Young-Sun Lee for her continuous support to provide both academic and professional guidance since the beginning of my graduate study. I deeply appreciate all the insightful words she has given me. I also thank other members of my dissertation committee, Professors Herbert Ginsburg, Stephen Peverly, and Ellen Lukens, who have helped me to realize the substantive implications of this study. I also thank my colleagues at the National Center for Disaster Preparedness and the Mailman School of Public Health for their support and for understanding my academic requirements to balance school with work. I thank Drs. Ju-Ho Lee and Song-Chang Hong who have inspired me to enter the field of measurement.

I dedicate this study to my family. I am truly grateful to my parents, Sang-Hyun Park and Hye-Young Jin, for believing in me and for providing everything I can imagine. I also thank my sister, Yoon Ji Park, and my brother-in-law, Daniel Joo, for their continuous encouragement. I thank my wife, Heeyoung Choi, who provided endless support and love; your smiles keep me motivated everyday.

Chapter I

INTRODUCTION

The use of constructed response (CR) items or performance tasks to assess test takers' ability has grown tremendously over the past decade. Examples of CR items in educational measurement range from essays, works of art, and musical performances. In particular, there is a growing prevalence of essays used in high-stakes decisions such as admissions tests; examinations such as the Medical College Admission Test (MCAT) and the Graduate Record Examination (GRE) demonstrate the popularity of CR items in the educational field. In addition, the use of CR items has also extended to certification programs. Examples include essays that determine eligibility for high school diploma, teaching certification, and medical practice (National Education Goals Panel, 1996; National Board for Professional Teaching Standards, 1993; Margolis & Ross, 1995).

The increased use of CR items can be attributed to its role in validity. According to Livingston (2009), there are important skills and knowledge, such as complex competencies, direct performances, or explication of reasoning that cannot be fully measured when only multiple choice (MC) items are used. CR items also measure the abilities of low- and high-performing students more accurately (Ercikan et al., 1998) and avoid testwiseness that can occur when only MC items are administered (Pollock, Rock, & Jenkins, 1992; Rodriguez, 2002). Therefore, when used selectively and scored with rigor, CR items provide valid information and insight into students' achievements.

However, unlike multiple-choice (MC) items that have predetermined options, CR items require test takers to construct their own answer. Although clear guidelines exist to

score MC items such as fixed timing, machine-scored answer sheets, equating different forms, and reporting scores on a continuous scale, items that require test takers to write essays, create pieces of art, dance, or record spoken language do not necessarily have a clear and objective answer (McClellan, 2010). As such, they require the judgment of multiple raters that are subject to differences in perception and prior knowledge of the material being evaluated. As with any scoring procedure, scores assigned by raters must be comparable over time and over different test administrations and forms; in other words, scores must be reliable and valid for all test takers, regardless of when an individual takes the test.

Rater drift refers to changes in rater behavior across different test administrations. Prior research has found evidence of rater drift (e.g., Wilson & Case, 2000; Congdon & McQueen, 2000). Although raters can drift within a testing occasion, this study considers rater drift between test administrations. More specifically, this study investigates patterns of rater drift from two or more raters scoring the same CR and examines how changes in their rating behavior can affect scores in the context of various rater models. Rater drift will be examined using simulations and analysis of real-world data.

1.1 Statement of the Problem

Inherent within the framework of CR scoring is the notion that objective scores are independent of the rater (Wright & Douglas, 1986), meaning that regardless of the person grading the performance, the same score will be given. However, contrary to this assumption, it has been noted in the literature that there are individual differences in perception and judgment (Thurstone, 1927), which embodies a subjective nature into

scoring CR items. This can lead to dire consequences for measuring an examinee's ability when differences in raters are ignored. For example, if the reliability of a rater is low, then there is a high likelihood that the same decision made by another rater will result in a different score. The volatility of decisions that vary across raters can become a problem, even a liability with legal consequences (Johnson, Penny, & Gordon, 2000). As such, 78.4% of state departments of educations that use CR items in their testing program employ two or more raters to help resolve reliability issues that may result from using the score from only one rater (Johnson, Penny, & Johnson, 1998). This section considers problems associated with CR scoring within the context of rater drift.

Rater Drift

Rater drift occurs when raters unintentionally redefine their scoring criteria or standards over time (Wheeler, Hartel, & Scriven, 1992, p. 12). A problem associated with rater drift is that it can lead to problems with scoring accuracy. For example, a rater can be strict or lenient depending on the testing occasion; that is, given two testing administrations, a rater may score stricter on the second test, giving an advantage to examinees that tested earlier. These changes in raters' scoring behavior can be attributed to a wide variety of errors or *rater effects* (Myford and Wolfe, 2003). Rudner (1992) classifies rater effects as (1) the halo effect, impressions that a rater forms about an essay, (2) stereotyping, impressions that a rater forms about a group of essays, (3) perceptional differences, viewpoints and past experiences of a rater that can affect interpretation of behaviors or context, (4) leniency or stringency error, systematically scoring higher or lower from lacking sufficient knowledge to make an objective rating, and (5) scale shrinking, preference in raters to avoid the end of a scale.

The measurement literature focuses mostly on rater drift due to leniency or stringency error – *rater severity* (e.g., Lumley & McNamara, 1995; Congdon & McQueen, 2000). For instance, in Lunz and Stahl (1990), rater severity was studied over three grading periods using essays and oral examinations. They found that there was significant instability in rater severity among two of the three periods. In a different study by Myford (1991), ratings of dramatic performances were examined over a month. It was again found that there were significant changes in the severity of raters regardless of their expertise. Rater drift has also been studied using a large-scale assessment that was graded by trained raters over seven rating days; the results from this study showed differences in rater severity for each rating day (Congdon & McQueen, 2000). In these studies, rater drift was examined using parameters from the FACETS model (Linacre, 1989) that indicated a level of rater stringency. Drift was measured as a change in the severity parameter over different occasions; it was also measured using fit statistics and residuals derived from the rater model that examined a level of agreement between the raters.

As these studies show, raters have a tendency to drift in their rating, which can be a problem for scores generated from models used in CR scoring. As such, the effect of rater drift on the accuracy of scores derived from different rater models needs further examination. The following section describes models used in CR scoring.

Models Used for CR Scoring

Various models have been developed to score CR items. However, it is unclear how rater drift affects model-based classification of scores as defined in the scoring rubric. This section considers two types of rater models (1) item response theory (IRT) models and (2) latent class signal detection theory (SDT) model.

Item response theory (IRT) models. In IRT, an examinee's response patterns are used as indicators to measure a latent ability (θ). Examples of IRT models for scoring CR items are the graded response (GR) model (Samejima, 1969), partial credit (PC) model (Masters, 1982), and generalized partial credit (GPC) model (Muraki, 1992). All three models have a threshold or step parameter (b_k), which can be used to estimate rater severity and to infer information about rater effects. Both the GR and the GPC models also have a discrimination parameter (a) that measures the ability of raters to discriminate among essays of different quality. The PC model is a simplified version of the GPC model in that the former does not incorporate a discrimination parameter. Furthermore, the GR model and the GPC model differ in how they parameterize differences in scoring categories. Another IRT model commonly used to score CR items is the FACETS model (Linacre, 1989) that measures both rater severity as well as item difficulty; both rater and item effects comprise the "facets" of the model. However, for a single CR item, the FACETS model is equivalent to the PC model.

Latent class signal detection theory (SDT) model. In the latent class signal detection theory (SDT) model, CR scoring is viewed as a psychological process. The SDT approach to CR scoring uses a latent class extension (DeCarlo, 2002, 2005), where raters are viewed as attempting to discriminate between latent classes of essays. Here, latent classes are defined by the scoring rubric, because the rubric provides a description of latent categories that raters attempt to discriminate. For example, if there are 4 scores defined in the scoring rubric, it is assumed that there are 4 latent classes that raters attempt to discriminate. The latent class SDT model provides a measure of a rater's precision in terms of how well they discriminate between the latent classes (d). It also

estimates their use of response criteria (c_k), which reflects rater effects such as how lenient or strict they score as well as shrinkage and other effects.

Using patterns of rater scores and estimated rater parameters, the latent class SDT model classifies essays into latent classes defined by the scoring rubric. A unique aspect of the latent class SDT model is that it allows an examination of the quality of classification. This is measured by *classification accuracy* (see DeCarlo, 2002, 2005), which is used in this study to examine the effect of rater drift on model-based classification from the latent class SDT model.

The study of rater drift also requires the use of incomplete designs which are used in practice. Large-scale assessments such as Praxis and the Test of English as a Foreign Language (TOEFL) use incomplete designs with 2 raters per essay (DeCarlo, 2008). In both cases, when there is a discrepancy of two or more points, a third rater adjudicates differences in the scores (Xi & Mollaun, 2009). The use of only two raters per essay raises issues about rater designs. Both simulations and empirical data analysis can be used to evaluate whether rater drift can be adequately estimated under incomplete designs and how it affects classification accuracy.

1.2 Purpose of the Study

The purpose of this study can be divided into two main goals. This study investigates the effect of rater drift on model-based classification of constructed responses into latent categories defined by the scoring rubric; that is, this study examines the effect of different patterns of rater drift on classification accuracy. Moreover, the ability of different rater models to detect drift is examined, as parameters used to describe

rater severity and discrimination may differ between models. To address these issues, this study is divided into two parts – empirical and simulation studies.

Empirical study. In the empirical study, two real-world data sets are used: a teacher certification test and a high school writng test. The analysis consists of the following:

(1) identify patterns of rater drift and

(2) examine the effects of rater drift on model-based classification.

Patterns of rater drift are summarized using IRT models (GR and GPC models) and the latent class SDT model. Rater drift is examined using plots of parameter estimates reflecting rater severity and rater precision from rater models over several testing occasions. Parameters that represent rater severity (threshold or step parameter in IRT models and the criteria in the latent class SDT model) are investigated for drift. This study also examines drift in rater discrimination. Most studies that examined rater drift (e.g., Congdon & McQueen, 2000) have concentrated on rater effects such as rater severity over time; however, not many have examined changes in rater discrimination over time.

The effect of rater drift on model-based classification is examined using classification accuracy statistics derived from the latent class SDT model, which measures the quality of classification. Measures of classification accuracy can be created for each scoring occasion to examine changes in latent scores due to drift. These measures provide information about the effect of rater drift on the quality of model-based classifications.

Simulation study. Simulation studies are conducted to examine the effect of rater drift on classification accuracy. Simulations allow the researcher to test different conditions by manipulating rater severity and discrimination over time to assess how rater drift can affect classification. The simulation study examines the relationship between rater drift and model-based classification using the latent class SDT model. Rater drift is examined by changing rater behavior across two time points. This study examines the following:

(1) the effect on classification accuracy when some raters become stricter or lenient,

(2) the effect on classification accuracy when raters become more discriminating,

(3) the ability of IRT and the latent class SDT models to detect rater drift, and

(4) the impact of changing latent class sizes on rater parameter estimates.

First, simulations are used to examine the effect of rater drift on classification accuracy using data generated from the latent class SDT model. In this model, one type of rater severity occurs when raters' criteria locations shift. If the criteria all shift up, then raters are stricter, because they tend to give lower scores. If they shift down, then raters are more lenient. The conditions above allow an examination of rater severity on classification accuracy when rater effects are present across two testing administrations. The simulation also investigates changes in classification accuracy when rater discrimination increases between the testing administrations.

Simulations are also used to examine how well IRT models detect drift when data are generated using the latent class SDT model. The effect on IRT parameter estimates are studied when raters are more lenient and strict. Parameters are also examined when

the distribution of scores are non-normal, meaning a concentration of scores in the mid-scoring categories with very few scores in the extreme categories. Shifting the latent class sizes and assessing this effect on classification is also examined. For example, this can occur when there is a greater use of higher scoring categories in the second scoring occasion than in the first scoring occasion.

Summary. The empirical and simulation studies comprise an investigation of how rater drift affects classification accuracy. The empirical analysis investigates patterns of drift in real-world data, and whether rater severity and discrimination affect classification. This is accompanied by examining classification accuracy over the testing administrations. The simulation study investigates the relationship between rater drift and model-based classification. The combination of both studies will inform researchers on the effects of rater drift and its implication for rater models. The results from this analysis will also indicate new and important understanding of CR scoring and issues that emerge in practice.

Chapter II

LITERATURE REVIEW

This chapter reviews studies in educational measurement used to assess rater drift. Efforts to reduce rater effects through the use of feedback and training are examined in the context of rater drift. A description of incomplete designs, which are commonly used in practice to allocate CR to raters, is also included. The remaining sections of the chapter describe models used for rater effects: IRT models and the latent class SDT model.

2.1 Rater Drift

Rater drift refers to changes in rater behavior over different testing administrations. The literature on rater drift documents its occurrence as a change in rater scoring over time. More specifically, studies have focused on drift due to changes in *rater severity*, which refers to the general leniency or harshness of a rater (Linacre, 1989). On the other hand, the term *rater characteristic* is a more holistic term that encompasses both rater severity as well as other rater effects (McNamara & Adams, 1991). The consensus from most studies in the measurement literature is that rater drift persists, and it is difficult to eliminate tendencies in raters to drift. Although many studies have identified rater drift as a problem, not many have examined how it affects model-based classification.

Knowledge that there is variability in test scores due to rater factors dates as early as Edgeworth (1890). In general, there are two main problems with grading CR items: (1) different raters assign different scores to a particular essay and (2) the same rater may

assign different scores to the same CR on different occasions (Coffman, 1971). For example, in a classic study by Diederich, French, and Carlton (1961), where 300 essays were judged by 53 raters, it was found that 94% of the essays received at least 7 different scores from the raters.

This section reviews articles from the literature that have examined rater drift. Then, efforts to reduce rater effects through training and feedback, focused on rater severity are presented. These studies are important, because they attempt to alleviate the problems created by rater drift.

Studies on Rater Drift

Various studies have investigated the effects of rater severity on model-based scores. In Lunz, Wright, and Linacre (1990), a section of the certification examination was used to demonstrate the prevalence of rater severity using the FACETS model (Linacre, 1989). Two hundred and seventeen examinees' clinical assessments of fifteen histology slides were examined by eighteen raters that scored each slide on a 1 to 5 scale. There were 15 slides to examine, with a total possible score of 75 points. However, due to varying rater severity, some judges gave a score lower than others reflecting strictness; others scored higher, showing leniency. The study reported two fit statistics to indicate intra-judge consistency across items and examinee performances. The *infit* statistic is an information weighted mean-square residual difference between the observed and expected that measures the change from the expected value, and the *outfit* statistic is an unweighted mean-square residual, which is useful for identifying outlying deviations (Wright & Masters, 1982). The authors used these statistics to screen judges that were deviant from the rest.

The study found that the slides were graded consistently by the raters, as indicated by the *infit* statistic, but there were also severe or lenient graders, represented by the *outfit* statistic. That is, raters maintained their level of severity across slides and examinations, but the level of severity differed significantly between raters. The authors noted that using unadjusted rater scores without accounting for rater severity can create biased inferences about examinee performance. Moreover, results from the study supported the findings from the literature that differences in rater characteristics can bias examinee performance.

In Congdon and McQueen (2000), the FACETS model was used again to examine the stability of rater severity over time (i.e., rater drift) using the ratings of 16 judges on 8,285 elementary school students over seven rating days. Results showed that there were significant differences in rater severity between raters and also for the same rater during this period by separately fitting the FACETS model for each day. They also examined measures of agreement using the infit and outfit statistics, which demonstrated drift among raters. In other words, the findings suggested calibrating rater severity for each occasion, due to the variability of rater severity between raters and for the same rater at different time points. The authors also concluded that a possible extension of their study using the partial credit model (Masters, 1982) would be meaningful under variant multifaceted considerations.

In a study spanning three months, scores of "stable" raters were studied using a clinical skills assessment task (McKinley & Boulet, 2004). An analysis of covariance (ANCOVA) design was used to study rater severity over time, where the effect of an outcome was controlled using explanatory variables. Two measures of examinee ability

were used as adjustments over different time periods with rater scores as the outcome variable. This method was used to ensure that changes in rater behavior were not a function of examinee ability. The authors concluded that raters who were relatively stable across days or weeks may also drift in more extended periods; the authors also found that even from a sample of stable raters, there were some that drifted significantly. They concluded that drift among certain raters should be regarded as an important effect, because they can provide an unfair advantage to examinees.

Rater drift has also been examined under the generalizability theory (G-theory) framework. In Harik et al. (2009), the effectiveness of using estimated rater parameters to adjust for differences in rater severity was studied. They used a clinical skills examination data to assess whether the G-theory approach could eliminate rater-related error by statistical adjustment. The authors adjusted for sources of rater and item variability, which was found to improve the precision of the scores. Furthermore, they noted that adjusting for rater severity produced appropriate estimates within similar periods or between 1 to 2 months. However, the use of predetermined rater parameters to adjust for rater severity in as little as 5 to 6 months was ineffective and even counterproductive.

Using a simulation, Wolfe, Moulder, and Myford (2001) examined the recovery of parameters exhibiting rater drift. They generated data using the FACETS model (e.g., by setting examinee ability to be normally distributed with fixed mean and variance) that an examinee receives a particular score from a rater; they also specified population values of parameters to exhibit drift. For example, by shifting rater severity parameter over testing occasions, they generated a condition where raters were stricter. The infit and

outfit statistics were used to assess rater drift; these statistics were derived from the FACETS model as well as the recovery of parameters. By investigating rater severity over time, the authors recovered parameters for the same rater as well as variability across raters. Although this was one of few studies that conducted a simulation to examine the effects of drift on a rater model, the authors received criticism that their study was not generalizable or realistic (e.g., Harik et al., 2009). Their simulations did not encompass a condition where multiple patterns of drift occurred over time; rather, they only considered one condition per simulation.

In sum, the literature shows that rater drift is inevitable. These studies indicate the need for model-based approaches to scoring CR items that incorporates rater characteristics for measuring examinee ability. In light of these developments, other studies have investigated the effect of rater training and feedback as means to reduce rater drift. The following section describes studies that have examined training and feedback using measures of agreement in the context of rater drift.

Studies on Rater Drift with Efforts to Reduce Rater Effects

To improve consistency and to minimize rating errors, raters must familiarize themselves with the measures they are using, understand the sequence of operation, and explain how they interpret the data. Several empirical studies have shown the effectiveness of these strategies. For example, in Shohamy, Gordon, and Kraemer (1992), it was found that the overall reliability among raters were higher for trained raters than untrained raters, whereas prior experience did not affect their reliability. To a certain degree, rater training may help to alleviate rater differences. However, studies have also shown that completely overcoming them is difficult.

In the context of reducing rater effects over time, Lumley and McNamara (1995) used the FACETS model to examine rater drift with training. They used the Occupational English Test (McNamara, 1990) to examine six criteria on communicative effectiveness with a maximum score of 6 points for each criterion. Data were collected on three occasions, of which the first two comprised training sessions. They concluded that even with multiple training sessions across different occasions, rater severity could not be eliminated. Furthermore, they asserted that there were significant rater variations in severity. The authors concluded that rater severity must be calibrated at each administration to estimate examinee performance, and they called into question the practice of using unadjusted rater scores.

In Wilson and Case (2000), the impact of feedback using estimates of rater severity on half-day intervals from two scoring occasions was examined. They found that it was feasible to provide interpretable feedback to raters on given intervals. However, even with feedback, there were significant rater drift between periods. They also noticed that the effectiveness of the feedback varied from rater to rater. Hoskens and Wilson (2001) extended their study by providing real-time feedback to rater leaders. Feedback was provided using estimates of rater severity in five successive periods. A modified linear logistic test model (LLTM) was used to generate rater severity estimates. Although feedback seemed to draw raters closer to the mean, a controlled test showed that this was not successful. Given an attempt to reduce drift, the authors found that changes in rater behavior was inevitable in their study. Although these empirical findings demonstrate a reduction in measurement error due to training, not all variability in rater severity was

eliminated. These results reiterate that rater drift due to rater severity are difficult to overcome even with training.

Testing contextual effects have also been studied to reduce rater effects (Hughes & Keeling, 1984). Contextual effects refer to raters giving higher scores when an essay is preceded by a poor-quality essay. Using data from high school students scored by 156 first-year college students with model essays, the authors conducted a regression analysis controlling for context quality and scoring instructions. They found that contextual effects were neither reduced nor eliminated; the authors concluded that it was increasingly challenging to find practical methods to overcome context effects.

In a study conducted by Chase (1986), the impact of multiple factors affecting rater scores was examined. These factors included gender, race, reader expectation, and different qualities of penmanship. They conducted a multiple regression analysis controlling for these factors. They also allowed interactions between variables to examine whether there was a joint effect among the factors that explained the variability in scores. The authors found that all four factors had a significant effect on rater scores.

As demonstrated in these studies, rater effects including severity are difficult to eliminate even with repeated efforts to retrain raters and provide feedback; rater effects have also been found to exist for multiple factors affecting rater scores such as contextual effects and characteristics of the rater and the examinee. These studies reinforce the conclusion from the rater drift literature that the use of unadjusted rater scores can bias assessments of examinees. Therefore, studies have suggested the use of scores derived from rater models which accounts for rater effects such as rater severity. However, model-based classifications under rater drift have not been extensively studied in the

literature. This study contributes to this understanding by examining the effect of rater drift on classification accuracy.

The following sections introduce incomplete designs used in most large-scale assessments and present rater models that have incorporated rater severity to improve estimates of examinee's ability.

2.2 Incomplete Designs

Although fully crossed designs (i.e., all raters score all essays) are ideal, most large-scale assessments score CR items using the ratings of two raters. Designs that do not allocate each rater to every essay are known as *incomplete designs*. Variations of these rating designs are documented in Hombo, Donoghue, and Thayer (2001). Examples of incomplete designs for CR scoring include the balanced incomplete block design (BIB) and the unbalanced design.

Balanced incomplete block (BIB) design. The BIB is an efficient design for recovering parameter estimates under the latent class SDT model. The design is defined by a systematic method of allocating essays to each rater and the connectivity among raters, which are balanced under certain constraints (DeCarlo, 2008, 2010). The BIB is divided into n essays (i.e., blocks) of k raters that score each essay, where different raters are assigned to the same essay. There are g raters (i.e., treatments) each of which is grouped in r essays scored by each rater (i.e, blocks). Finally, any two treatments occur together in exactly λ essays scored by each pair of rater (i.e., blocks). In other words, the following properties must be met (DeCarlo, 2008):

$$rg = kn$$
$$\lambda(g-1) = r(k-1)$$
$$r > \lambda$$
$$n \geq g$$

Following the specification above, in a BIB design for 10 raters, each rater scores 216 essays; each of the 45 possible rater pairs scores 108 distinct essays. The uniform pattern exhibited by the BIB design allows rater characteristics to be estimated well (DeCarlo, 2008).

Unbalanced design. Another incomplete design used in large-scale assessments is the unbalanced design. In an unbalanced design, the restrictions specified above are ignored; that is, the number of essays scored by a rater can differ between raters as well as the number of essays scored by rater pairs. Moreover, all possible rater pairs do not have to be used. Table 1 presents an example of an unbalanced design for 10 rater pairs.

Table 1. Unbalanced Incomplete Design, 10 rater pairs

10 Pairs	Rater										Total
	1	2	3	4	5	6	7	8	9	10	
	20	20									20
		40	40								40
			80	80							80
				60	60						60
					140	140					140
						90	90				90
							220	220			220
								150	150		150
									250	250	250
	30									30	30
Total/Rater	50	60	120	140	200	230	310	370	400	280	1080

Each column represents the total number of essays scored by a rater for a total of 10 raters scoring 1080 essays; each row shows the number of essays scored by a pair of raters. As presented above, restrictions from the BIB design on raters are no longer

present; that is, there are an unbalanced number of essays that each rater pair scores as well as an unequal number of essays scored by each rater. DeCarlo (2010) examined different unbalanced designs using 10, 20, and 45 rater pairs using the latent class SDT model. Even under an unbalanced design, the recovery of parameters was good; however, the bias in parameter estimates and in standard error estimates were larger for raters that scored fewer essays.

The ensuing sections describe models used for CR scoring. The latent class SDT model is presented first, and IRT models follow. Differences between the model parameters used to describe rater characteristics as well as scores derived from the rater models are discussed.

2.3 Latent Class Signal Detection Theory (SDT) Model

In the latent class SDT model (DeCarlo, 2002), rating is conceptualized as a psychological process, where a rater's role in scoring a CR item is viewed as attempting to discriminate between latent classes of essays; the latent classes are defined as scores from the scoring rubric. That is, for a CR item with four scoring categories, a rater's task is to classify an essay into one of the four latent scores. In fact, the role of a rater is to discriminate between scores defined in the rubric, which is analogous to discriminating between latent classes.

The latent class SDT model has two parameters that explain the response of a rater: (1) discrimination (d) and (2) response criteria (c_k). Rater discrimination (d) refers to the ability of a rater to discriminate between latent classes of essays, and the response criteria (c_k) represents the internal criteria to which the rater uses to compare and judge

the essay score. Figure 1 presents a representation of the SDT, where four probability distributions of perceptions in essay quality are illustrated. There are three response criteria locations in the figure. These locations represent a rater's criteria for judging a particular score. For example, if an essay is thought to be between c_1 and c_2, then the rater gives the essay a "2." However, if a rater perceives the quality as over c_2, but below c_3, then the score now becomes "3." As such, the response criteria represent a decisional aspect of the rater. Furthermore, it can be inferred from this diagram that by shifting c_3 up, the rater becomes stricter, because this decreases the likelihood of getting a "4." Likewise, by shifting c_1 down, the rater becomes more lenient, because this increases the chance for a rater to assign a higher score. As noted, these shifts in raters' criteria locations represent rater effects, because they allow a rater to be lenient or strict. Furthermore, it can also account for the shrinkage effect in that if the criteria location for c_1 is shifted to the far left, then a rater's chance of assigning a score of "1" becomes very low.

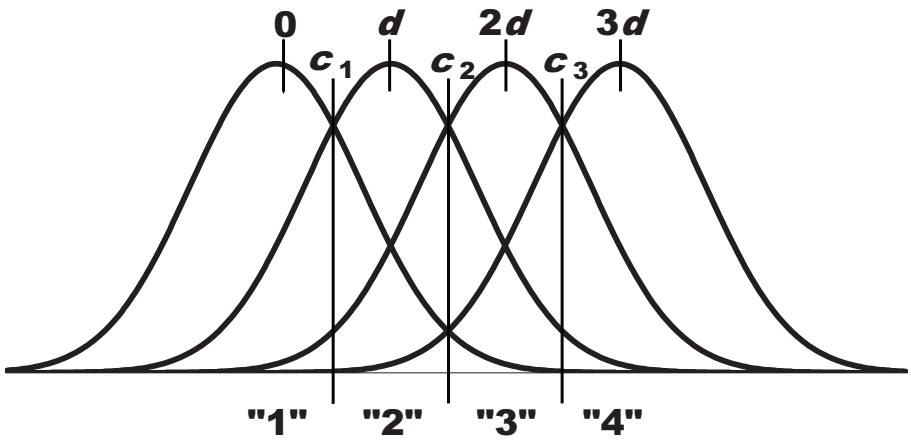


Figure 1. A representation of SDT for scoring categories 1 to 4

The discrimination parameter (d) represents the distance between the probability distributions and reflects a perceptual aspect of the rater. Rater discrimination represents how well a rater discriminates between latent classes of essays. When the distance

between distributions is larger, the rater has greater discrimination between the latent classes, because this means that the perceptions of each scoring category are more distinct. In other words, when *d* is larger, there is less overlap between the distributions and less error in terms of a rater's attempt to classify an essay. If the distance between distributions is small, the ability of a rater to differentiate between two latent classes of essays becomes less clear.

More formally, for *N* items, *J* raters, and *K* discrete scores (such that $1 \leq k \leq K$), the latent class SDT model is expressed as follows:

$$\Pr(Y_j \leq k \mid \eta_c) = F(c_{jk} - d_j \eta_c).$$

Here, Y_j is rater *j*'s observed response, and *F* is the logistic cumulative distribution function. The η_c represents the categorical latent classes, which are the discrete ordered scores of examinee ability defined by the scoring rubric. This model can be decomposed into ordered $1 \leq k \leq K$ categories as follows:

$$\Pr(Y_j = k \mid \eta_c) = F(c_{jk} - d_j \eta_c) \qquad k = 1$$

$$\Pr(Y_j = k \mid \eta_c) = F(c_{jk} - d_j \eta_c) - F(c_{jk-1} - d_j \eta_c) \qquad 2 < k < K$$

$$\Pr(Y_j = k \mid \eta_c) = 1 - F(c_{jk-1} - d_j \eta_c) \qquad k = K$$

Unlike measures of agreement that provide an overall estimate of rater reliability, the latent class SDT model estimates separate criteria locations (c_{jk}) and discrimination (d_j) for each rater. However, to compare criteria locations across raters, they should be standardized to the same scale. As such, the relative criteria can be used (DeCarlo, 2005):

$$\text{rel}\, c_{jk} = c_{jk} / (K-1) d_j.$$

The relative criteria standardize the criteria estimate by rescaling its estimate using rater discrimination and the number of scoring categories (i.e., one minus *K* categories).

Classification accuracy. One of the aims of latent class analysis is to classify examinees into a latent class using the observed response patterns (Dayton, 1998; Clogg, 1995). The posterior probability of the latent variable η_c can be used to measure the quality of this classification. For example, using the probabilities estimated from the above equations, the posterior probability of latent classification for three raters (Y_1,Y_2,Y_3) can be obtained (DeCarlo, 2002), which can be used as a measure of how well raters classify an essay into the latent classes:

$$\Pr(\eta_c \mid Y_1, Y_2, Y_3) = \frac{\Pr(\eta_c)\Pr(Y_1, Y_2, Y_3 \mid \eta_c)}{\sum_{c=1}^{C} \Pr(\eta_c)\Pr(Y_1, Y_2, Y_3 \mid \eta_c)} = \frac{\Pr(\eta_c)\Pr(Y_1 \mid \eta_c)\Pr(Y_2 \mid \eta_c)\Pr(Y_3 \mid \eta_c)}{\sum_{c=1}^{C} \Pr(\eta_c)\Pr(Y_1 \mid \eta_c)\Pr(Y_2 \mid \eta_c)\Pr(Y_3 \mid \eta_c)}.$$

Two measures for classification accuracy are presented for these purposes. These measures are used in this study to reflect the accuracy of classification derived from the latent class SDT model under rater drift. First, the expected proportion of cases correctly classified (P_c) is calculated as follows (DeCarlo, 2002):

$$P_c = \sum_s [n_s \times \max \Pr(\eta_c \mid Y_1, Y_2, ..., Y_J)] / N .$$

Here, s indicates the unique response patterns and n_s corresponds to the frequency of each pattern. Furthermore, $\max \Pr(\eta_c \mid Y_1, Y_2, ..., Y_J)$ is the maximum posterior probability across the latent classes for a given response pattern, and *N* is the total number of cases. In addition to the proportion correctly classified statistic (P_c), the lambda statistic (λ) is considered, which accounts for classification that can occur by chance. This statistic can

be important when there is a latent class with a large size (DeCarlo, 2002). The lambda statistic is calculated as follows:

$$\lambda = \frac{P_c - \max \Pr(\eta_c)}{1 - \max \Pr(\eta_c)} \ .$$

Both proportion correctly classified (P_c) and the lambda statistic (λ) are used in this study to study classification accuracy.

2.4 Item Response Theory (IRT) models

Item response theory (IRT) models use response patterns as indicators of latent ability. The models presented in this section estimate rater discrimination and rater threshold parameters by considering raters as items (Wilson & Case, 2000; Masters & Wright, 1997). In general, rater discrimination is a measure of how well raters discriminate between different qualities of essays, and the threshold parameter expresses information on rater effects such as rater severity. This section provides an overview of four IRT models used for CR scoring: graded response (GR) model (Samejima, 1969), partial credit (PC) model (Masters, 1992), generalized partial credit (GPC) model (Muraki, 1992), and the FACETS model (Linacre, 1989).

Graded Response Model

The graded response (GR) model (Samejima, 1969) considers scores as ordered polytomous categories. In other words, for a given latent ability (θ), the GR model estimates the conditional probability that an examinee successfully masters a task up to a particular score. For J raters and K scoring categories, we have the following equation (McDonald, 1999):

$$\Pr(Y_j \le k_j \mid \theta) = F[a_j(\theta - b'_{jk})] = F(a_j\theta - b_{jk}).$$

The *F* represents a logistic cumulative density function (cdf), which characterizes the ordered nature of the model; θ is a continuous latent ability variable. There are two parameters in the model. The discrimination parameter (a_j) measures the ability of raters to discriminate between essays of different quality, and the threshold parameter (b_{jk}), which is a product of the b'_{jk} and a_j, measures rater effects. The latent class SDT model is related to the GR model in that the discrimination parameters of the two models are analogous and that the threshold parameter is related to the criteria parameter. However, the difference lies in the latent ability variable. In the GR model, θ is continuous, whereas in the latent class SDT model, η_c is discrete. The latent class SDT model can be viewed as a semi-parametric version of the GR model (DeCarlo, 2005).

Partial Credit Model and Generalized Partial Credit Model

The partial credit (PC) model (Masters, 1982) is another IRT model that is used to score essays (Wright & Masters, 1982). Rather than considering responses as cumulative (e.g., GR model), the PC model calculates the conditional probability of adjacent scoring categories (i.e., scoring in category $k + 1$ versus category k). The following is the PC model:

$$\log[\frac{\Pr(Y_j = k_j + 1 \mid \theta)}{\Pr(Y_j = k_j \mid \theta)}] = \theta - b_{jk}.$$

Unlike the GR model, the PC model uses adjacent category logits as shown above (Agresti, 2002). The parameter b_{jk} is the item step or difficulty parameter. This parameter (b_{jk}) is the location in the continuous latent ability scale where two adjacent categories intersect.

The generalized partial credit (GPC) model (Muraki, 1992) is an extension of the PC model that incorporates the discrimination parameter as follows:

$$\log[\frac{\Pr(Y_j = k_j + 1 \mid \theta)}{\Pr(Y_j = k_j \mid \theta)}] = a_j(\theta - b_{jk}).$$

The GPC model has the same interpretation as the PC model with the exception that it has an additional discrimination parameter (a_j).

In a study by Boughton, Klinger, and Gierl (2001), the GPC model and the GR model were compared for their utility in scoring essays. Using a simulation, they found that the GR model was better than the GPC model in terms of estimation and parameter recovery; they compared various numbers of scoring response categories ranging from 4, 6, and 8, and found that as the number of scoring categories increased, estimation improved. Furthermore, they noted that for both models, estimation was poor for the threshold parameters that were at the extremes. For example, for the four-point scale, the b_{j1} and the b_{j3} parameters were poorly estimated. The authors noted that this was due to the sparseness of essays scored at extreme categories. The study also assessed the effects of rater error on estimation; rater error was generated by changing rater scores to incorrect values. They found that the GPC model was better than the GR model under rater error.

FACETS Model

One of the most commonly used models to evaluate rater performance is the FACETS model (Linacre, 1989). This model includes item and examinee parameters to incorporate additional additive effects on the logit scale known as *facets*. That is, each additive effect is measured as a facet, where the model encompasses an item facet, an examinee facet, and a rater facet. The FACETS model estimates the conditional

probability that rater *j* scores item *m* in category *k* given examinee's ability, $\Pr(Y_{mj} = k \mid \theta)$, as the following:

$$\log\left[\frac{\Pr(Y_{mj} = k+1 \mid \theta)}{\Pr(Y_{mj} = k \mid \theta)}\right] = \theta - b_m - \gamma_k - c_j .$$

The equation above results in a three-facet model (i.e., examinee, item, and rater facets). Here, the parameter b_m represents item difficulty, and the parameter γ_k is the item step parameter. The parameter c_j estimates rater severity, which represents how lenient or strict a rater scores; c_j also determines the magnitude of shift in the item response function along the ability scale. An advantage of the model is that it places all parameters in the common linear log-odds scale, centered at a common origin (Lunz, Wright, & Linacre, 1990). However, the FACETS model assumes discrimination to be constant across raters. This means that the model ignores the possibility that some raters may discriminate better than others.

For a single CR item (i.e., two-facet design) the FACETS model can be written as follows:

$$\log[\frac{\Pr(Y_j = k+1 \mid \theta)}{\Pr(Y_j = k \mid \theta)}] = \theta - f_k - c_j .$$

Here, θ is the examinee ability, c_j is the severity of rater *j*, and f_k is the difficulty of the step from category $k + 1$ to k. The single CR item FACETS model is related to the PC model, in that the step parameter (b_{jk}) of the PC model combines the effect of the rater severity (c_j) and the difficulty step parameter (f_k); furthermore, the FACETS model also uses adjacent category logits.

One note to consider for parameters of polytomous IRT (e.g., GR, GPC, and PC) models is that they confound rater effects with item and examinee effects. In tests that use multiple raters, the item response has a three-way interaction between examinee, item, and raters (Tate, 1999). Rather than viewing parameters as rater parameters, they should be considered as *item/rater parameters*; therefore, direct estimates of rater effects cannot be obtained in IRT models. Tate (1999) proposed several methods to separate item and rater parameters that involve linking item parameters between test administrations. However, DeCarlo (2011) showed that using the latent class SDT model, rater parameters can be recovered without linking item parameters. This was demonstrated by generating data with specific item and rater parameters that changed between two occasions. The latent class SDT model correctly recovered the generating rater parameters. Unlike IRT models that scale the ability parameter, the discrete latent classes of the SDT model allow direct estimation of rater parameters.

Chapter III

METHODS

This study investigates the effect of rater drift on model-based classifications. In the empirical study, parameters from rater models were used to identify patterns of drift using estimates of rater severity (response criteria for the latent class SDT model and threshold or step parameter for the IRT models) and rater discrimination. Furthermore, parameter estimates were compared to examine patterns of drift indicated by different rater models. When rater characteristics deviate between testing sessions due to random shifts in rater perception or due to training, the accuracy of model-based classifications may change. The empirical analysis investigates patterns of rater drift for the same rater and also examines how drift affects the latent classification of scores.

The effect of drift on classification accuracy was further examined using simulations by varying levels of rater severity and discrimination using the latent class SDT model. The ability of rater models to detect drift was also examined. This chapter describes methods that were employed in the real-world data analysis as well as in the simulation.

3.1 Empirical Study

The empirical analysis examined the effect of rater drift on the classification of latent scores. Drift was assessed using parameters estimated from different rater models. Parameters c_{jk} and d_j were examined from the latent class SDT model, and b_{jk} and a_j were examined from the IRT models to determine drift in rater severity and rater

discrimination, respectively. Furthermore, changes in the accuracy of latent scores due to drift were studied. More specifically, the empirical study investigated the following research questions:

(1) What patterns of rater drift appear in a large-scale assessment?

(2) How do parameters that measure rater severity differ across rater models?

(3) What is the variability of drift in rater severity and discrimination over time?

(4) How does rater drift affect classification accuracy?

IRT models (GR and GPC models) and the latent class SDT model were used to fit data from different scoring occasions for two real-world data sets. To investigate patterns of drift in real-world data, parameter estimates from the models were examined; the threshold or step parameters (b_{jk}) in IRT models and the response criteria (c_{jk}) in the latent class SDT model represent rater effects. As such, shifts in these parameters indicate a change in rater severity or deviations in category usage. In the latent class SDT model and the GR model, when parameters reflecting rater effects shift up, raters are stricter; if these parameters shift down, raters are more lenient. The discrimination parameters (a_j for the IRT models and d_j for the latent class SDT model) represent how well raters discriminate essays of different quality.

The specific interpretations of parameters in each model also differ. For example, the discrimination parameter of the latent class SDT model indicates the ability of a rater to discriminate between discrete latent classes of essays, whereas the discriminate parameter for the GR model shows how well a rater discriminates between different qualities of essays that are measured in a continuous latent scale. The IRT models and the latent class SDT parameters were estimated using Latent Gold 4.5 (Vermunt & Magidson,

2007), which uses an EM algorithm then switches to the Newton-Raphson iterative process to finalize the estimation process. To avoid boundary estimation problems that are often found in latent class models, posterior mode estimation was used (Galindo-Garre & Vermunt, 2006).

Estimates of rater parameters were plotted for each rater to assess patterns of drift. For instance, a rater can be more lenient over different scoring occasions or stricter; a rater can also have higher discrimination between tests. These patterns for rater effects and discrimination provide information about different trends in raters' behavior over time. Moreover, by examining different plots of parameters, implications of rater behavior can be examined for different rater models. The latent class SDT model and the IRT models differ in that the former uses discrete ordered categories, whereas the latter uses a continuous scale to estimate ability. These differences can show variations in how the models detect rater drift. Overall trends in rater parameters were summarized using regression, where the parameter estimates were examined for linear and nonlinear trends. This was conducted by regressing time on rater parameter estimates so that the slope of this regression indicates drift in the parameters.

To examine how rater drift affects scoring, the proportion correct (P_c) and the lambda (λ) statistics were calculated using posterior probability estimates from the latent class SDT model to measure classification accuracy. The P_c statistic measures the overall quality of the classification, and the λ statistic measures the increase in classification accuracy from using the model, over classifying an essay into the largest latent class (Dayton, 1998). By examining scoring accuracy measures, the impact of rater drift on scores was evaluated.

The real-world data used for the empirical study were taken from two sources: a teacher certification test and a high school writing test. Both are large-scale assessments. The teacher certification exam was scored in a 1 to 6 scale with 45 raters. The high school writing test was scored in a 1 to 4 scale with 28 raters. The teacher certification exam covered seven testing administrations, whereas the high school writing test spanned twelve scoring occasions for each month of the year. The differences in the number of scoring categories and the number of raters as well as the substantive context of these assessments motivates the comparison of two real-world data examples.

In summary, the empirical study examined patterns of drift in rater severity and discrimination for real-world data. Rater drift was assessed using plots of parameters derived from IRT models and the latent class SDT model. Overall trends in parameter estimates were summarized using regression to examine changes in rater behavior. Furthermore, to investigate the effect of drift on latent scores, classification accuracy from the latent class SDT model was used to evaluate how drift affected the quality of classification.

3.2 Simulation Study

The simulation study was conducted to investigate changes in classification accuracy under different conditions of rater drift. Moreover, parameters from the IRT model was examined to assess whether rater drift generated from the latent class SDT model can be detected. The simulation study addressed four research questions:

(1) How do changes in rater severity affect classification accuracy?

(2) How do changes in rater discrimination affect classification accuracy?

(3) Can an IRT model detect rater drift generated from a latent class SDT model?

(4) Does the normality in latent category distribution affect parameter estimates?

(5) Does a shift in the latent class sizes affect parameter estimates?

To answer these questions, two simulation studies were conducted. The first study examined changes in latent class SDT model parameters – both criteria and discrimination – to determine whether drift affected classification accuracy measured using the proportion correct (P_c) and the lambda statistics (λ). In the second study, data generated from the latent class SDT model were fit using an IRT model. The GR model was used as it resembles the latent class SDT model in many ways discussed previously such as the use of cumulative logits to parameterize rater effects.

Study 1: Examining Changes in Classification Accuracy due to Rater Drift

Study 1 was divided into two subsections. The first section examined changes in classification accuracy due to drift in rater effects. In the latent class SDT model, this is represented by the criteria parameter. There were three conditions used in this simulation study. Ten raters with normally distributed rater discrimination with a mean of 4 were generated (see Table 2). Raters' criteria values were generated at the mid-point criteria locations using equidistant spacing. For example, for rater 1 with a rater discrimination population value of 2, the criteria locations for the five locations were 1, 3, 5, 7, and 9, respectively.

Table 2 shows two scoring occasions that represent drift in rater severity. Condition 1 shows a shift down in the response criteria for raters 4 to 9 between the two scoring occasions; this indicates leniency in raters between the two test administrations. In condition 2, response criteria were raised for raters 4 to 9, making raters stricter. In

condition 3, response criteria were shifted up for raters 1, 4, 5, and 8; response criteria were shifted down for raters 3, 6, 7, 9, and 10. This condition allowed raters to be both lenient and strict when compared to the first scoring occasion.

Table 2. Conditions for study of rater drift

Condition	Rater	Rater Parameters											
		First Scoring Occasion						Second Scoring Occasion					
		d_j	c_{j1}	c_{j2}	c_{j3}	c_{j5}	c_{j5}	d_j	c_{j1}	c_{j2}	c_{j3}	c_{j5}	c_{j5}
1 (more lenient; shift down in c_j for some raters)	1	2	1	3	5	7	9	2	1	3	5	7	9
	2	3	1.5	4.5	7.5	10.5	13.5	3	1.5	4.5	7.5	10.5	13.5
	3	3	1.5	4.5	7.5	10.5	13.5	3	1.5	4.5	7.5	10.5	13.5
	4	4	2	6	10	14	18	4	1	5	9	13	17
	5	4	2	6	10	14	18	4	1	5	9	13	17
	6	4	2	6	10	14	18	4	1	5	9	13	17
	7	4	2	6	10	14	18	4	1	5	9	13	17
	8	5	2.5	7.5	12.5	17.5	22.5	5	1.5	6.5	11.5	16.5	21.5
	9	5	2.5	7.5	12.5	17.5	22.5	5	1.5	6.5	11.5	16.5	21.5
	10	6	3	9	15	21	27	6	3	9	15	21	27
2 (stricter; shift up in c_j for some raters)	1	2	1	3	5	7	9	2	1	3	5	7	9
	2	3	1.5	4.5	7.5	10.5	13.5	3	1.5	4.5	7.5	10.5	13.5
	3	3	1.5	4.5	7.5	10.5	13.5	3	1.5	4.5	7.5	10.5	13.5
	4	4	2	6	10	14	18	4	3	7	11	15	19
	5	4	2	6	10	14	18	4	3	7	11	15	19
	6	4	2	6	10	14	18	4	3	7	11	15	19
	7	4	2	6	10	14	18	4	3	7	11	15	19
	8	5	2.5	7.5	12.5	17.5	22.5	5	3.5	8.5	13.5	18.5	23.5
	9	5	2.5	7.5	12.5	17.5	22.5	5	3.5	8.5	13.5	18.5	23.5
	10	6	3	9	15	21	27	6	3	9	15	21	27
3 (both lenient and strict; shifts in c_j for some raters)	1	2	1	3	5	7	9	2	2	4	6	8	10
	2	3	1.5	4.5	7.5	10.5	13.5	3	1.5	4.5	7.5	10.5	13.5
	3	3	1.5	4.5	7.5	10.5	13.5	3	0.5	3.5	6.5	9.5	12.5
	4	4	2	6	10	14	18	4	3	7	11	15	19
	5	4	2	6	10	14	18	4	3	7	11	15	19
	6	4	2	6	10	14	18	4	1	5	9	13	17
	7	4	2	6	10	14	18	4	1	5	9	13	17
	8	5	2.5	7.5	12.5	17.5	22.5	5	3.5	8.5	13.5	18.5	23.5
	9	5	2.5	7.5	12.5	17.5	22.5	5	1.5	6.5	11.5	16.5	21.5
	10	6	3	9	15	21	27	6	2	8	14	20	26

Note: Condition 1 specified a shift down in the response criteria for raters 4 to 9 to allow raters to be more lenient. In condition 2, response criteria were raised, making raters stricter. In condition 3, a combination of rater effects were implemented.

The simulation study examined changes in classification accuracy under rater drift using the latent class SDT model. Classification accuracy was measured using the proportion correct (P_c) and the lambda (λ) statistics.

The second part of this section examined changes in classification accuracy when rater discrimination increased. Table 3 shows the population values for this simulation:

Table 3. Condition for drift in rater discrimination

Rater	Rater Parameters											
	First Scoring Occasion						Second Scoring Occasion					
	d_j	c_{j1}	c_{j2}	c_{j3}	c_{j5}	c_{j5}	d_j	c_{j1}	c_{j2}	c_{j3}	c_{j5}	c_{j5}
1	0.5	0.25	0.75	1.25	1.75	2.25	2	1	3	5	7	9
2	1	0.5	1.5	2.5	3.5	4.5	3	1.5	4.5	7.5	10.5	13.5
3	1	0.5	1.5	2.5	3.5	4.5	3	1.5	4.5	7.5	10.5	13.5
4	2	1	3	5	7	9	4	2	6	10	14	18
5	2	1	3	5	7	9	4	2	6	10	14	18
6	2	1	3	5	7	9	4	2	6	10	14	18
7	2	1	3	5	7	9	4	2	6	10	14	18
8	3	1.5	4.5	7.5	10.5	13.5	5	2.5	7.5	12.5	17.5	22.5
9	3	1.5	4.5	7.5	10.5	13.5	5	2.5	7.5	12.5	17.5	22.5
10	4	2	6	10	14	18	6	3	9	15	21	27

Note: In the first scoring occasion, rater discrimination (d) had population values that were normally distributed with mean of 2; in the second scoring occasion, d was raised to be normally distributed with mean of 4.

Between the first and second scoring occasions, rater discrimination increased by two units for all raters, except rater 1; the discrimination for rater 1 had a population value of 0.5 for the first scoring occasion to indicate a value close to 0. Mid-point criteria locations were used with equidistant spacing; as rater discrimination increased, criteria locations also changed. In the first scoring occasion, rater discrimination had population values that were normally distributed with a mean of 2; in the second scoring occasion, this was raised to be normally distributed with a mean of 4. To estimate classification accuracy for the conditions described, the latent class SDT model was fit separately for

the two time points. Classification accuracy measures were examined for each scoring occasion to assess whether drift affected scoring accuracy.

Study 2: Detecting Drift using Rater Models

In this section, data were generated using the latent class SDT model following specifications from Table 2 (p. 33) and from Table 3 (p. 34). This data were fit using the GR model to examine whether drift in rater severity and in rater discrimination can be detected by an IRT model. That is, this investigated whether the GR model was sensitive to detect data indicating drift in rater severity and discrimination.

Differences in latent class sizes between two time periods were also investigated to examine how this affected parameters in the latent class SDT model and in the GR model. Table 4 shows three conditions using 6 and 4 scoring categories. In both scoring occasions, mid-point criteria were used for raters with population values of the discrimination normally distributed with a mean of 4. The difference between the two scoring occasions is in the latent class sizes. For the first condition with 6 categories, the first scoring occasion had normally distributed sizes with 0.08, 0.17, 0.25, 0.25, 0.17, and 0.08 for the six latent classes, respectively. However, in the second scoring occasion, this was changed to a non-normal distribution with sizes of 0.03, 0.03, 0.40, 0.40, 0.10, and 0.04 with a concentration of density at the middle classes, 3 and 4. The second condition with 4 categories followed a similar pattern. The first scoring occasion had latent class sizes of 0.17, 0.33, 0.33, and 0.17 to represent a normal distribution of scores. The second occasion had a non-normal distribution with class sizes of 0.07, 0.43, 0.43, and 0.07 for the four classes, respectively.

Table 4. Conditions for differences in latent class sizes over two scoring occasions

# of categories and condition	Parameters											
	First Scoring Occasion						Second Scoring Occasion					
	Class	LC1	LC2	LC3	LC4	LC5	LC6	LC1	LC2	LC3	LC4	LC5	LC6
	Size	0.08	0.17	0.25	0.25	0.17	0.08	0.03	0.03	0.40	0.40	0.10	0.04
6	Rater	d_j	c_{j1}	c_{j2}	c_{j3}	c_{j5}	c_{j5}	d_j	c_{j1}	c_{j2}	c_{j3}	c_{j5}	c_{j5}
	1	2	1	3	5	7	9	2	1	3	5	7	9
Condition 1:	2	3	1.5	4.5	7.5	10.5	13.5	3	1.5	4.5	7.5	10.5	13.5
	3	3	1.5	4.5	7.5	10.5	13.5	3	1.5	4.5	7.5	10.5	13.5
Change in normality of latent class sizes	4	4	2	6	10	14	18	4	2	6	10	14	18
	5	4	2	6	10	14	18	4	2	6	10	14	18
	6	4	2	6	10	14	18	4	2	6	10	14	18
	7	4	2	6	10	14	18	4	2	6	10	14	18
	8	5	2.5	7.5	12.5	17.5	22.5	5	2.5	7.5	12.5	17.5	22.5
	9	5	2.5	7.5	12.5	17.5	22.5	5	2.5	7.5	12.5	17.5	22.5
	10	6	3	9	15	21	27	6	3	9	15	21	27
	Class	LC1	LC2	LC3	LC4			LC1	LC2	LC3	LC4		
	Size	0.17	0.33	0.33	0.17			0.07	0.43	0.43	0.07		
4	Rater	d_j	c_{j1}	c_{j2}	c_{j3}			d_j	c_{j1}	c_{j2}	c_{j3}		
	1	2	1	3	5			2	1	3	5		
Condition 2:	2	3	1.5	4.5	7.5			3	1.5	4.5	7.5		
	3	3	1.5	4.5	7.5			3	1.5	4.5	7.5		
Change in normality of latent class sizes	4	4	2	6	10			4	2	6	10		
	5	4	2	6	10			4	2	6	10		
	6	4	2	6	10			4	2	6	10		
	7	4	2	6	10			4	2	6	10		
	8	5	2.5	7.5	12.5			5	2.5	7.5	12.5		
	9	5	2.5	7.5	12.5			5	2.5	7.5	12.5		
	10	6	3	9	15			6	3	9	15		
	Class	LC1	LC2	LC3	LC4			LC1	LC2	LC3	LC4		
	Size	0.07	0.5	0.4	0.03			0.03	0.4	0.5	0.07		
4	Rater	d_j	c_{j1}	c_{j2}	c_{j3}			d_j	c_{j1}	c_{j2}	c_{j3}		
	1	2	1	3	5			2	1	3	5		
	2	3	1.5	4.5	7.5			3	1.5	4.5	7.5		
Condition 3:	3	3	1.5	4.5	7.5			3	1.5	4.5	7.5		
	4	4	2	6	10			4	2	6	10		
Shift in density	5	4	2	6	10			4	2	6	10		
	6	4	2	6	10			4	2	6	10		
	7	4	2	6	10			4	2	6	10		
	8	5	2.5	7.5	12.5			5	2.5	7.5	12.5		
	9	5	2.5	7.5	12.5			5	2.5	7.5	12.5		
	10	6	3	9	15			6	3	9	15		

Note: Three conditions are presented using 6 and 4 scoring categories for two time points. In the first two conditions, the first scoring occasion had normally distributed latent class sizes; the second scoring occasion had a distribution where class sizes were concentrated in the middle categories. For the third condition, there was a shift in the class sizes.

The rationale for using these conditions was that many large-scale assessments use either 4 or 6 scoring categories as presented in the two empirical data sets in this study. In addition, many IRT models such as the GR model assume a normal distribution of examinee ability. As such, the implications of fitting a non-normally distributed condition using IRT can be investigated; the effect of latent class sizes on parameter estimates was designed to assess whether IRT models can detect non-normal distributions of scores.

The simulation also includes a condition for a shift in latent class sizes, meaning that there was a change in the proportion of scores. This is presented in condition 3 of Table 4. For the first scoring occasion, the latent class sizes were generated using population values of 0.07, 0.50, 0.40, and 0.03; in the second scoring occasion, it changed to 0.03, 0.40, 0.50, and 0.07 for the four latent classes, respectively. This condition also presents a change in the proportions at the end categories, while maintaining a high concentration of scores in the middle categories.

The data generated from these conditions were fit using both the latent class SDT model and the GR model to examine whether changes in latent class sizes can be recovered. Both BIB and unbalanced designs (see specification in Table 1 for the unbalanced design) were used, which represent designs used in many large-scale assessments. Unbalanced designs are used for assessments such as Praxis and the TOEFL (DeCarlo, 2008).

A SAS macro used in DeCarlo (2010) was implemented to create fully-crossed data sets with 6 latent classes (or 4 latent classes depending on the condition) for 10 raters and 1,080 essays using the latent class SDT model. The population values for class sizes

were 0.08, 0.17, 0.25, 0.25, 0.17, and 0.08, respectively, for all data generation except for conditions specified in Table 4 of Study 2. These class sizes represent a normal distribution of scores. Following the generation of data, fully-crossed data sets were transformed into incomplete designs. A SAS macro generated 100 replications of the conditions with corresponding Latent Gold input files and a DOS batch file. A different macro summarized the results from the replication and provided information on classification, parameter recovery, and standard errors of the simulated data.

In summary, the simulation studies presented in this chapter examined the effect of drift on classification accuracy using the latent class SDT model. These simulations examined rater drift using two testing occasions to assess classification accuracy in the latent class SDT model. These conditions were studied within the framework of incomplete designs specified by the BIB and the unbalanced design to resemble rating formats used in many large-scale assessments. The results from this study can be used to understand the relationship between rater drift and classification accuracy. These findings were also used to investigate implications for rater training in the literature that have been focused on rater severity.

The second part of the simulation study examined the ability of rater models to detect drift. This was conducted by generating data using the latent class SDT model for drift in rater effects and in rater discrimination. Variations of latent class sizes between testing occasions were also examined. Both normal and non-normal distributions in latent class sizes were generated for 6 and 4 scoring categories; shifts in latent class sizes were also generated. Data were fit using the GR model to assess whether data generated from the latent class SDT model could be detected by an IRT model. Furthermore, results

from these studies can be used to provide researchers with a greater understanding of the effects that rater drift have in the context of different rater models.

Chapter IV

RESULTS

This chapter presents findings from the empirical and simulation studies, which are both divided into two separate sections. In the empirical section, results from the teacher certification test and the high school writing test are presented. The simulation study presents findings on the effects of rater drift on classification accuracy. Simulation results also indicate how well an IRT model such as the GR model, detects drift when data were generated using the latent class SDT model. The effects of changing the distribution of latent class sizes on parameter estimation were also examined for both the latent class SDT model and the GR model.

4.1 Empirical Study: Teacher Certification Test

This section uses the latent class SDT model and IRT models to examine patterns of drift in a teacher certification test used nationally to license instructors entering the teaching profession. The essay section from this test was used in the analysis. Among 45 raters that scored the essays, the ratings of 32 raters were used; these 32 raters were selected on the basis that they consistently scored on 6 or more administrations of the test (there were 7 total administrations of the test). The CR item was scored on a 1 to 6 scale, with a higher score representing greater mastery. For each of the 7 administrations, there were 3326, 10659, 4804, 6257, 7014, 5450, and 3387 examinees (Mean=5842, SD=2528), respectively.

Teacher Certification Test: Rater Effects

Plots of rater parameters. Figure 2 present plots of the relative criteria parameters for the latent class SDT model. Figures 3 and 4 show plots of the threshold and step parameters for the two IRT models. For each plot, the *X*-axis represents the 7 administrations, and the *Y*-axis represents the relative criteria or the threshold values in the latent class SDT model and the IRT models, respectively. The estimates of the relative criteria in Figure 2 rescale the criteria locations for each rater so that the relative criteria are between 0 and 1; this allows criteria locations to be comparable between raters (DeCarlo, 2008). For 6 distributions (since the essay is scored in a 1 to 6 scale), the means are located at 0.0, 0.2, 0.4, 0.6, 0.8 and 1.0. Horizontal lines were added at intersection-point criteria locations of the six distributions. They are the midway points between the means and are therefore at 0.1, 0.3, 0.5, 0.7 and 0.9. Relative criteria estimates above this line indicate a stricter rating; estimates below indicate a more lenient rating. As such, these lines serve as a relative guide to indicate rater effects such as severity and scale shrinkage. It is noted here that intersection-point locations cannot be derived for IRT models, as they do not have the same conceptualization of latent classes, which are used to create these markers.

Using the intersection-point criteria as a relative guide, plots in Figure 2 can be examined for rater effects. In general, most raters were consistent in their scoring; that is, their plots lied mostly on the intersection-point locations. This indicates that the level of severity among most raters were constant during the 7 scoring administrations. However, the plots can also be used to identify raters that were strict or lenient.

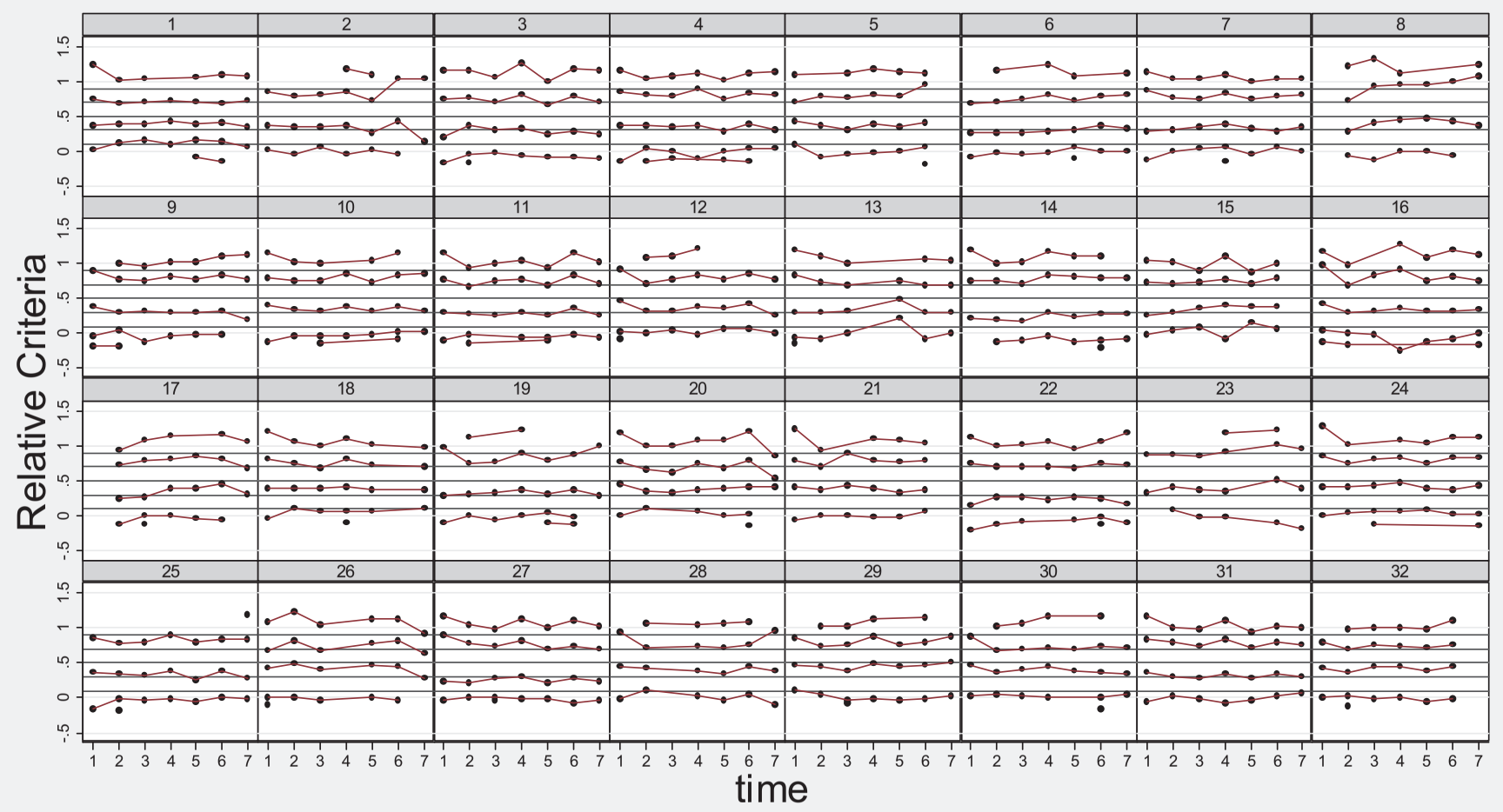


Figure 2. Teacher certification test: Plot of individual rater's relative criteria (latent class SDT)

Note: Only raters that scored 6 or more administrations were included (16 raters eliminated). Horizontal lines represent optimal cut-point criteria at 0.1, 0.3, 0.5, 0.7, and 0.9.

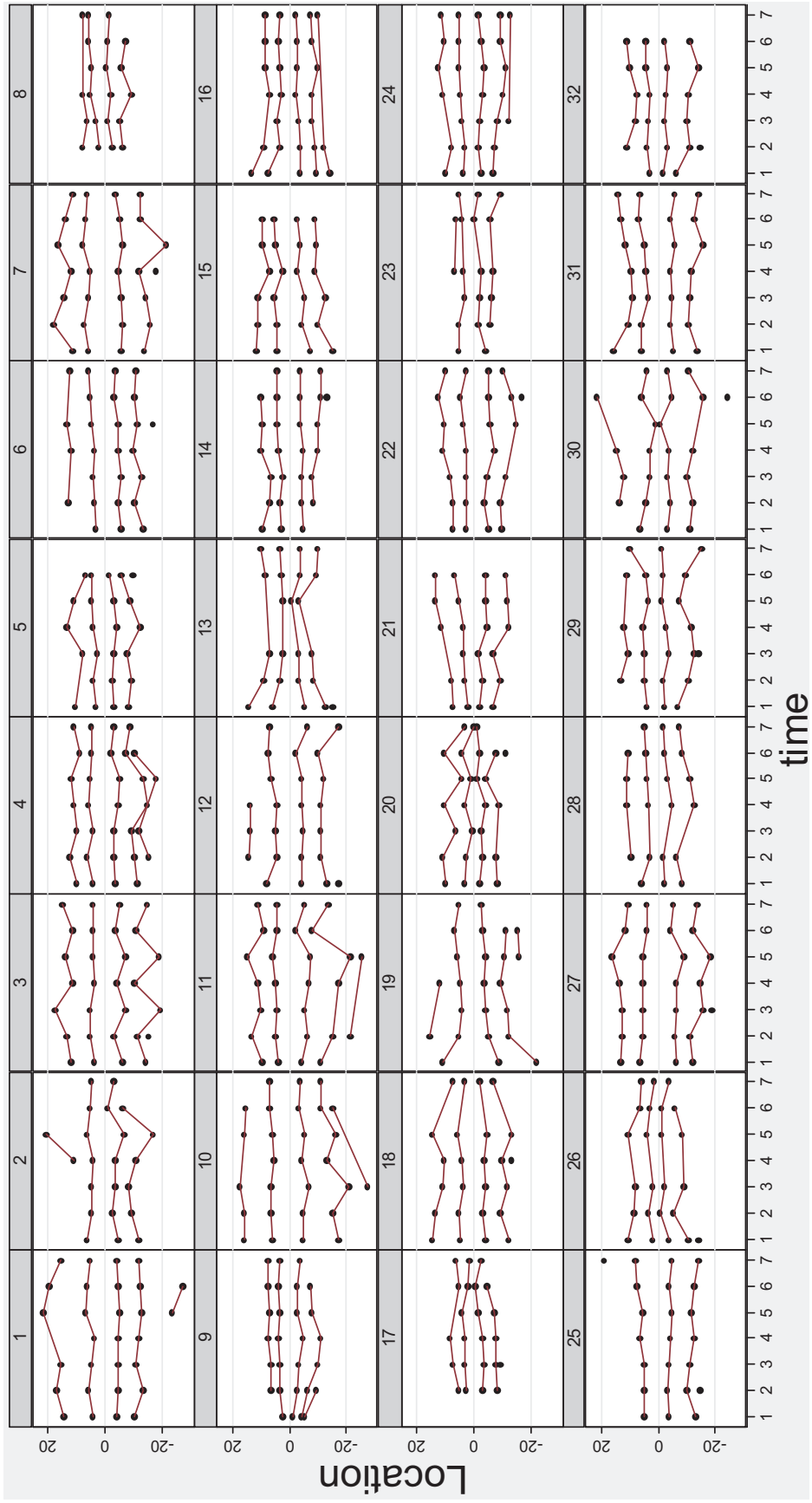


Figure 3. Teacher certification test: Plot of individual rater's location (graded response model)

Note: Only raters that scored 6 or more administrations were included (16 raters eliminated)

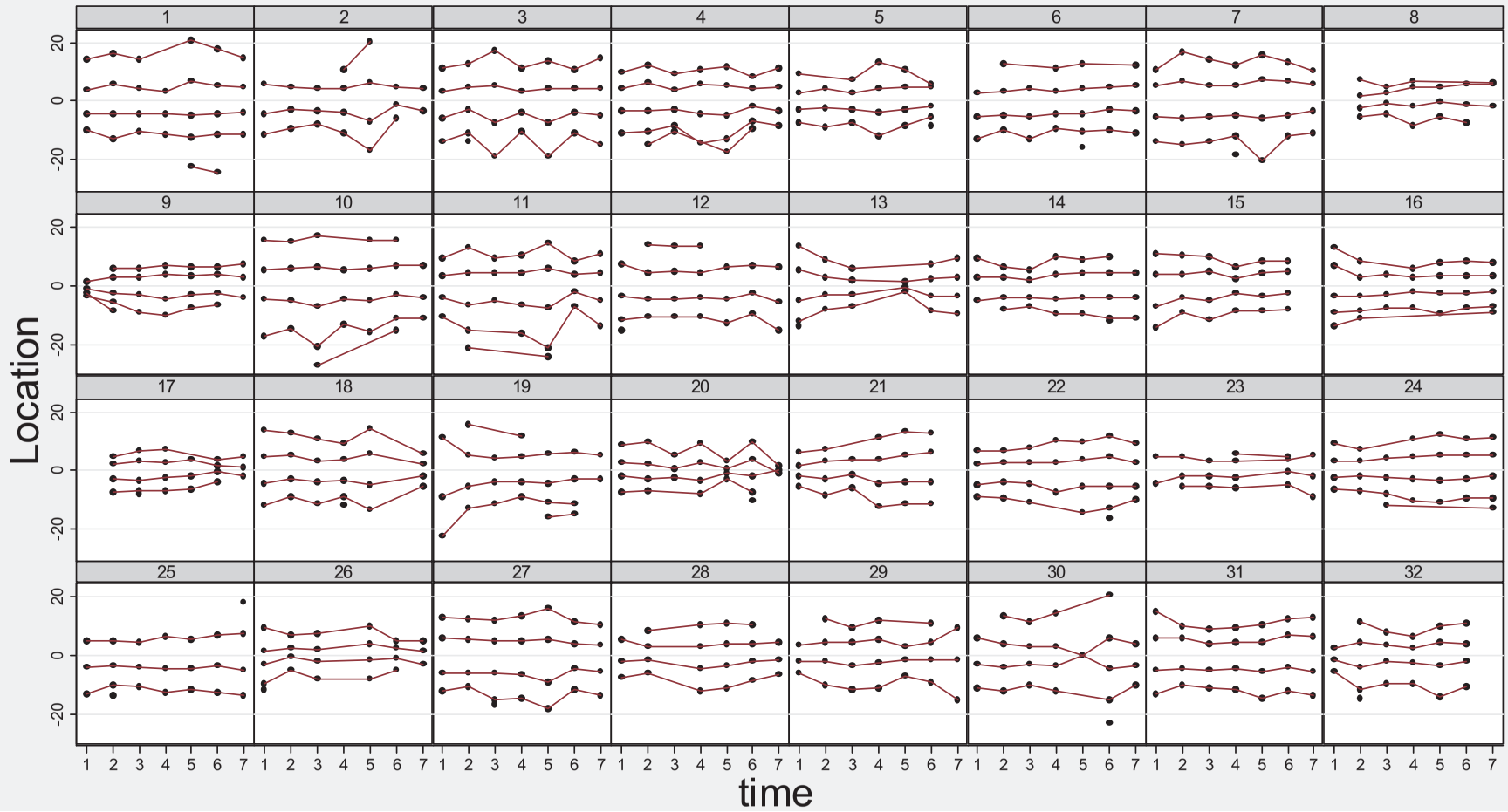


Figure 4. Teacher certification test: Plot of individual rater's location (generalized partial credit model)

Note: Only raters that scored 6 or more administrations were included (16 raters eliminated)

Although minor, some raters showed strictness in their use of the "6" category. This can be identified by criteria locations that were higher than the intersection-criteria locations for the fifth criteria estimates. Furthermore, scale shrinkage effects were also identified for raters that do not have five criteria estimates for each administration.

For the purposes of presenting an example, Rater 3 is used to discuss rater effects. In Figure 2, Rater 3's use of the fifth criteria was above the intersection-point location, indicating that this rater was strict on the use of scoring a "6." In addition, for all administrations except the second scoring occasion, this rater only used scores from 2 to 6. This demonstrates a scale shrinkage effect. Examining the relative location of the second criteria, which is below the horizontal line, rater 3 was also lenient in scoring a "2." Based on this figure, this rater tended not to use the "1" category and was stricter on the higher scores, while lenient on the use of the lower scoring categories. As demonstrated from this example, these plots provide informative detail about a rater's scoring behavior over the seven administrations.

Rater effects also appear in the IRT plots of threshold locations and step parameters for the GR and GPC models in Figures 3 and 4, respectively. Plots of parameters from these IRT models were similar. However, in comparison to the relative criteria locations using the latent class SDT model, the parameters from the IRT models were difficult to interpret as there are no natural intersection-points of reference. Although the IRT plots were also fairly stable across the seven scoring administrations, there were differences when compared to the latent class SDT model in Figure 2. Using Rater 3 as an example again, the IRT plots indicate that there were drift in this rater's use of the "2" category in that it fluctuated between the administrations. However, the

relative criteria plots for the lowest category were stable throughout the 7 occasions. These results indicate that the latent class SDT model and the IRT models differ in their presentation of rater drift – but only in what appear to be minor ways.

Parameter estimates summarized using regression. To examine the overall trend in the parameter estimates, a regression was used to summarize changes in rater parameter estimates. For example, changes in rater criteria were summarized by using c_j as outcomes in a linear regression. Table 5 presents these results by separately fitting a regression to summarize linear trends for each parameter (nonlinear trends were also examined, but the results showed no trends and thus are not presented); that is, each row presents the slope and intercept of a parameter estimate that is regressed on time. The coefficient of variation, which allows a comparison of residual variance between models, is also used to indicate the variability in parameter estimates. Significant estimates in the slope would indicate a linear trend for rater effects such as rater severity for the latent class SDT and the IRT models, respectively.

Results indicated that for the latent class SDT model, there was no linear increase in all five relative criteria parameter estimates as indicated by slopes that are near zero. A similar trend was found for the IRT models in that most slope estimates were not significant. The third location or step parameter for the GR and the GPC models were significant, but given the small parameter estimate of 0.1, this indicates a minor increase. The coefficient of variation, a measure of model residual, was smaller for the latent class SDT model, except for the second parameter estimate. In summary, the rater effects from the plots show drift for some raters, but as indicated by the regression that summarized

the parameters, there was no significant evidence of overall rater drift for parameters that describe rater severity.

Table 5. Regression results to summarize parameter estimates in rater effects (c_j for the latent class SDT and b_j for IRT models) over 7 administrations

Model	Parameter	Slope	Coefficient of Variation
LC-SDT	c_1	–0.002 (0.003)	0.279
	c_2	0.003 (0.002)	7.916
	c_3	–0.002 (0.002)	0.208
	c_4	0.002 (0.003)	0.106
	c_5	–0.003 (0.003)	0.079
GR	b_1	–0.336 (0.406)	0.334
	b_2	–0.057 (0.106)	0.313
	b_3	0.103 (0.049)	0.432
	b_4	0.085 (0.047)	0.342
	b_5	–0.020 (0.119)	0.308
GPC	b_1	–0.397 (0.421)	0.369
	b_2	–0.059 (0.110)	0.335
	b_3	0.105 (0.049)	0.452
	b_4	0.079 (0.048)	0.365
	b_5	–0.042 (0.125)	0.336

Note: Values in parenthesis represent standard errors. LC-SDT model refers to the latent class SDT model. Coefficient of variation represents the ratio of the root mean squared error to the mean of the parameter estimate.

Teacher Certification Test: Rater Discrimination

Mean rater discrimination. Table 6 shows the mean rater discriminations across the 7 administrations. The results show that the overall mean rater discriminations were similar for the 7 scoring occasions. Although the mean discrimination was greatest for the fifth administration, as indicated in both the latent class SDT and the IRT models, this scoring occasion also had the greatest variability of discrimination across the raters. In general, the distribution of discrimination estimates for each scoring administrations show a normal distribution of the parameters based on the skewness and kurtosis, regardless of the rater model used. That is, based on the distribution of rater discrimination parameters from the latent class SDT model, there were differences in

raters' ability to discriminate between latent classes of essays. Similarly, the IRT models also show that there were differences in raters' ability to discriminate between different qualities of essays.

Table 6. Mean rater discrimination for each administration

Model (parameter)	Administration	Mean	Variance	Skewness	Kurtosis
LC-SDT (*d*)	1	3.029	0.481	–0.369	2.884
	2	3.620	0.723	0.149	2.385
	3	3.521	1.093	0.266	2.200
	4	3.243	0.467	0.163	3.026
	5	3.985	1.708	–0.140	2.527
	6	3.244	1.001	0.313	2.270
	7	3.327	0.993	0.388	2.788
GPC (*a*)	1	4.847	1.561	–0.065	3.366
	2	4.341	1.624	0.479	2.699
	3	4.573	2.577	0.644	2.595
	4	4.524	0.770	–0.258	2.977
	5	5.103	3.100	–0.130	2.065
	6	4.697	2.210	0.554	3.050
	7	4.646	1.857	0.341	2.311
GR (*a*)	1	4.542	1.716	–0.049	3.926
	2	4.171	1.644	0.279	2.332
	3	4.338	2.600	0.612	2.740
	4	4.346	0.864	–0.339	2.632
	5	4.900	3.208	–0.266	2.253
	6	4.454	2.129	0.171	2.477
	7	4.375	1.910	0.002	2.361

Note: Formula for kurtosis used: $(m_4 - m_2^2) - 3$, where $m_i = \sum (X - \overline{X})^i / N$. LC-SDT model refers to the latent class SDT model.

Plots of rater parameters. Figures 5, 6, and 7 present plots of the discrimination parameter estimates for the latent class SDT, GR, and the GPC models, respectively. A best-fit linear line was added to the figure for each rater's discrimination to summarize the overall trend. Similar to the criteria and location parameters presented above, the latent class SDT and the IRT models showed similar trends and parameter estimates for the discrimination parameter.

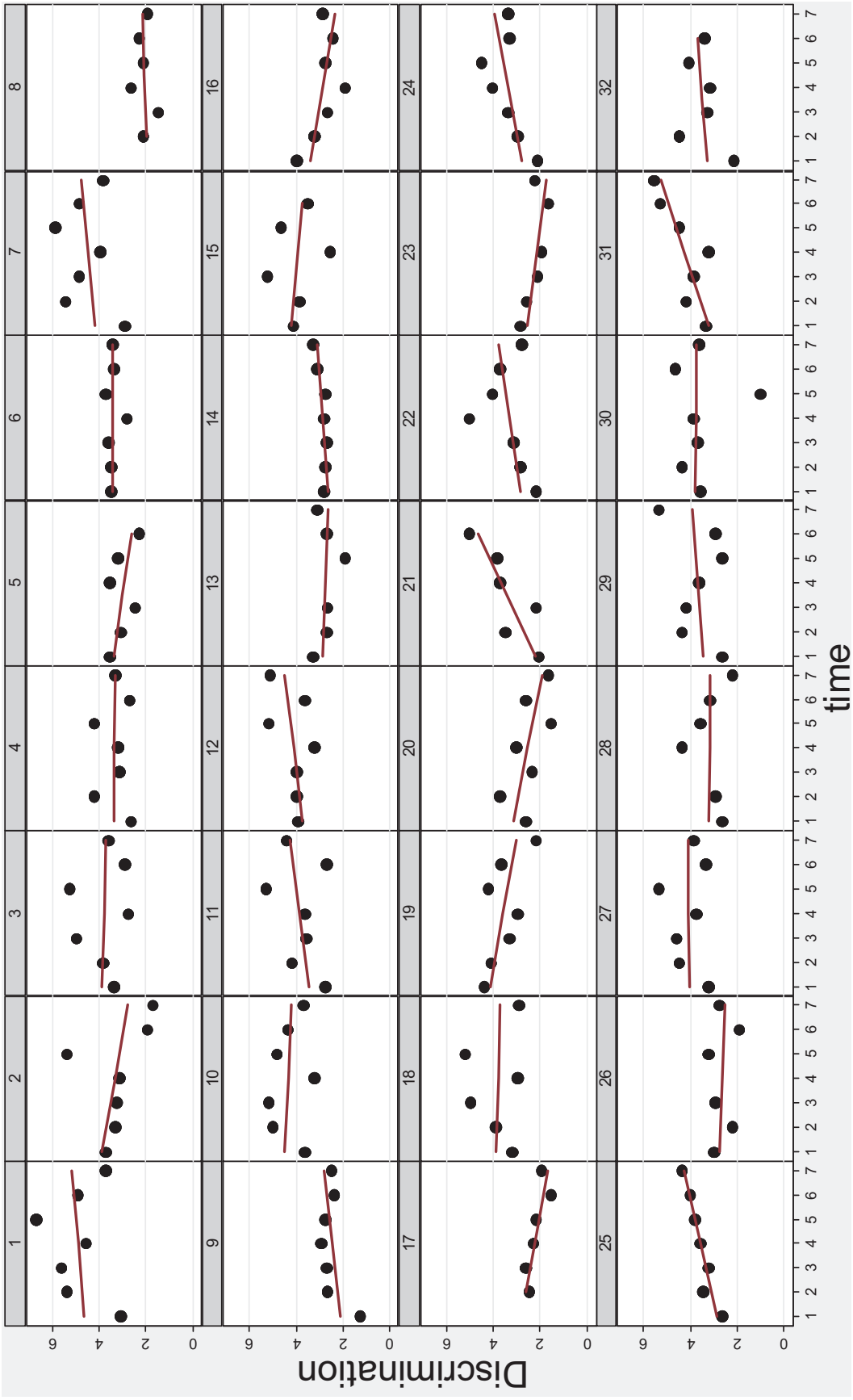


Figure 5. Teacher certification test: Plot of individual rater's discrimination (latent class SDT)
Note: Only raters that scored 6 or more administrations were included

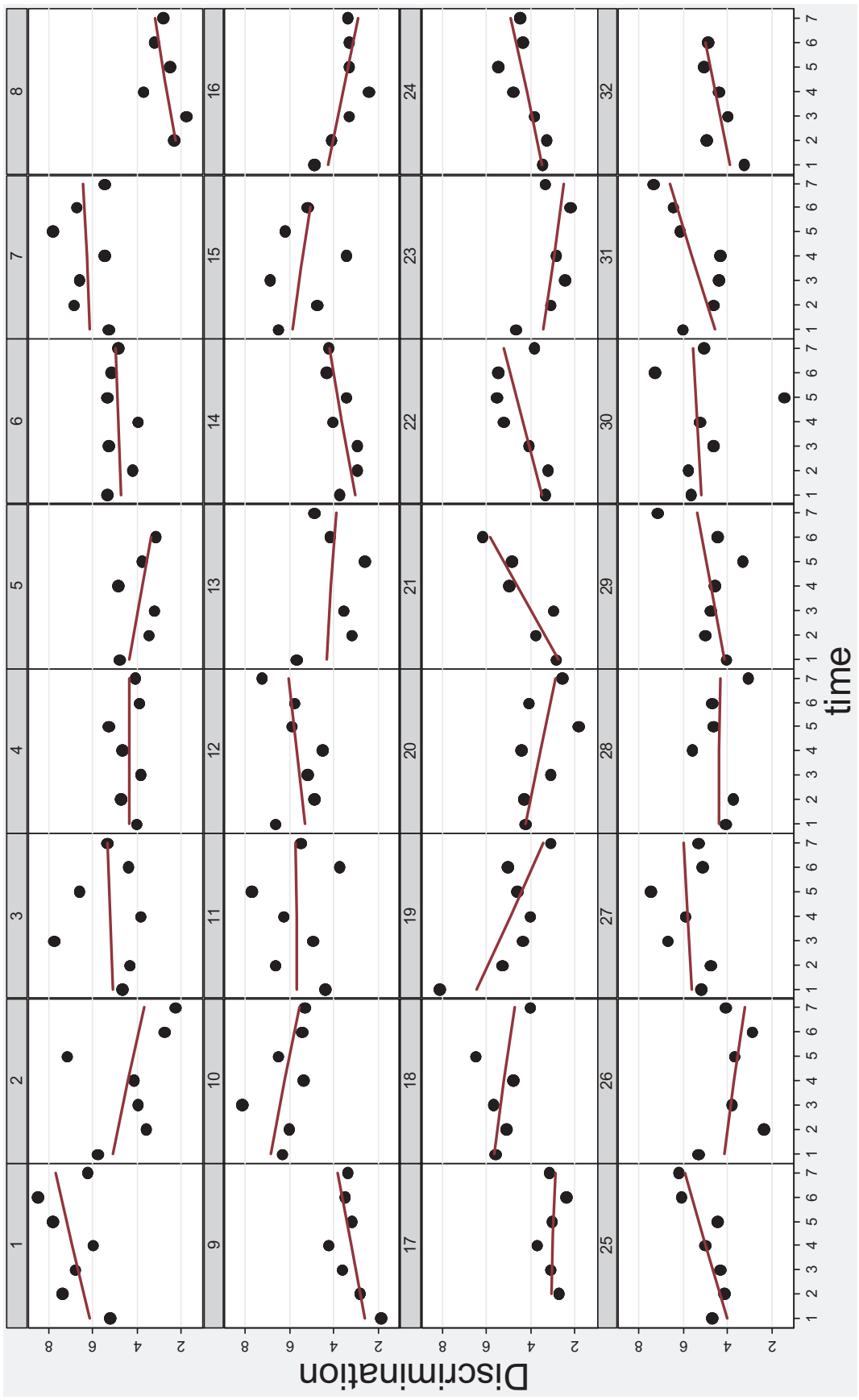


Figure 6. Teacher certification test: Plot of individual rater's discrimination (graded response model)
Note: Only raters that scored 6 or more administrations were included

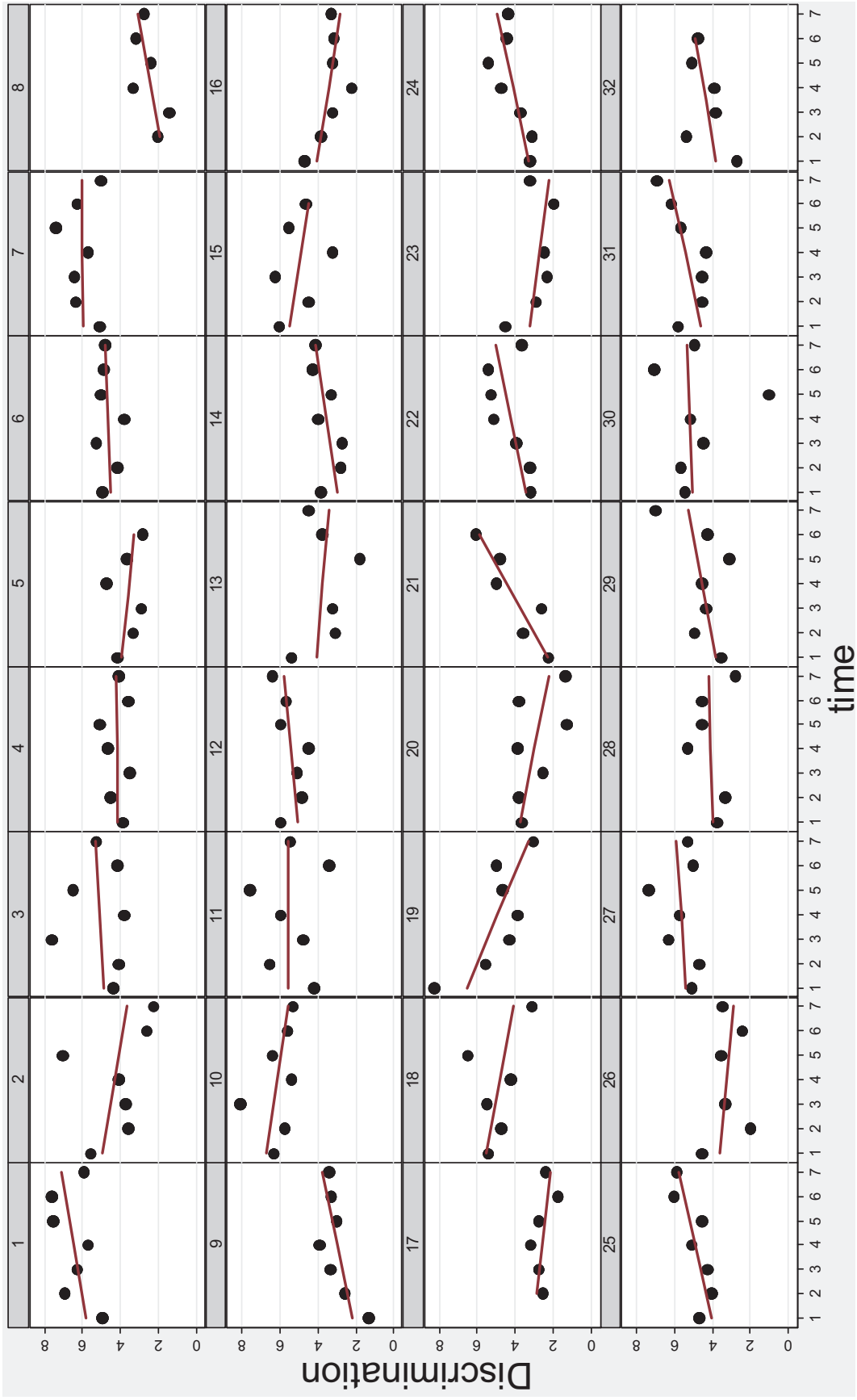


Figure 7. Teacher certification test: Plot of individual rater's discrimination (generalized partial credit model)
Note: Only raters that scored 6 or more administrations were included

For raters that showed an increase in discrimination for the latent class SDT models, the parameter estimates also increased for the IRT models. This was also the case for parameters that remained stable and for parameters that decreased. These plots also show that rater discrimination increased for some raters, while decreased for others. This shows that the level of discrimination differed between raters. For example, Rater 1's discrimination estimates increased, while Rater 17's discrimination estimates were lower and decreased between the scoring occasions.

Parameter estimates summarized using regression. Regression was used to summarize changes in the discrimination parameter estimates over time. Similar to Table 5, the slope represents the linear growth in the discrimination parameter. Based on regression slopes, there were no significant linear trends in discrimination. The coefficient of variation, which shows a measure of model residual, was about 0.3 for all three models. The results from the regression indicate that there were no significant linear trends in the three rater models.

Table 7. Regression results to summarize parameter estimates in rater discrimination (d_j for the latent class SDT and a_j for IRT models) over 7 administrations

Model	Parameter	Slope	Coefficient of Variation
LC-SDT	*d*	0.022 (0.015)	0.292
GR	*a*	0.035 (0.022)	0.303
GPC	*a*	0.030 (0.022)	0.322

Note: Values in parenthesis represent standard errors. LC-SDT model refers to the latent class SDT model. Coefficient of variation represents the ratio of the root mean squared error to the mean of the parameter estimate.

Teacher Certification Test: Latent Class Sizes and Classification Accuracy

Latent class sizes. Figure 8 presents the latent class sizes for the 7 administrations. The *X*-axis presents the six latent classes, and the *Y*-axis shows the latent

class sizes. As shown in Figure 8, for every administration, there was a concentration of scores in the fourth scoring category.

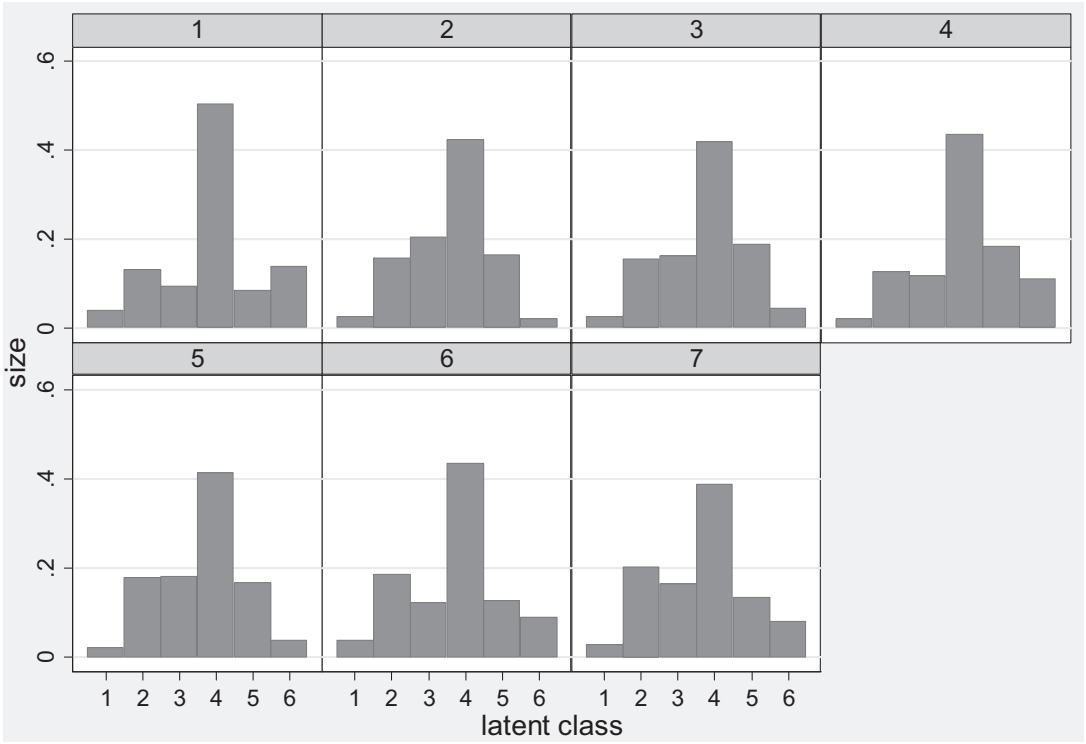


Figure 8. Teacher certification test: Histogram of latent class sizes

The large proportion of scores distributed to this category is contrasted with scores in the end categories. Based on Figure 8, the distribution of latent class sizes were similar across the seven administrations.

Classification accuracy. To examine the quality of classification, Figure 9 presents a plot of the proportion correctly classified and the lambda statistics. As noted earlier in the literature review, the latent class SDT model classifies essays into latent classes. That is, the posterior probability and the latent class sizes can be used to estimate the quality of classification. These are presented by the classification accuracy statistics. Two classification accuracy statistics are considered. The proportion correct (P_c) is the expected proportion of cases correctly classified, and the lambda (λ) is a statistic that

accounts for classification that can occur by chance, which is motivated when there is a latent class with a large class size. The plots show that classification accuracy from the P_c statistic was around 75%; there was relatively small deviation in both classification accuracy statistics.

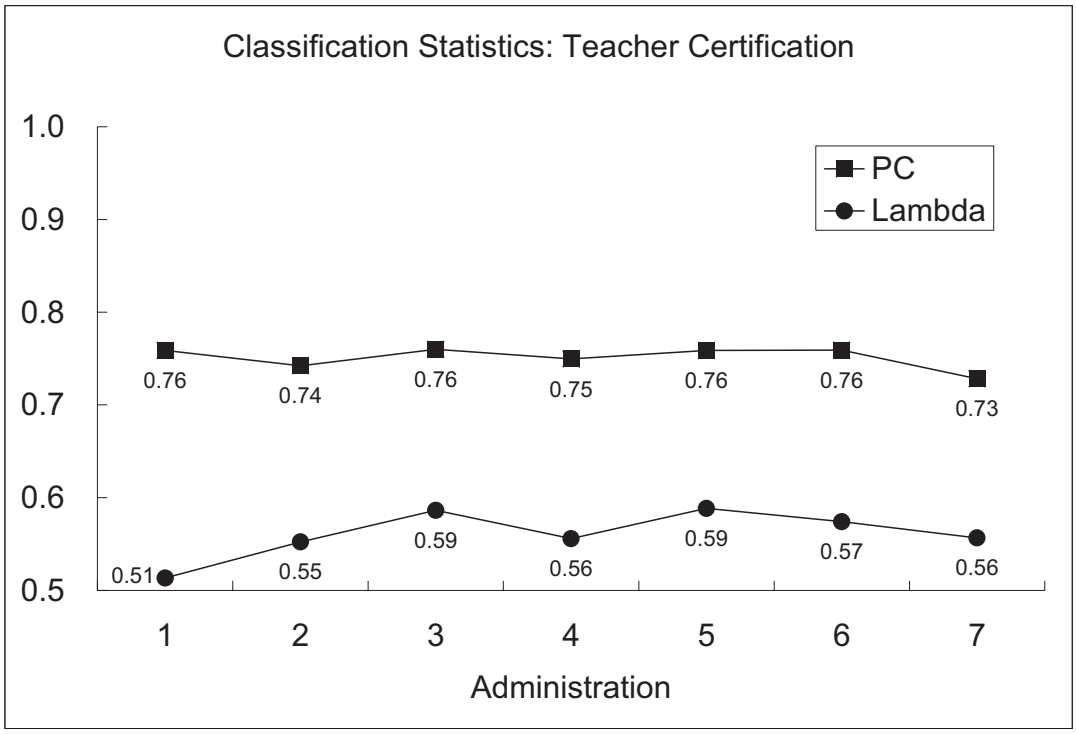


Figure 9. Teacher certification test: Classification statistics

Overall summary of the teacher certification test. The empirical results of the teacher certification exam showed that changes in classification accuracy over 7 scoring occasions were minimal. Plots showed individual variation in drift, where some raters were stricter and others lenient; there were raters that increased in discrimination, while others decreased during the seven administrations. However, the results summarized by regression indicated that overall, there was no significant linear trend in the parameter estimates. Given these results, the classification accuracy was also stable for the seven administrations. The empirical findings from this analysis indicate that although there

were variations between raters and also for the same rater, there were minimal effects on classification accuracy.

4.2 Empirical Study: High School Writing Test

The second data set used for this study comes from a national assessment of writing ability at the high school level. There were 28 raters used for this study that scored on a 1 to 4 scale, where a higher score represented greater writing mastery. This test was administered continuously throughout the year. The combined administrations were analyzed on a monthly basis making twelve points of analysis. Based on this method, 18 raters were selected that scored on 6 or more months throughout the testing year. Consequently, there were 11697, 14508, 15428, 17924, 16772, 14756, 11415, 9169, 12278, 14234, 12788, and 11320 examinees for each of the twelve consecutive months (Mean=13524, SD=2527), respectively. The presentation of the results for the high school writing test follows similarly from the teacher certification exam.

High School Writing Test: Rater Effects

Plot of rater parameters. To examine drift in rater effects, Figures 10, 11, and 12 illustrate plots of raters' relative criteria and locations for the three CR models. In Figure 10, the plots of the latent class SDT model are presented. Similar to the teacher certification test, intersection points were added to the figure to help identify reference locations. Since there were 4 distributions, the means for the locations were at 0.00, 0.33, 0.67, and 1.00; therefore, the intersection points lie at 0.17, 0.50, and 0.83. These three points provide a reference to indicate the severity of rating as well as category usage.

Relative criteria estimates above the optimal location indicate a stricter rating and an estimate below implies a more lenient rating.

Figure 10 shows that for most raters, their criteria locations for the second and the third criteria were above the optimal locations; for the first relative criteria, parameter estimates were below the intersection-point location. This meant that raters were harsher for giving higher scores and more lenient for lower scores (based on the relative reference indicated by the intersection-points). However, given that all raters had similar relative criteria locations that were above the intersection points for the second and third criteria and below the intersection point for the first criteria, this indicated consistent stringency.

The general patterns of rater severity from the latent class SDT model and the IRT models showed that raters were stable in their ratings. Raters 2, 4, 5, 7, and 14 had very stable parameter estimates; they showed minimal drift in their criteria locations. However, for raters such as 8, 17, and 18, there was drift; there were months when their criteria and locations shifted up and others when they shifted down. For raters 17 and 18, the relative criteria estimates for the first and the third parameter were lower and higher than other raters, respectively. Unlike the teacher certification test, there were fewer instances of category shrinkage, (i.e., raters avoiding to use a specific scoring category). Across the twelve months, all 18 raters used each of the four scoring categories, implying that there were no raters that systematically avoided using a scoring category. The GR and the GPC models showed plots that were nearly identical, whereas the plots of the latent class SDT and the IRT models were slightly different. For example, there were spikes in the plots of the IRT models for Rater 10 and 11, but this was not present in the latent class SDT plots. In general, these plots show that parameter estimates reflecting rater severity were stable.

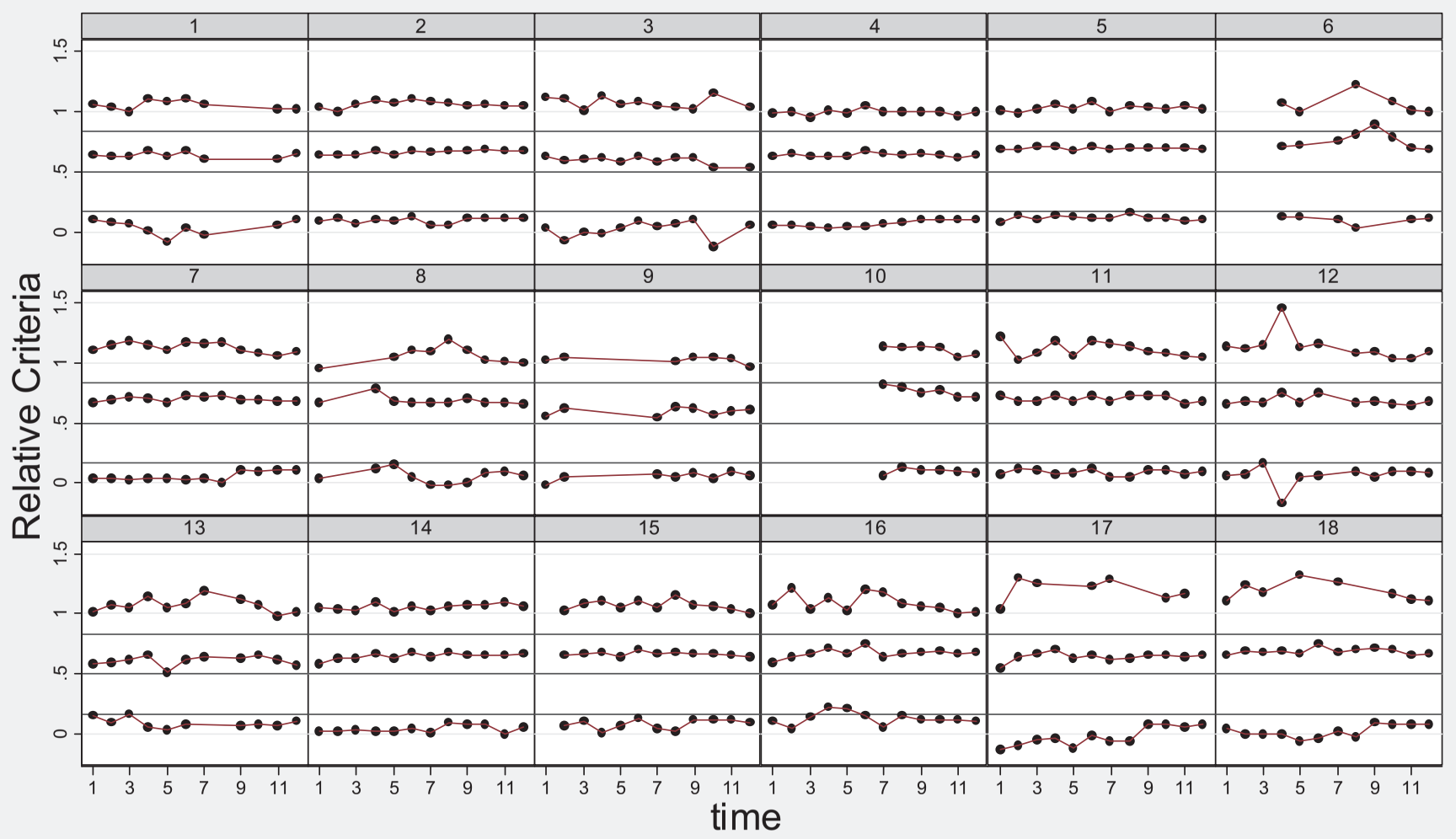


Figure 10. High school writing test: Plot of individual rater's relative criteria (latent class SDT)

Note: Only raters that scored over 6 administrations were included (10 raters eliminated). Horizontal lines represent intersection-point criteria at 0.167, 0.500, and 0.833.

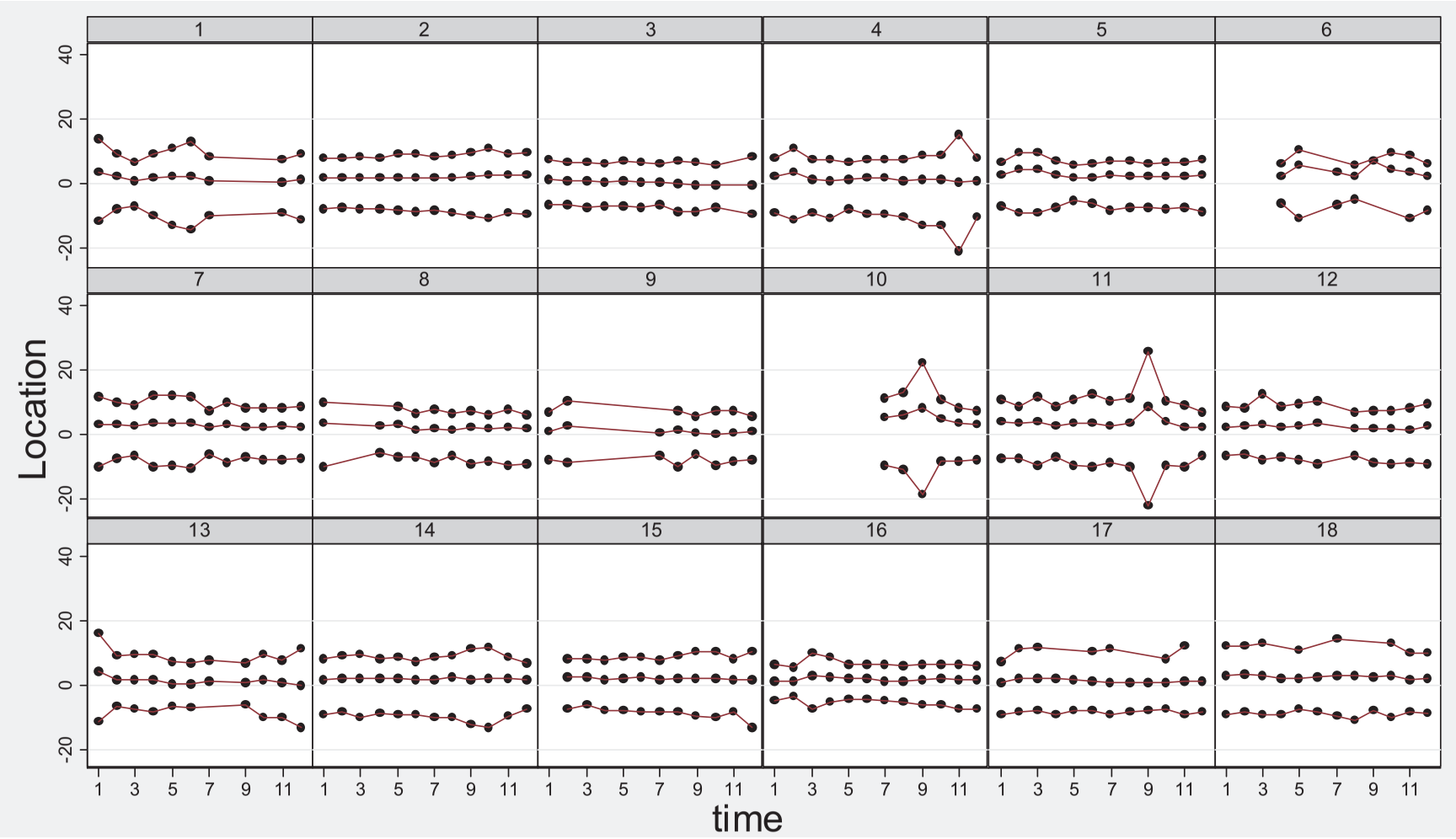


Figure 11. High school writing test: Plot of individual rater's location (graded response model)

Note: Only raters that scored over 6 administrations were included

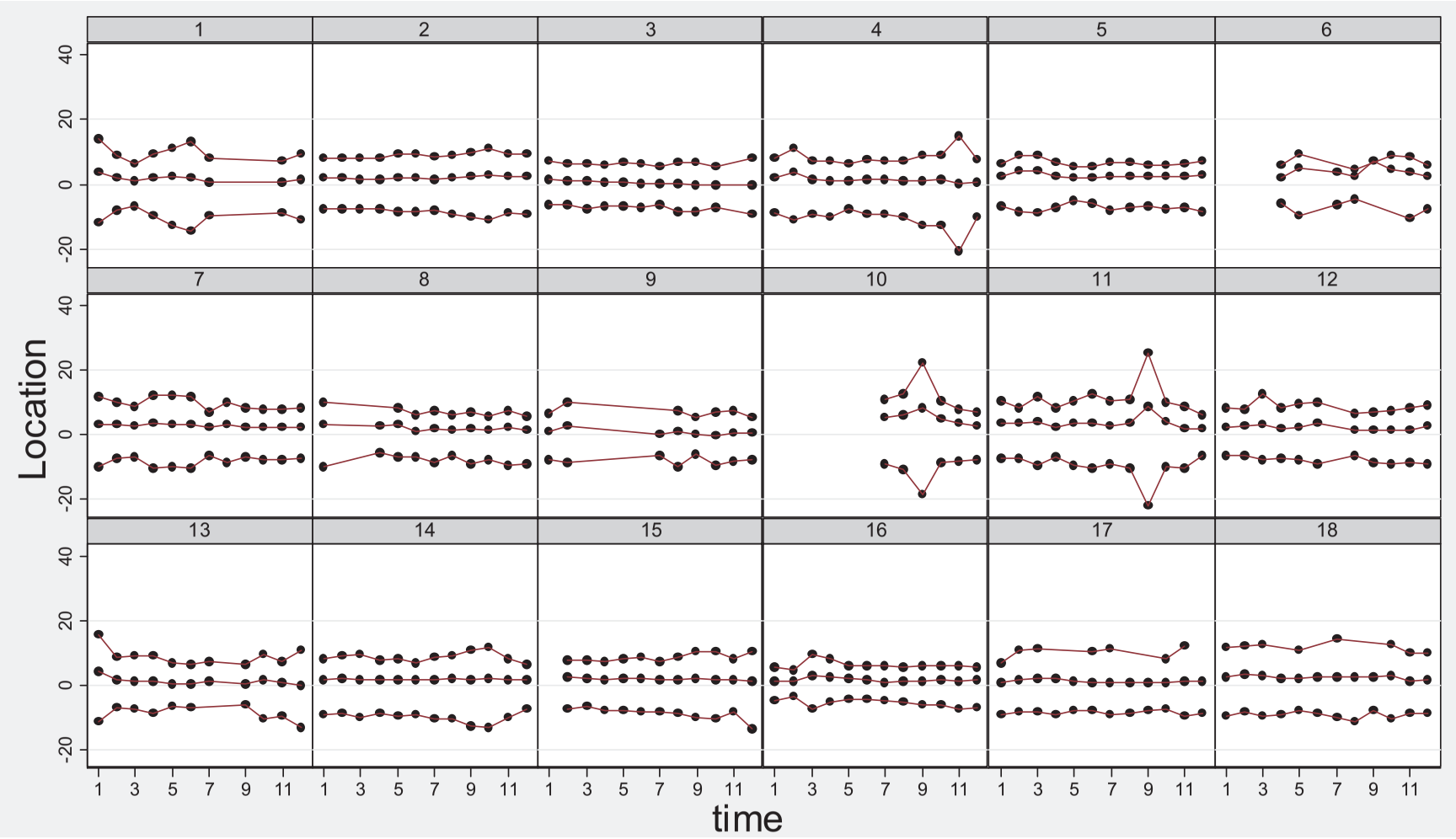


Figure 12. High school writing test: Plot of individual rater's location (generalized partial credit model)

Note: Only raters that scored over 6 administrations were included

Parameter estimates summarized using regression. Parameters from the latent class SDT model and the IRT models were summarized using regression. This presentation is similar to the results presented earlier in the teacher certification test to examine overall trends in parameter estimates that describe rater effects. Table 8 presents the changes in rater effects as a regression on time. The two IRT models produced nearly identical results. However, they differed from the latent class SDT model. The slope of the relative criteria in the latent class SDT model showed a significant increase in the first and the second locations; in the third relative criteria, there was a significant decrease in the parameters. However, these estimates in slope were very small, representing a change of less than 0.005. For the IRT models, the first location parameter showed a significant decrease on average. These differences in direction between the two models can mean contradicting interpretations. An increase in the relative criteria location in the latent class SDT model or in the location estimates for IRT models reflects a rater becoming stricter. The coefficient of variation, which shows a measure of model residual, was similar among the IRT models, but differed with the latent class SDT model.

Table 8. Regression results to summarize parameter estimates in rater effects (c_j for the latent class SDT and b_j for IRT models) over 12 months

Model	Parameter	Slope	Coefficient of Variation
LC-SDT	c_1	0.004 (0.001)	0.904
	c_2	0.002 (0.001)	0.079
	c_3	–0.004 (0.002)	0.068
GR	b_1	–0.160 (0.048)	0.126
	b_2	–0.048 (0.027)	0.460
	b_3	–0.057 (0.055)	0.203
GPC	b_1	–0.156 (0.049)	0.273
	b_2	–0.047 (0.027)	0.607
	b_3	–0.060 (0.057)	0.310

Note: Values in parenthesis represent standard errors. LC-SDT model refers to the latent class SDT model. Coefficient of variation represents the ratio of the root mean squared error to the mean of the parameter estimate.

High School Writing Test: Rater Discrimination

Mean rater discrimination for each administration. Table 9 shows the mean rater discrimination for each month, which differed between the latent class SDT model and the IRT models. The latent class SDT model showed an increase in the mean discrimination from 5.8 to 8.9 between January and December. On the other hand, the mean discrimination estimates in the IRT models remained constant throughout the year. This difference in results can be important, because the discrimination parameter in the latent class SDT model reflects the level of precision in raters; that is the ability for raters to discriminate between different classes of essays. For the latent class SDT model, there was an increase in rater precision, while the discrimination for IRT models was stable.

Table 9. Mean rater discrimination for each month

Model (parameter)	Month	Mean	Variance	Skewness	Kurtosis
LC-SDT (*d*)	1	5.826	2.593	0.249	1.696
	2	5.697	2.220	–0.290	2.825
	3	6.478	0.836	–0.169	2.138
	4	5.912	1.907	–0.682	2.494
	5	6.344	2.145	–0.056	1.955
	6	6.461	1.366	–0.796	2.912
	7	5.843	1.806	0.041	2.328
	8	6.369	2.674	–0.327	3.258
	9	9.466	5.461	–0.124	1.865
	10	9.142	6.364	–0.484	2.953
	11	9.074	4.013	0.192	1.970
	12	8.949	5.136	–0.471	2.405
GR (*a*)	1	3.802	1.657	1.775	5.642
	2	3.334	0.655	0.019	3.066
	3	3.471	0.485	–0.009	1.733
	4	3.272	0.290	0.290	2.771
	5	3.359	0.613	0.764	2.850
	6	3.246	0.717	1.156	4.321
	7	3.178	0.400	–0.475	2.605
	8	3.157	0.770	0.304	2.625
	9	3.784	4.125	1.544	4.362
	10	3.432	0.814	0.004	2.017
	11	3.615	1.700	3.077	11.953
	12	3.429	0.748	1.137	3.415

GPC (*a*)	1	3.746	1.636	1.665	5.249
	2	3.281	0.654	–0.080	3.382
	3	3.420	0.478	–0.037	1.717
	4	3.227	0.304	0.274	2.730
	5	3.281	0.556	0.390	2.004
	6	3.205	0.785	1.048	3.973
	7	3.127	0.402	–0.555	2.607
	8	3.098	0.778	0.221	2.529
	9	3.726	4.147	1.501	4.238
	10	3.366	0.853	–0.015	1.883
	11	3.560	1.653	3.055	11.900
	12	3.369	0.813	1.139	3.355

Note: Formula for kurtosis used: $(m_4 - m_2^2) - 3$, where $m_i = \sum (X - \overline{X})^i / N$. LC-SDT model refers to the latent class SDT model.

There was also a gradual increase in variance for the discrimination parameter in the latent class SDT model.

Plots of rater discrimination. To examine rater-specific trends in the discrimination parameter, Figures 13, 14, and 15 show plots of the rater discrimination over the twelve months for the latent class SDT, GR, and GPC models, respectively. Best-fit lines were added to describe the trend of the parameter estimates. For the latent class SDT model (Figure 13), nearly all raters (except rater 14) showed an increase in discrimination over time, which is consistent with the results presented in the previous section.

However, in the IRT models (Figures 14 and 15), nearly all raters showed stability. Only rater 3, 4, and 15 showed an increase in discrimination; other raters such as rater 10 exhibited a decrease in discrimination (in comparison to an increase in the latent class SDT model). The results from the rater discriminations indicate that there is a discrepancy between the CR models. Differences in these results between the two rater models are further discussed in the Discussion section of the study.

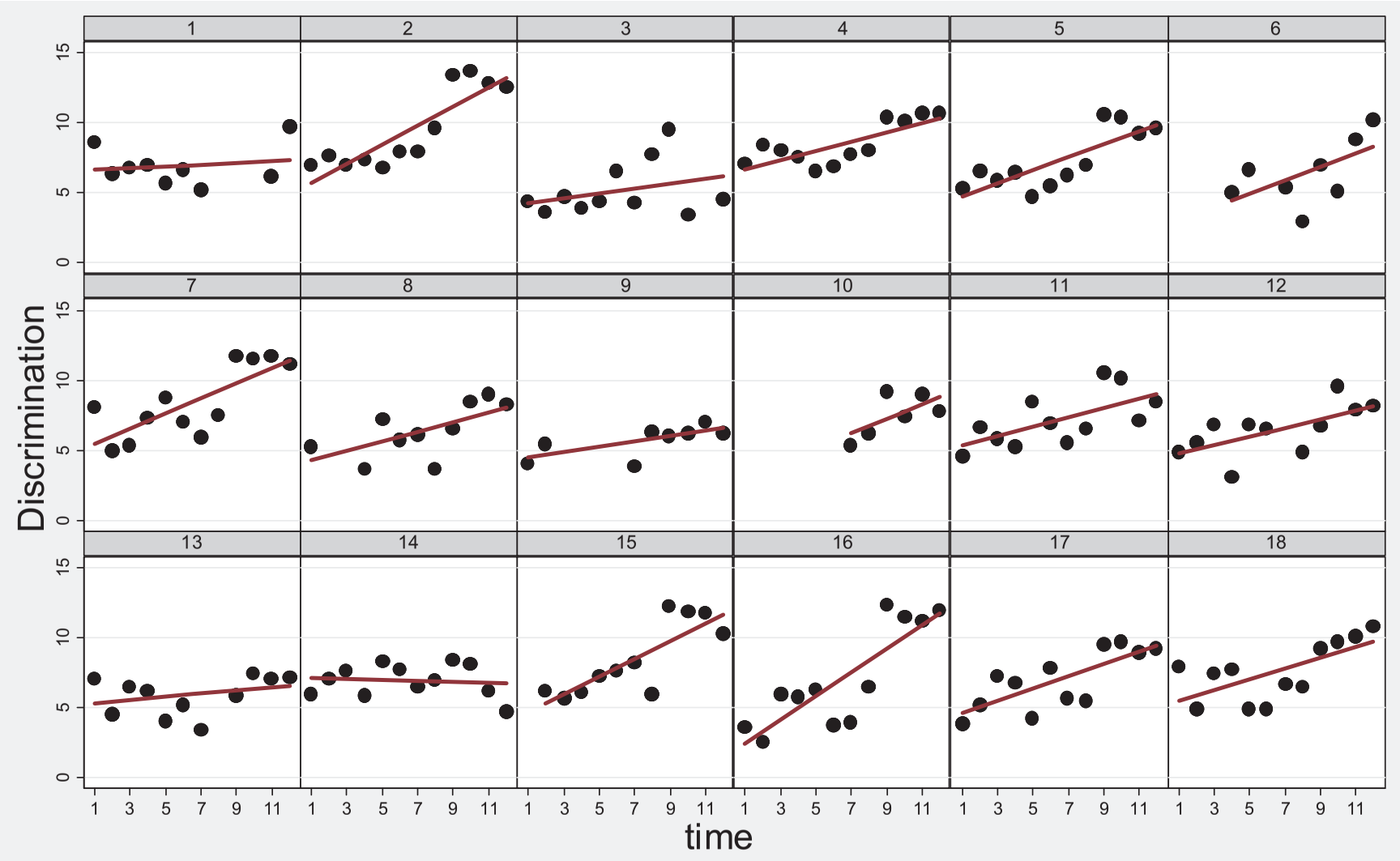


Figure 13. High school writing test: Plot of individual rater's discrimination (latent class SDT)
Note: Only raters that scored over 6 administrations were included (10 raters eliminated)

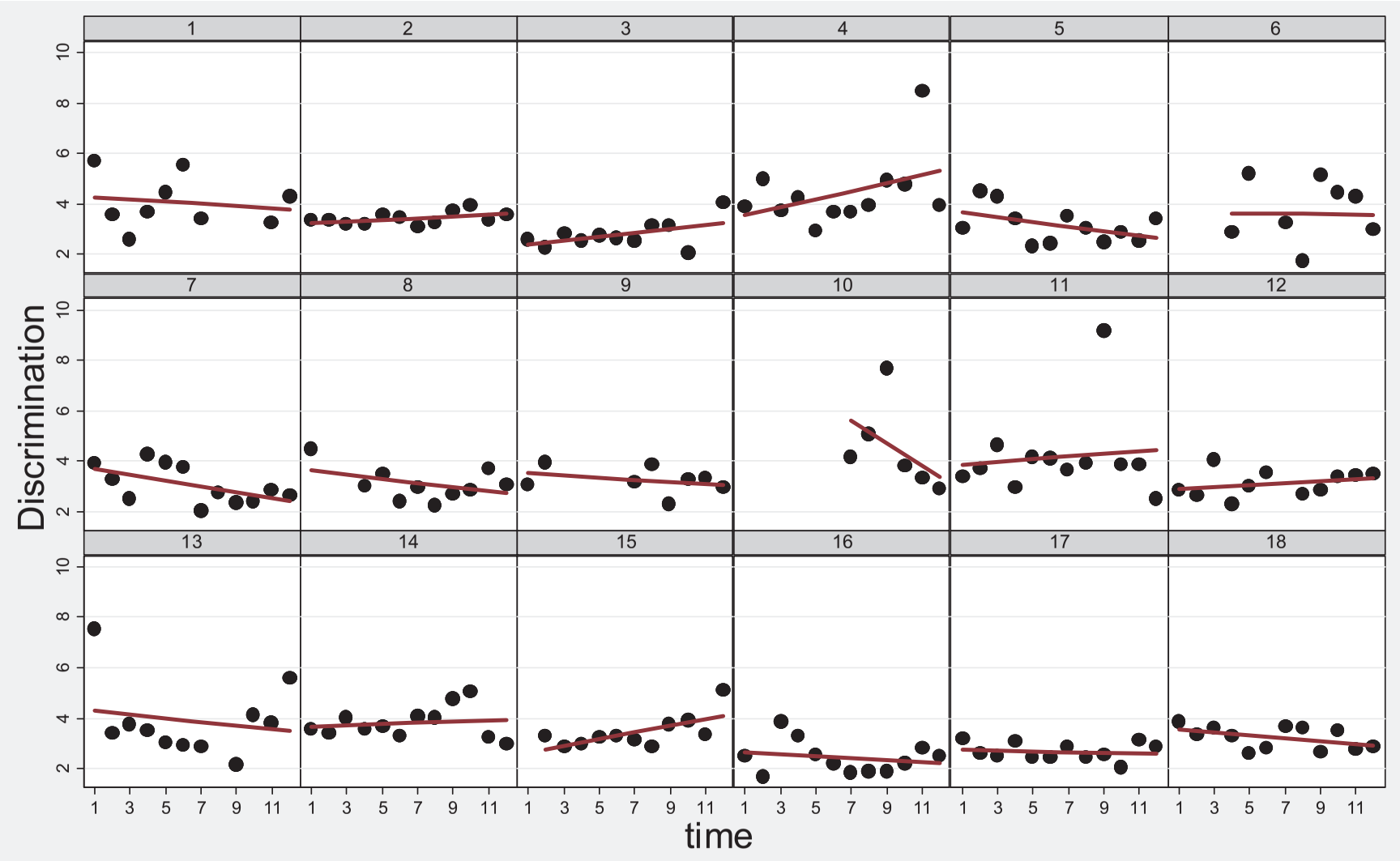


Figure 14. High school writing test: Plot of individual rater's discrimination (graded response model)
Note: Only raters that scored over 6 administrations were included (10 raters eliminated)

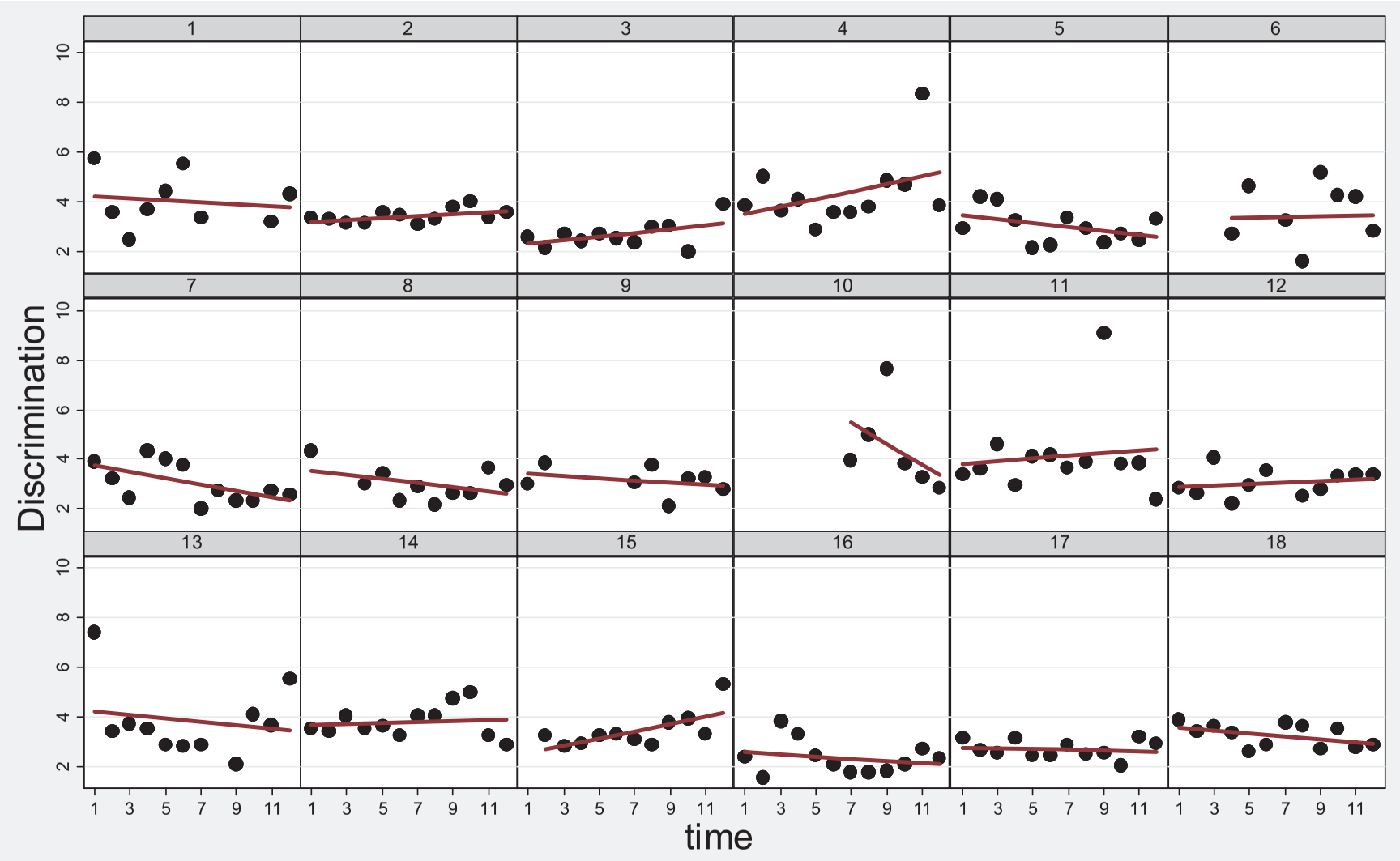


Figure 15. High school writing test: Plot of individual rater's discrimination (generalized partial credit model)
Note: Only raters that scored over 6 administrations were included (10 raters eliminated)

Parameter estimates summarized using regression. A regression was used to summarize the rater discrimination parameters. Table 10 shows these results, which indicates the discrimination parameter increased significantly for the latent class SDT model (slope = 0.356), whereas in the two IRT models, the slope parameter was close to zero. Furthermore, the coefficient of variation, which represents a measure of residual variance, was similar for the three models. As indicated from previous results, the results in this section present a contradicting picture between the latent class SDT model and the IRT models. The following section examines the latent class sizes and the classification accuracy.

Table 10. Regression results to summarize parameter estimates in rater discrimination (d_j for the latent class SDT and a_j for IRT models) over 12 months

Model	Parameter	Slope	Coefficient of Variation
LC-SDT	*d*	0.356 (0.023)	0.268
GR	*a*	0.001 (0.013)	0.308
GPC	*a*	0.001 (0.013)	0.314

Note: Values in parenthesis represent standard errors. LC-SDT model refers to the latent class SDT model. Coefficient of variation represents the ratio of the root mean squared error to the mean of the parameter estimate.

High School Writing Test: Latent Class Sizes and Classification Accuracy

Latent class sizes. Figure 16 shows the distribution of latent class sizes, which shows a highly non-normal distribution of class sizes with light tails (kurtosis of about 1 for each month). The class sizes were mostly concentrated in the second and the third latent classes. In contrast, the first and the fourth latent classes had minimal sizes below 0.03 and 0.13 at any given month, respectively. There was also a shift in the class sizes between the second and the third latent classes. For example, until May, the second latent class had the largest class size; however, this was reversed in the following months.

The histogram of latent class sizes indicates a level of non-normality in the distribution of scores. Given that there were 4 scoring categories for this assessment, the non-normality may be important as most IRT models assume a normal distribution of the latent trait.

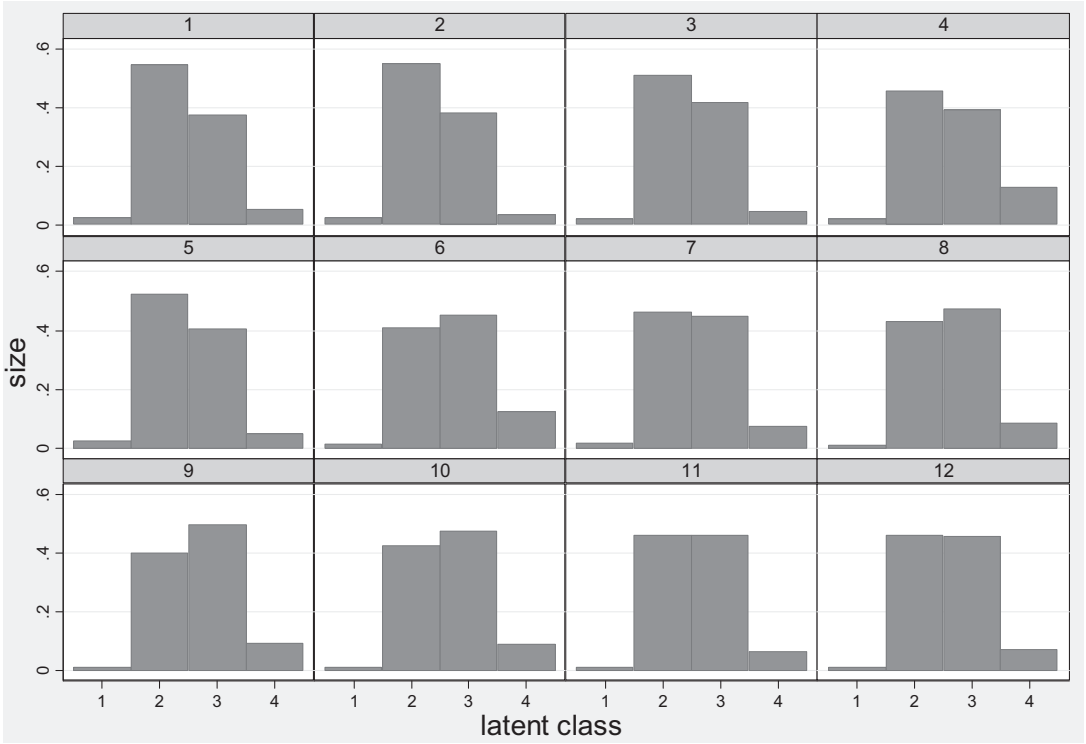


Figure 16. High school writing test: Histogram of latent class sizes

Classification accuracy. Figure 17 shows the classification accuracy statistics for each month. As presented earlier, the same classification accuracy measures – proportion correctly classified (P_c) and the lambda statistic (λ) – were used to examine the quality of classification. For P_c, there were minimal changes between the 12 months. The lambda statistic may be motivated here, because it corrects for latent classes with large sizes. Although the P_c was stable, the λ decreased nearly 0.2 points from January to September. In contrast to the teacher certification exam, the classification accuracy statistics presented here had greater deviation. On average, classification was lowest in June,

September, and in October. Given that classification accuracy statistics are derived from model parameters and the latent class sizes, these factors played a role in affecting these estimates. Further discussion of this result is provided in the Discussion section of this study.

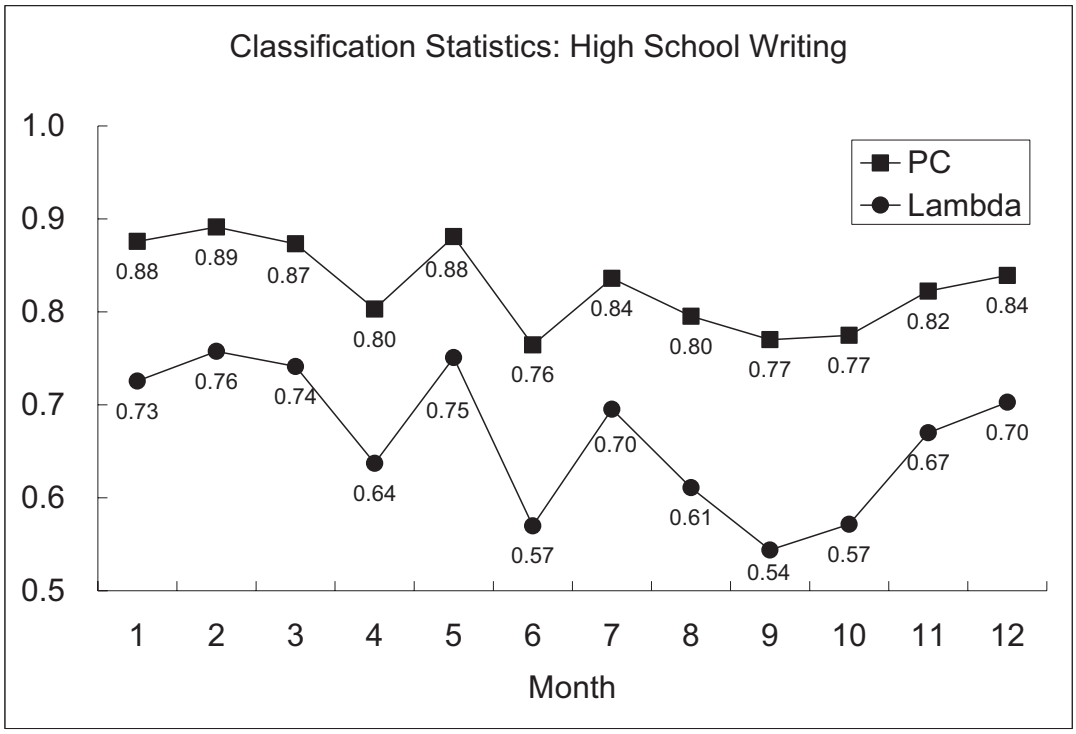


Figure 17. High school writing test: Classification statistics

Overall summary of the high school writing test. The empirical analysis of the high school writing test showed patterns of drift that differed from the teacher certification test in that the former assessment had greater indicators of overall drift. Most notably, the discrimination parameters of the latent class SDT model showed a significant increase in the parameter estimates, whereas the IRT models showed no significant trend. Moreover, unlike the rater criteria estimates, the variability of rater discrimination was over three times greater in the latent class SDT model when compared to IRT models. The estimated latent class sizes of the high school writing test had a larger concentration

in the middle-two categories. This non-normal distribution in latent class sizes differed from the more normally distributed class sizes of the teacher certification test. Overall, the high school writing test generated results that differed between the rater models that contradicted in the interpretation of rater drift.

4.3 Simulation Study 1: Examining Changes in Classification Accuracy due to Rater Drift

Simulations were conducted to examine how changes in rater behavior affect classification. Data reflecting rater drift were generated by varying population values of rater criteria and discrimination parameters from the latent class SDT model. The rater criteria parameter (c_k) indicates rater effects such as rater severity and scale usage. The rater discrimination parameter (d) reflects how well a rater discriminates between latent classes of essays. This section presents results from different conditions of rater drift and their effect on classification accuracy. Changes in population values from two parameters of the latent class SDT model (criteria and discrimination) were used to simulate rater drift, which were used to assess their effect on classification accuracy.

Classification accuracy was measured using the proportion correctly classified (P_c) statistic and the lambda (λ) statistic. This section compares estimates of classification accuracy for different conditions of rater drift. The BIB (i.e., all raters score the same number of essays) and the unbalanced (i.e., raters and pairs of raters score different number of essays) designs were used in the simulation; they were used to examine differences in classification accuracy between the two designs.

Classification Accuracy: Rater Effects

Table 11 presents classification accuracy statistics for three conditions with varying rater severity (i.e., raters are more lenient, stricter, and both lenient and strict) following population values in Table 2 (p. 33). Two incomplete designs, BIB and the unbalanced, were specified to reflect common frameworks used to conduct large-scale assessment tests. In condition 1, six raters among the ten total raters scored more leniently between the two scoring occasions by shifting their response criteria down. Condition 2 represents raters that were stricter, which was implemented by raising the response criteria of six raters. Finally in condition 3, some raters were stricter, while others were more lenient.

Table 11. Classification accuracy due to drift

Parameter	Design	Condition (2^{nd} scoring occasion)	Time 1		Time 2	
			P_C	λ	P_C	λ
Criteria	BIB	More lenient	0.843	0.789	0.823	0.762
		Stricter	0.843	0.789	0.825	0.764
		Both	0.843	0.789	0.820	0.757
	Unbalanced	More lenient	0.852	0.800	0.827	0.766
		Stricter	0.852	0.800	0.829	0.769
		Both	0.852	0.800	0.823	0.761
Discrimination	BIB	More discriminating	0.619	0.477	0.843	0.789
	Unbalanced	More discriminating	0.641	0.510	0.852	0.800

Note: Simulations ran with 100 replications. BIB refers to the balanced incomplete block design.

The three conditions were repeated for the BIB and the unbalanced design for ten rater pairs (as specified in Table 1, p. 18). The results for both designs were similar. As shown in Table 11, between time 1 and time 2, classification accuracy (P_C) changed from 84% to 82% for the BIB design for all three conditions, whereas for the unbalanced design, it changed from 85% to about 83%. These results indicate that a shift in raters' response criteria for all three conditions had minimal impact on classification accuracy. That is,

changes in rater's leniency, strictness, or both have only a small effect on classification accuracy. This result has implications for rater training and feedback, which are discussed in greater detail in the next chapter. In short, these results show that changes in rater severity had only a small effect on classification accuracy.

Classification Accuracy: Rater Discrimination

Table 11 (p. 70) also shows the results of classification accuracy for changes in rater discrimination following the specification in Table 3 (p. 34). In the first scoring occasion, raters' discrimination was normally distributed with a mean of 2; in the second scoring occasion, raters' discrimination increased to a mean of 4. As such, rater discrimination increased by two points (except for rater 1's discrimination, which increased by 1.5 points). This was simulated for both the BIB and the unbalanced designs for 10 rater pairs.

Results showed that for the BIB and the unbalanced conditions, both classification statistics P_c and λ increased from about 0.6 to about 0.8. These conditions show a contrast with classification accuracy resulting from changes in rater severity, which had minimal effects on classification. In sum, the findings from these simulation results show that classification accuracy is largely driven by changes in rater discrimination, rather than shifts in rater effects such as rater severity. This again has implications for rater training. These issues are also discussed in greater detail in the next chapter.

4.4 Simulation Study 2: Detecting Drift using Rater Models

The simulations in this section present results that examine how well IRT models detect drift when data were generated using the latent class SDT model. In the first part of

this simulation study, conditions specified in Table 2 (p. 33) were fit using the GR model. This was done to examine the effect on parameter estimates (b_k) reflecting rater severity. Condition from Table 3 (p. 34) that increased rater discrimination was also fit using the GR model. The combination of these conditions together indicate whether IRT models such as the GR model can detect drift in either rater effects and in rater discrimination, if the data are generated according to the latent class SDT model.

The second part of this study examined whether the distribution of the latent classes affects IRT parameters when data are generated using the latent class SDT model. Conditions for this simulation follow from Table 4 (p.36), where the first scoring occasion has normally distributed latent class sizes. In the second scoring occasion, there was a concentration in the third and in the fourth latent classes. The simulation study changed latent class sizes to examine how the normality of data affect discrimination estimates in the latent class SDT and in the IRT models. For all simulations in this section, the BIB design was used to allow a balanced number of essays to be scored by each rater. Results of the mean parameter estimates and standard errors are presented in Appendix A.

Detecting Drift using the GR model

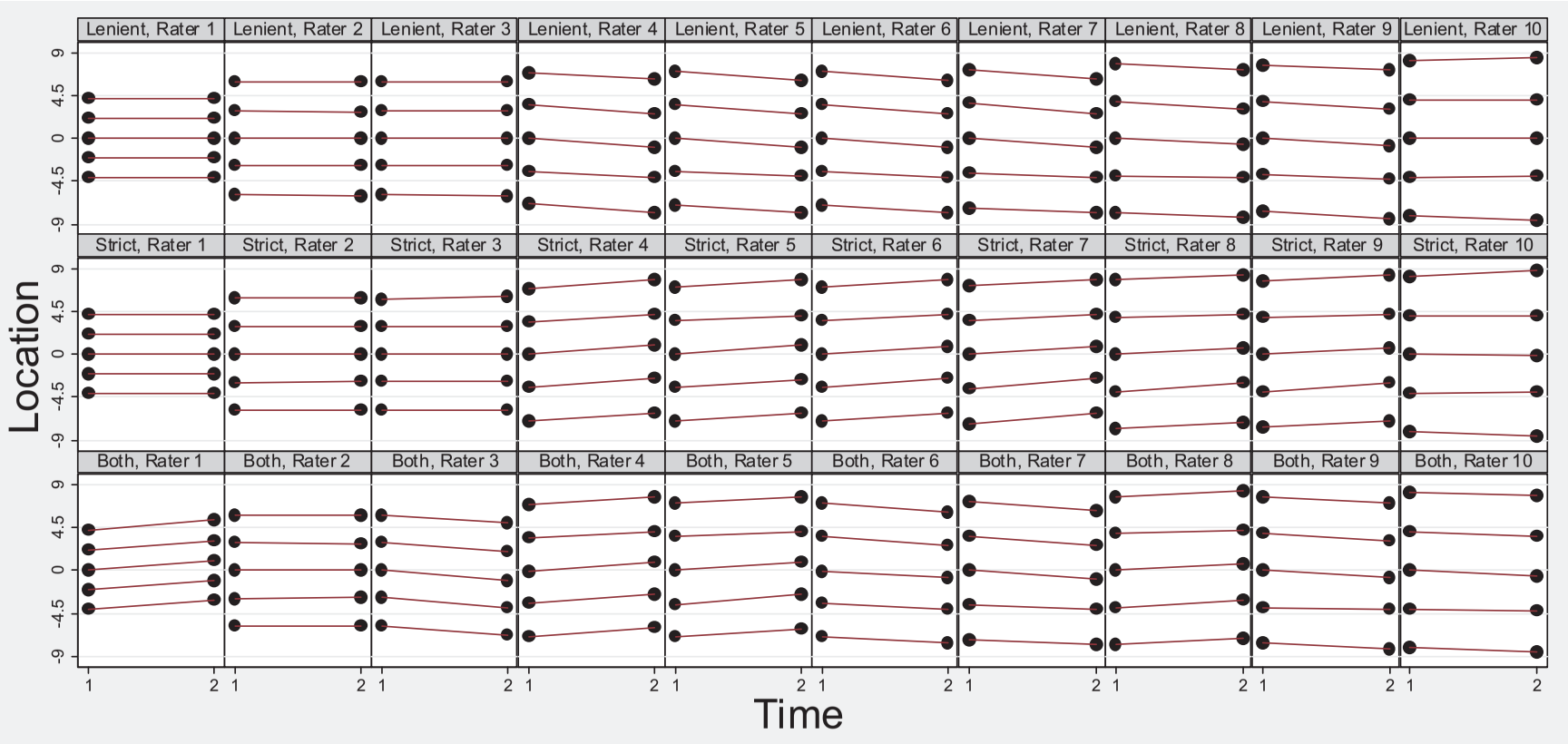
This section presents the results of IRT parameters when data were generated reflecting drift in rater criteria and in rater discrimination using the latent class SDT model.

Drift in rater effects. Table A1 (see Appendix) shows the results of the three conditions that reflect raters that are more lenient, stricter, and both lenient and strict between two scoring occasions. The first block shows the mean parameter estimates and the mean standard errors of the first scoring occasion, where raters had discrimination

population values that were normally distributed with a mean of 4 and criteria population values at the mid-point locations. The following blocks show the mean parameter estimates and standard errors of each condition – reflecting raters that were more lenient, stricter, and both lenient and strict.

Figure 18 graphically illustrates the mean parameter estimates in Table A1 (see Appendix) to present the effect of parameter changes between the scoring occasions. The *X*-axis represents the two scoring occasions, and the *Y*-axis represents the location (threshold) parameter estimated using the GR model. Similar to the presentation in Table A1, the first row presents the mean parameter estimates for the condition representing raters becoming more lenient; raters 4 to 9 had parameter estimates that shifted down to demonstrate this effect. The second row presents the case where raters were stricter; raters 4 to 9 had parameter estimates that shifted up to reflect stricter ratings. The third row presents both leniency and strictness for the raters; for this case, raters 3, 6, 7, 8, and 10 had parameters that shifted up to show strictness in rating, while raters 1, 4, 5, and 8 had parameter estimates that shifted down to reflect leniency.

In general, the results showed that the GR model was able to detect drift in rater severity. For the raters that were more lenient, the location parameters shifted down. This was shown for raters 4 to 9 that had their criteria shifted down by 1 point in the generating values. This was also found in the condition where raters were stricter and also in the condition where raters were both stricter and more lenient. Although a 1-point increase or decrease in the generating value of the latent class SDT model did not necessarily result in a 1-point difference in the estimates, an overall shift was present in the GR model parameter estimates.



Note: Row 1: Raters 4 to 9 (more lenient) downward shift in GR item location (threshold) parameters
Row 2: Raters 4 to 9 (stricter) upward shift in GR item location (threshold) parameters
Row 3: Raters 3, 6, 7, 9, and 10 (stricter); Raters 1, 4, 5, and 8 (lenient)

Figure 18. Plots of individual rater's item locations using the GR model: Lenient, strict, and both conditions

Drift in discrimination. Table A2 (see Appendix) presents the mean parameter and standard error estimates of the GR model when there was an increase in discrimination generated from the latent class SDT model. The conditions used for generating the two scoring occasions were taken from Table 3 (p.34), where raters' discrimination was normally distributed with a mean of 2 at the first scoring occasion and increased to a mean of 4 in the second scoring occasion.

The mean parameter estimates indicate that the GR model was able to detect drift resulting from changes in rater effects and in rater discrimination. That is, when data were fit for the two scoring occasions representing an increase in mean discrimination, estimates of discrimination from the GR model increased. This indicates that the GR model was able to detect changes in rater discrimination from the latent class SDT model. Although the population values had a two-point increase in discrimination (*d*), the GR discrimination parameters (*a*) did not necessarily increase by two units. For raters with lower discrimination (*d*) population values (raters 1, 2, and 3), their discrimination (*a*) increased by more than two points; for raters with higher discrimination (*d*) population values (raters 8, 9 and 10), their discrimination (*a*) increased by less than two points. For raters with discrimination (*d*) population values of 2 and 4 (raters 4, 5, 6, and 7), their discrimination (*a*) increased by about two points. Because mid-point criteria values were used to generate data, the GR location parameters also shifted as discrimination increased.

Effect on IRT parameters for Normal and Non-Normal Class Sizes

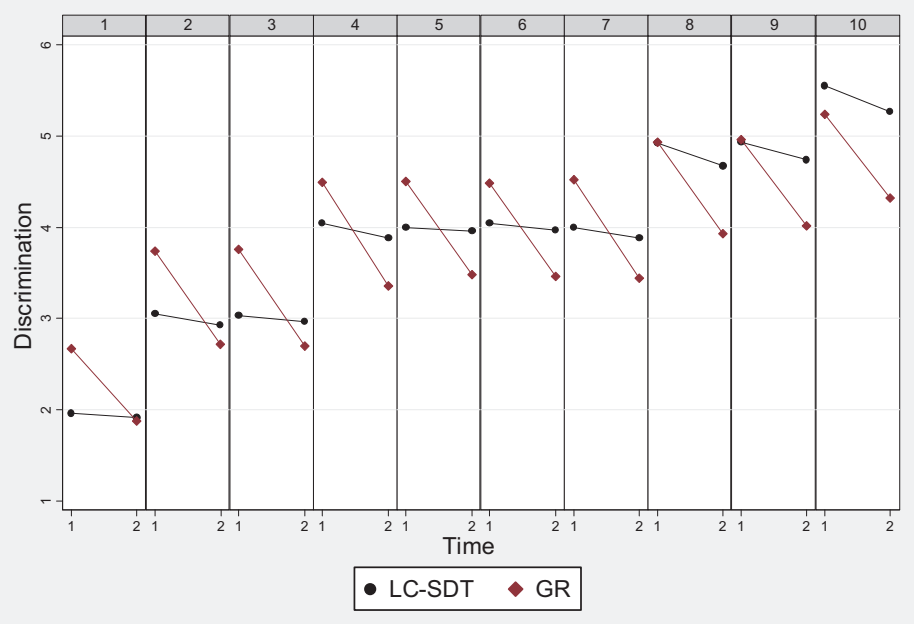
This section presents results of GR parameter estimates when class sizes changed between two scoring occasions. The data used for this simulation were generated using the latent class SDT model. Conditions described in Table 4 (p. 36) were used. In the first

condition using 6 categories, the first scoring occasion used latent class sizes that were normally distributed; in the second scoring occasion, the latent class sizes were non-normal in that there was a greater concentration of class size in the third and fourth latent classes. The BIB design was used in this simulation.

Normal and non-normal latent class sizes for 6 scoring categories. In this section, simulations were conducted to examine the effect of parameter estimates when the distribution of latent class sizes was non-normal. This is presented in Table A3 (first scoring occasion) and in Table A4 (second scoring occasion). The left column shows the generating conditions, and the column to the right shows the mean parameter estimates and standard errors for the latent class SDT model and the GR model. In general, when latent class sizes were non-normal, discrimination parameter (*a*) estimates in the GR model were underestimated.

The parameters for the latent class SDT model closely resemble the population values; that is, the parameters were recovered well with low bias. The latent class sizes were also recovered well for this condition. Table A4 (see Appendix) shows the results of the same condition presented in Table A3, but only changing the latent class sizes. Between the two conditions, there was a small decrease in the discrimination parameters for the latent class SDT model. For rater 10 that had the highest rater discrimination (*d*=6), discrimination decreased by 0.28 points; for other raters, their discrimination on average decreased by 0.13 points. However, taking into account the range of estimates using the standard error, the differences between the two scoring occasions were not significant. This showed that changing the normality of the latent class sizes did not have a significant effect on affecting parameter estimates for the latent class SDT model.

This result contrasts with the GR model, where the average decrease in rater discrimination was about 1.00 point. Figure 19 illustrates the mean change in discrimination parameters between the two scoring occasions. Here, the *X*-axis represents the two scoring occasions, and the *Y*-axis represents the discrimination parameter (*d* for the latent class SDT model and *a* for the GR model).



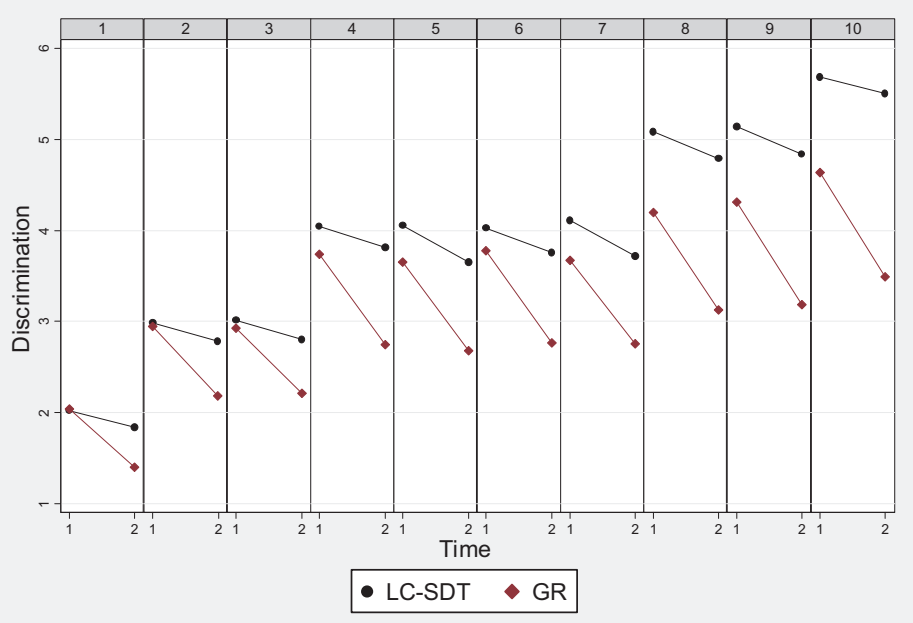
Note: The *X*-axis represents the two scoring occasions. The *Y*-axis represents *d* for the latent class SDT model and a for the GR model.

Figure 19. Plots of the discrimination parameters for the latent class SDT model and the GR model with 6 scoring categories (Condition 1: Change in normality of latent class sizes: non-normal condition for the second scoring occasion)

For the criteria parameters, the second and third criteria that had the largest latent class sizes also decreased by 0.41 and 0.45 points on average, respectively, for the latent class SDT model. In contrast, the second and the third location parameters of the GR model decreased by 1.51 and 0.40 points on average, respectively.

Normal and non-normal latent class sizes for 4 scoring categories. The condition examined in this section follows from the previous simulation; the effect of

non-normal distribution in latent class sizes on parameter estimates were investigated for 4 scoring categories. Tables A5 and A6 show the results for the 4 category condition. This is graphically illustrated in Figure 20, where the *X*-axis presents the two scoring occasions, and the *Y*-axis shows the discrimination parameter estimates. In the condition with 4 scoring categories, the mean discrimination parameter estimates for both latent class SDT model and the GR model decreased. However, similar to the condition using 6 categories, the decrease in parameter estimates from the latent class SDT model was not significant; moreover, the level of decrease was greater in the GR model. For the latent class SDT model, the mean decrease in parameter estimate was 0.3 points. The decrease for the GR model was 0.9 on average (see Figure 20).



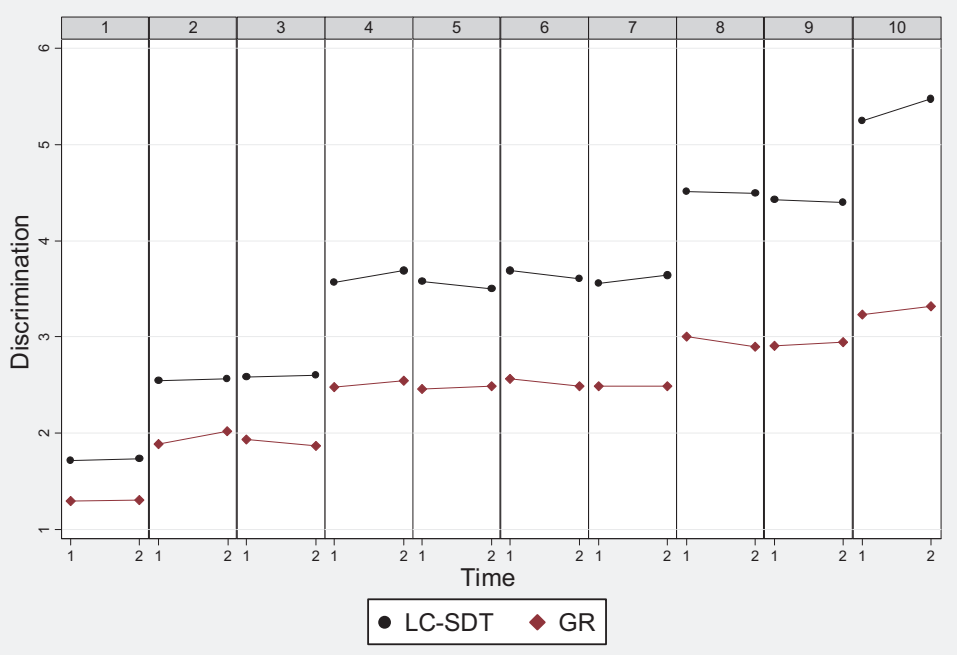
Note: The *X*-axis represents the two scoring occasions. The *Y*-axis represents *d* for the latent class SDT model and a for the GR model.

Figure 20. Plots of the discrimination parameters for the latent class SDT model and the GR model with 4 scoring categories (Condition 2: Change in normality of latent class sizes: non-normal condition for the second scoring occasion)

A notable difference from the 6 category condition was the large bias in class sizes from the latent class SDT model. The class size estimates were overestimated with estimates of

0.106, 0.397, 0.393, and 0.104 for the four classes, respectively (the population values were 0.07, 0.33, 0.33, and 0.07).

Shift in latent class sizes. The results of the third condition, which shifted the densities of the latent classes –from 0.07, 0.50, 0.40, and 0.03 for the four classes to 0.03, 0.40, 0.50, and 0.07, respectively – are presented in Tables A7 and A8 (see Appendix). Similar to the others, the only condition that changed in the generating values was the latent class sizes, not the parameter values. Figure 21 shows a graphical representation of the changes in discrimination parameters for the two models between the two scoring occasions. For this condition, the mean discrimination parameter estimates for the latent class SDT model and the GR model seemed to be stable between the two scoring occasions.



Note: The *X*-axis represents the two scoring occasions. The *Y*-axis represents *d* for the latent class SDT model and a for the GR model.

Figure 21. Plots of the discrimination parameters for the latent class SDT model and the GR model with 4 scoring categories showing shift in density (Condition 3: Shift in density between the scoring occasions)

Unlike previous conditions, where non-normality affected discrimination estimates in both the latent class SDT and the GR models, a shift in latent class sizes affected the criteria estimates. The criteria and location parameters were affected by the shift in class sizes. The mean criteria parameter estimates (c_k) for the three locations changed from 1.15, 5.54, and 10.02 for the first scoring occasion to 0.69, 5.20, and 9.61 for the second scoring occasion. This presents a shift down in the criteria parameters. Likewise, for the GR parameters, there was also a downward shift in the threshold parameters (b_k); for the first scoring occasion, the mean parameter estimates for the ten raters were –3.65, 0.46, and 4.54, while for the second scoring occasion, it changed to –4.53, –0.46, and 3.65, respectively. The downward shift in the parameters, which shows leniency among raters, is consistent with the shift in the latent class sizes that increased the proportion of scores in the lower categories.

Similar to the non-normal condition with 4 categories in the previous section, the latent class sizes were overestimated. The recovery of latent class sizes were 0.109, 0.446, 0.340, and 0.106 for the first scoring occasion, and 0.096, 0.346, 0.446, and 0.111 for the second scoring occasion, respectively.

Summary. The results from this simulation study shows that changes in latent class sizes can affect the discrimination parameter of the GR model. For the latent class SDT model, the generating parameters were well recovered, with a small decrease in parameter estimates. However, for the GR model, the difference in latent class sizes shifted discrimination by nearly 1 point. Furthermore, for the last condition that shifted the latent class sizes, the discrimination parameters were not affected; rather, they shifted down the estimates for the criteria parameter for the latent class SDT model and the

location parameter for the GR model. The effect of shift in class sizes is consistent with the interpretation of the rater parameters reflecting rater severity; that is, the shift in density indicated greater class size for the 4th category, meaning more lenient scores as reflected in the criteria (c_k) and in the threshold (b_k) parameter estimates.

4.5 Parameter Recovery: Rater Parameters, Latent Class Sizes, and Standard Errors from the Latent Class SDT model.

This section presents results for the recovery of rater parameters and latent class sizes for the simulated data discussed above. Simulated data were generated using the latent class SDT model. Appendix B presents the population value, mean estimate, bias, percent bias, and mean squared error (MSE) of the parameters.

Estimates of standard errors of rater discrimination and latent class sizes were also evaluated; asymptotic theory was used by examining the inverse of the observed information matrix (for details see Vermunt & Magidson, 2005; DeCarlo, 2010). Bias was calculated by taking the difference of the standard deviation of the parameter estimates across the 100 replications to the mean of the estimated standard errors. Appendix C presents the standard deviation, mean standard error, bias, and the percent bias for conditions used in the simulation.

As the focus of the simulation study was on examining the effect of rater drift on classification accuracy and on parameter estimates derived from the GR model, a detailed account of results from the parameter recovery is not presented – as these results were consistent with findings from earlier work conducted in DeCarlo (2008) and in DeCarlo

(2010). The following sections present key findings and summaries of the recovery in parameter estimates and standard errors.

Rater Parameters and Latent Class Sizes

Table B1 presents the parameter estimates, bias, percent bias, and MSE for 10 raters, where rater discrimination was distributed with a mean population value of 4 (condition from the second scoring occasion in Table 3, p. 34). The condition used in Table B1 represents mid-point criteria locations as population values, which reflect raters that do not exhibit rater effects such as severity or scale shrinkage. Tables B2, B3, and B4 show the results when population values were specified for more lenient, stricter, and both lenient and strict raters, respectively, following conditions presented in Table 2 (p. 33). These conditions were specified for the BIB design, where all 10 raters score the same number of essays. Tables B5 to B8 replicates the same conditions using the unbalanced design, where raters and pairs of raters score different number of essays.

Tables B9 and B10 show the results for rater discrimination normally distributed with population mean value of 2 for the BIB and unbalanced designs, respectively. Table B11 presents the results when latent class sizes were non-normally distributed; that is, population values were specified to create a concentration of class sizes in the middle classes 3 and 4.

Tables B12 to B15 replicate similar conditions using 4 scoring categories. In Table B12, discrimination was normally distributed with mean population value of 4; in Table B13, a non-normal distribution was used to generate data with a concentration in classes 2 and 3. Tables B14 and B15 show the results for a shift in latent class sizes

where there was a larger class size for classes 1 and 2 in Table B14 and a larger class size for classes 3 and 4 in Table B15.

In general, parameters were underestimated for the first criteria estimate (e.g., c_{11}, c_{21}, … , c_{101}) with percent bias that was higher than other parameters. For example, in Table B1, the percent bias ranged between 10% to 25% for the first criteria estimate; the remaining parameters had percent bias that was less than 5%. MSE was greater for the fourth and the fifth criteria estimates. Raters with higher population values of discrimination also had greater percent bias and also MSE. This trend was consistent for all conditions. For conditions where raters were stricter (Tables B2 and B6), most parameters were underestimated with a greater percent bias for the first category ranging between 27% and 57% (for the rater with the highest discrimination). When raters were lenient (Tables B3 and B7), both percent bias and MSE were smaller than Table B1. There was a mixed result for the condition with raters exhibiting both leniency and strictness (Table B4 and B8); that is, raters had greater percent bias when they were stricter for the first criteria estimate.

The percent bias in latent class sizes were over 10% for the end categories (i.e., 1 and 6). For conditions where raters were stricter, the percent bias for the first category was 35%; likewise, for the condition where raters were lenient, the percent bias was over 36% for the last category. These differences in latent class sizes show that when raters are stricter, the class sizes increase for the lower scoring categories; when raters are lenient, the class sizes increase for the higher categories.

The difference between BIB and unbalanced designs were the inflated percent bias for raters that score a smaller number of essays. In the unbalanced design, raters 2

and 5 only score 50 and 60 essays, respectively. Their percent bias estimates were consistently higher for all conditions. For example, comparing Table B1 and B5, which reflect conditions without rater severity, the percent bias for rater 2 and 5 were less than 2.5% (excluding the first criteria); however, they were over 6% for rater 2 and over 10% for rater 5 in the unbalanced design. These results are consistent with findings from DeCarlo (2010).

Tables B9 and B10 show results for normally distributed discrimination with mean of 2 for the BIB and the unbalanced designs. In general, these results have similar findings as discrimination distributed with mean of 4 in that the first criteria estimates had consistently higher percent bias than other criteria estimates. However, when compared to Table B1 that had discrimination distributed at mean 4, the percent bias were higher. The percent bias in the latent class sizes was also higher; the end categories had percent bias over 50%. The inflation of percent bias for raters 2 and 5 were also found in the unbalanced design.

When discrimination was distributed at mean value of 2 (Tables B9 and B10), percent bias was greater for all parameters than when discrimination was distributed at mean value of 4 (Table B1). Although raters with higher discrimination also had greater bias, the percent bias for the first criteria was over 38% in the BIB design; it was greater for raters 2 and 5 that scored less in the unbalanced design. Furthermore, the percent bias in latent class sizes were over 50% for the end categories.

Tables B12 to B16 present the results when 4 categories were used. In general, similar patterns were found from conditions that used the 6 categories. Raters with higher discrimination had greater percent bias, and MSE was greater for higher criteria location

estimates. Moreover, the percent bias was greater for the first criteria estimate. The end categories of the latent class sizes were overestimated with percent bias of about 6% and 8% for the first and fourth latent class sizes, respectively, which were lower than the percent bias from the 6 category condition. When non-normality was considered by concentrating the latent class sizes in class 2 and 3 (population values of 0.43 each), the percent bias was about 50% for the end categories. Finally, when class sizes were generated to shift from a higher concentration in class 1 and 2 to a higher concentration in class 3 and 4 (Tables B14 and B15), the percent bias was over 250% in fourth category in the first condition; it was over 220% in the second condition.

Standard Errors

Appendix C presents the standard deviation, mean standard error, bias, and percent bias for the conditions generated in this study. In general, the percent bias in standard errors was greater for raters with higher discrimination. Furthermore, percent bias was greater for raters that scored fewer essays in the unbalanced design.

The different conditions examined in this study seemed to affect percent bias in the standard errors of the latent class sizes. For example, when raters were stricter or more lenient, the percent bias of standard errors increased for the end categories when compared to Table C1, which represented the condition with population values of parameters without rater severity. Similarly, non-normality of the latent class sizes also increased the percent bias for the end categories, while the shift in the class sizes had greater percent bias for latent class sizes that did not have the concentration of class size. For example when classes 1 and 2 were larger, classes 3 and 4 had a greater percent bias than classes 1 and 2, which was also the case for the bias in class size estimates. As these

results indicate, the findings from the standard error estimates were also consistent with findings from DeCarlo (2008) and from DeCarlo (2010).

Chapter V

SUMMARY AND DISCUSSION

5.1 Summary

The use of CR items to evaluate examinee ability has increased over the years, which can be attributed to its role in validity. There are important skills that cannot be fully measured when only MC items are used (Livingston, 2009). CR items ask test takers to construct their own answer, which requires the use of raters. This introduces a subjective layer into scoring CR items, because scores given by the same rater can also differ across scoring occasions. Yet, scores generated from CR items must be reliable and valid, regardless of when an individual takes the test.

Differences in rater scores between testing administrations raise the issue of rater drift, which occurs when raters change their scoring behavior over different scoring occasions. Studies have found evidence of rater drift in real-world data (e.g., Congdon & McQueen, 2000) and have suggested the use of rater models (e.g., IRT models and the latent class SDT model) to adjust for rater effects such as rater severity when scoring CR items. However, the effect of rater drift on model-based classifications of essays into latent classes defined by the scoring rubric has not been studied comprehensively. To address these issues, this study had two main goals: (1) to examine how changes in rater behavior – rater drift – affect model-based classification and (2) to investigate the ability of different rater models to detect rater drift. These objectives were addressed using an analysis of real-world data and simulation studies.

Empirical study 1: Teacher certification test. In the empirical study, a teacher certification test and a high school writing test were used to identify patterns of rater drift using the latent class SDT model and IRT models. Parameter estimates from the rater models were used to detect patterns of rater drift. The teacher certification test was scored by 32 raters over 7 testing administrations on a 1 to 6 scale.

Plots of rater parameters showed minor individual variation in drift. These changes in rater behavior reflected variations in rater severity and in rater discrimination. Regression was used to summarize rater severity, which showed no significant linear (and nonlinear) trends; there were no significant trends for rater discrimination. Measures of classification (i.e., proportion correctly classified and lambda) showed stable estimates of classification accuracy for the seven testing administrations. Although there was evidence of rater drift in rater severity and in rater discrimination, these variations had a minimal effect on classification accuracy.

Empirical study 2: High school writing test. In the second phase of empirical analysis, the high school writing test was used to examine the effect of rater drift on classification accuracy and also to investigate patterns of rater drift using different rater models. This data differed from the teacher certification test in that there were 18 raters scoring over 12 months on a 1 to 4 scale.

This study produced results that were unexpected; one of the most notable results was that the discrimination parameters from the latent class SDT model showed a significant increase in parameter estimates, whereas the IRT models showed stable estimates across the scoring occasions. The estimated latent class sizes showed a non-normal distribution, with a greater class size in the middle scoring categories (i.e., 2 and

3). Estimates of classification accuracy showed minor changes over the 12 scoring occasions. Unlike the teacher certification test, results from the high school writing test showed differences between the latent class SDT model and IRT models that contradicted with respective to measures of rater discrimination.

Simulation study 1: Effect of rater drift on classification accuracy. Two simulation studies were conduced. In the first study, the effect of rater drift on classification accuracy was investigated. Using the latent class SDT model, data reflecting raters becoming stricter, more lenient, and a combination of raters that were both stricter and more lenient were generated over two scoring occasions. A separate condition was created that showed an increase in rater discrimination between two scoring occasions. Results showed that changes in rater severity had a minimal effect on classification accuracy. On the other hand, rater discrimination had a greater effect on classification accuracy – for an average increase in rater discrimination of two units, classification accuracy increased by about 20%.

Simulation study 2: Effect of rater drift on parameters of rater models. In the second simulation study, the effect of rater drift on parameter estimates of the GR model was examined using data generated from the latent class SDT model. Results showed that the GR model was able to detect changes in rater severity and in rater discrimination. This indicated that the GR model was sensitive to detect changes in both rater severity and in rater discrimination using data generated from the latent class SDT model.

The effect of different latent class sizes using data generated from the latent class SDT model on parameter estimates of the GR model was also examined. In general,

when the distribution of latent class sizes were non-normal with a greater concentration of class size in the middle scoring categories, the GR model underestimated rater discrimination.

Finally, the effect of shifting latent class sizes on parameter estimates of rater models was examined; this represented a greater concentration of scores in the higher scoring categories during the second scoring occasion than in the first scoring occasion, thereby creating a shift in the latent class sizes. This condition affected estimates of the criteria parameter for the latent class SDT model and the location parameters of the GR model to shift down. However, estimates of rater discrimination remained stable. This effect was consistent with the interpretation of the latent class sizes, where there were greater proportions of scores in the higher scoring categories, reflecting leniency among raters.

5.2 Discussion

Implication for rater training: Rater discrimination. This study showed that rater training focused on rater severity is an ineffective method to improve classification accuracy. Test developers and assessment agencies invest enormous amounts of time and energy to train raters using measures of agreement based on rater severity. This study reiterates an important result that has implication for rater training – raters should begin to focus on improving how well they discriminate between latent classes defined by the scoring rubric, because this plays an important role in determining how well raters classify an essay. This finding had been noted in previous studies (e.g., DeCarlo, 2002), but the literature on CR scoring is still dominated by training focused on rater severity.

As this study showed, changes in rater severity affected classification accuracy in only minor ways.

This study is one of the first to examine rater discrimination over time. Not many studies have examined rater discrimination in the context of rater drift. In fact, most rater models such as the FACETS model and the PC model do not estimate rater discrimination; these rater models constrain rater discrimination to be equal across all raters. Yet, empirical results from both the teacher certification test and the high school writing test showed notable differences in rater discrimination; in fact, rater discrimination was normally distributed between raters. Moreover, there were differences in rater discrimination for the same rater over time.

As demonstrated in this study, the use of rater discrimination to identify rater behavior is important and cannot be ignored – rater discrimination cannot be assumed to be equal across all raters. Given the significant role that rater discrimination plays on improving the quality of classification, the inclusion of rater discrimination in rater models is both empirically and theoretically motivated.

Classification of constructed responses. This study showed that a latent class SDT framework to study rater drift is useful as it presents additional insights into the behavior of raters. This approach differs from traditional rater models such as IRT that ranks constructed responses into a continuous latent trait. In the latent class SDT model, the interpretations of the latent classes are derived from the scoring rubric, which provides a natural context for conceptualizing CR scoring. For example, if an essay is classified into a "2," the scoring rubric provides a detailed description of the ability that the examinee demonstrated through the constructed response. The description provided in

the scoring rubric is important, because raters are trained on the basis of its description to score essays. As such, these classifications provide diagnostic feedback to examinees that are reflected in the scoring rubric used by the raters. On the other hand, it may be difficult to interpret latent scores derived from other rater models, because its relationship with the scoring rubric may not be clear.

Another benefit in using the latent class approach is the derivation of intersection criteria locations. As illustrated in the plots of rater criteria estimates in this study, the intersection criteria locations provide a relative guide on the severity of raters. Estimates of the relative criteria above the intersection criteria location may imply stricter rating, while estimates below may indicate lenient scoring. Although this location may be subjective, the close resemblance it showed with parameter estimates in the two real-world examples indicates its usefulness for diagnosing rater severity. These locations cannot be derived using an IRT framework, because the conceptual approach is different in that there are no clear locations to distinguish intersections.

The use of latent classes to examine rater behavior also allows an examination of classification. This measure can be used to compare different patterns of rater drift as demonstrated in this study. In the CR scoring setting, where raters are assumed to classify essays into a score defined by the scoring rubric, classification accuracy provides an important statistic that examines the quality of classification. Given the natural inclination in CR scoring to measure the classification accuracy of raters, this approach has not been studied in the context of rater drift. In light of these findings, this study adds to the literature in its understanding and implications for rater drift.

Treatment of examinee ability: Discrete versus continuous measures. This study also showed an important distinction between IRT models and the latent class SDT model as models for studying rater drift. Although the main difference between the two models lies in the treatment of examinee ability – whether to treat them as discrete or as a continuous latent measure – this distinction has led to important implications in assessing rater behavior.

This study found that the latent class SDT model is a useful model to examine rater drift. The latent class SDT model was able to detect differences in rater behavior that was comparable to IRT models. However, this study also found that when examinee ability was non-normal, parameter estimates of rater discrimination can lead to greater bias when using IRT models in comparison to the latent class SDT model. In both the latent class SDT model and the GR model, rater discrimination is the slope parameter of examinee ability. When examinee ability is treated as a continuous latent variable, the variance of examinee ability can affect parameter estimates of rater discrimination, which can subsequently also affect estimates of rater severity. As demonstrated in the high school writing test, examinee ability can be non-normal in that there can be a greater concentration of scores in the middle categories. Given that most IRT models assume examine ability to be normally distributed with fixed variance, this assumption must be checked for determining the type of rater model to use.

Studies of rater drift constitute an important and practical aspect of educational measurement. The use of CR items to measure examinee ability is increasing, and an attempt to understand errors resulting from human scoring behavior serves as an important step to refining how CR items should be measured. The study of rater drift is

important with respect to this growing area of CR scoring, because rater errors associated with changes in their behavior and their effect on model-based scores have not been studied comprehensively. With respect to these considerations, this study adds to the growing literature on assessing student ability based on subjective measures of rater scores. As this study concludes, the findings from this study provide new and important understanding of CR scoring and issues that emerge in practice, especially in exploring the effect rater drift has on different rater models.

5.3 Limitations and Future Research

There are several limitations to this study that could be addressed in future research. For example, in the high school writing test, there was a discrepancy in the results between the latent class SDT model and the IRT models. The discrimination parameter estimates were increasing over the twelve testing administrations for the latent class SDT model, while they were stable in the IRT models. Based on results from the simulation study, an increase in rater discrimination should also increase classification accuracy. However, classification accuracy remained stable and only fluctuated in minor ways. Several conditions including an examination of non-normality in latent class size distributions have been investigated using simulations, yet a clear understanding of the stability in classification accuracy has not been fully resolved. This requires further study.

The inference generated from the simulation study and results from the empirical study are only valid for specific conditions and substantive settings motivated in this study. A wider range of values can be examined for the simulation study that includes the effect of classification accuracy for other rater errors resulting from rater drift. Rater

effects such as scale shrinkage or null categories, where a rater refuses to use a particular scoring category, may have implications on classification accuracy. The literature on rater drift is mostly dominated by studies of rater severity, yet other forms of rater errors can be studied.

This study also ignored characteristics of the item, such as item difficulty or item discrimination, which may affect rater scores. DeCarlo (2010) examined the use of a hierarchical rater model using signal detection theory to implement item characteristics into the latent class SDT model. Extensions of the hierarchical rater model to examine the effect of rater drift on classification can be investigated in future research.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Boughton, K., Klinger, D., & Gierl, M. (2001). *Effects of random rater error on parameter recovery of the generalized partial credit model and graded response model.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, Seattle, WA.

Chase, C. I. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement*, *23*, 33-41.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311-359). New York: Plenum Press.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*, 163-178.

Dayton, C. M. (1998). *Latent class scaling analysis*. Thousand Oaks, CA: Sage Publications.

DeCarlo, L. T. (2002). A latent class extension of signal detection theory, with applications. *Multivariate Behavioral Research*, *37*, 423-451.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement*, *42*, 53-76.

DeCarlo, L. T. (2008). *Studies of a latent-class signal-detection model for constructed response scoring* (ETS Research Rep. No. RR-08-63). Princeton NJ: ETS.

DeCarlo, L. T. (2010). *Studies of a latent-class signal-detection model for constructed response scoring II: Incomplete and hierarchical designs* (ETS Research Report No. RR-10-08). Princeton NJ: ETS.

DeCarlo, L. T. (2011, April). *Implications of a signal detection rater model for the study of rater drift*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

Diederich, P.B., French, J. W., & Carlton, S.T. (1961). *Factors in the judgment of writing quality.* Princeton, NJ: Educational Testing Service.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society*, *53*, 460-75 and 644-63.

Ercikan, K., Schwarz, R. R., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, *35*, 137-154.

Galindo-Garre, F. G., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by bayesian posterior mode estimation. *Behaviormetrika*, *33*, 43-59.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, *46*(*1*), 43-58.

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation* (ETS Research Rep. No. RR-01-05). Princeton, NJ: ETS.

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State Examination. *Journal of Educational Measurement*, *38*(*2*), 121-145.

Hughes, D., & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement*, 21, 227-281.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, *13*(*2*), 121-138.

Johnson, R. L., Penny, J., & Johnson, C. (1998, June). Score resolution in the rating of performance assessments: Practices and issues. Paper presented at the council of Chief State School Officers National Conference on Large Scale Assessment, Colorado Springs, CO.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago: MESA Press.

Livingston, S. A. (2009). *R&D Connections — Constructed-response test questions: Why we use them; how we score them*. Princeton, NJ: Educational Testing Service.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, *12*, 54-71.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, *13*, 425-444.

Lunz, M. E., Wright, B. D., & J. M. Linacre (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*, 331-345.

Margolis, M., & Ross, L. (1995, April). *Halo and related effects in ratings by standardized patients in clinical evaluation*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of the modern item response theory* (pp. 101-121). New York: Springer.

McClellan, C. A. (2010). *R&D Connections — Constructed-response scoring—Doing It Right*. Princeton, NJ: Educational Testing Service.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McKinley, D., & Boulet, J. R. (2004). Detecting score drift in a high-stakes performance-based assessment. *Advances in Health Sciences Education*, *9*, 29–38.

McNamara, T. F. (1990). *Assessing the second language proficiency of health professionals*. Unpublished doctoral dissertation, University of Melbourne.

McNamara, T. F., & Adams, R. J. (1991). Exploring rater characteristics with Rasch techniques. In *Selected papers of the 13th Language Testing Research Colloquium (LTRC)*. Princeton, NJ: Educational Testing Service, International Testing and Training Program Office.

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

Myford, C. M. (1991, April). *Judging acting ability: The transition from notice to expert*. Paper presented at the American Educational Research Association, Chicago.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386-422.

National Board for Professional Teaching Standards. (1993). *Candidate guide*. San Antonio, TX: Author.

National Education Goals Panel. (1996). *Profile of 1994–1995 state assessment systems and reported results*. Washington, DC: Author.

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213-231). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Rudner, L. M. (1992). Reducing errors due to the use of judges. *Practical assessment, Research & Evaluation*, 3(3).

Samejima, F. (1969). Estimation of latent ability using a response of graded scores. *Psychometrika Monograph No*. *17*, 34.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effects of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, 76, 27-33.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement*, *36*, 336-346.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273-286.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent Gold 4.0: Basic and Advanced*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2007, February). *LG-Syntax™ User's Guide: Manual for Latent Gold 4.5 Syntax Module.* Belmont, MA: Statistical Innovations Inc.

Wheeler, P., Haertel, G., & Scriven, M. (1992). *Teacher Evaluation Glossary*, Kalamazoo, MI: CREATE Project, The Evaluation Center, Western Michigan University.

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson & G. Engelhard (Eds.), *Objective Measurement: Theory into Practice* (Vol. V, pp. 113-133). Stamford, CT: Ablex.

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement*, *2*, 256–280.

Wright, B. D., & Douglass, G.A. (1986). The Rating Scale Model for Objective Measurement. *MESA Research memorandum No. 35*. Chicago: University of Chicago.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Xi, X., & Mollaun, P. (2009) *How do raters from India perform in scoring the TOEFL iBT™ speaking section and what kind of training helps?* (ETS Research Report No. RR-09-31). Princeton, NJ: ETS.

Appendix A

Mean Parameter Estimates and Standard Errors of Study II

Table A1. Fitting the GR model to data generated from latent class SDT model using conditions from the criteria parameter

Time	Rater	b_1		b_2		b_3		b_4		b_5		a	
Baseline (*d*=norm4)	1	–4.134	(0.457)	–2.031	(0.309)	0.046	(0.229)	2.089	(0.314)	4.202	(0.463)	2.841	(0.342)
	2	–5.841	(0.724)	–2.926	(0.469)	0.026	(0.268)	2.944	(0.470)	5.879	(0.728)	3.942	(0.502)
	3	–5.814	(0.716)	–2.875	(0.463)	0.027	(0.268)	2.913	(0.466)	5.857	(0.725)	3.928	(0.500)
	4	–6.914	(0.910)	–3.484	(0.568)	–0.011	(0.287)	3.471	(0.574)	6.938	(0.922)	4.573	(0.596)
	5	–7.009	(0.933)	–3.534	(0.588)	0.047	(0.288)	3.572	(0.587)	7.077	(0.953)	4.674	(0.613)
	6	–7.021	(0.937)	–3.499	(0.579)	–0.031	(0.288)	3.535	(0.586)	7.082	(0.949)	4.653	(0.608)
	7	–7.245	(0.984)	–3.605	(0.605)	0.024	(0.289)	3.631	(0.608)	7.186	(0.974)	4.780	(0.633)
	8	–7.727	(1.070)	–3.889	(0.643)	0.000	(0.301)	3.933	(0.645)	7.793	(1.084)	5.009	(0.658)
	9	–7.570	(1.042)	–3.881	(0.642)	0.026	(0.296)	3.902	(0.647)	7.725	(1.076)	5.007	(0.659)
	10	–8.127	(1.148)	–4.059	(0.678)	0.002	(0.306)	4.091	(0.673)	8.207	(1.150)	5.221	(0.686)
More lenient (criteria for raters 4 to 9 shifted 1 down)	1	–4.154	(0.460)	–2.013	(0.288)	0.013	(0.230)	2.022	(0.287)	4.109	(0.455)	2.697	(0.324)
	2	–5.979	(0.743)	–2.816	(0.394)	–0.029	(0.314)	2.812	(0.391)	5.890	(0.735)	3.790	(0.489)
	3	–6.045	(0.751)	–2.861	(0.398)	0.013	(0.318)	2.859	(0.402)	5.918	(0.743)	3.826	(0.494)
	4	–7.788	(0.989)	–4.062	(0.580)	–0.872	(0.374)	2.595	(0.403)	6.187	(0.872)	4.358	(0.596)
	5	–7.756	(0.980)	–4.020	(0.565)	–0.865	(0.371)	2.530	(0.397)	6.142	(0.865)	4.350	(0.594)
	6	–7.745	(0.981)	–4.080	(0.574)	–0.906	(0.369)	2.537	(0.395)	6.086	(0.850)	4.317	(0.586)
	7	–7.798	(0.999)	–4.066	(0.573)	–0.877	(0.377)	2.623	(0.405)	6.222	(0.874)	4.398	(0.601)
	8	–8.322	(1.072)	–4.126	(0.584)	–0.682	(0.409)	3.024	(0.442)	7.137	(0.985)	4.735	(0.645)
	9	–8.409	(1.086)	–4.261	(0.605)	–0.736	(0.421)	3.081	(0.450)	7.136	(0.991)	4.806	(0.656)
	10	–8.528	(1.129)	–3.992	(0.553)	–0.021	(0.472)	3.942	(0.572)	8.426	(1.127)	5.179	(0.709)
Stricter (criteria	1	–4.081	(0.453)	–2.005	(0.287)	0.004	(0.231)	2.054	(0.291)	4.194	(0.461)	2.682	(0.321)
	2	–5.864	(0.732)	–2.770	(0.391)	0.014	(0.313)	2.924	(0.402)	5.973	(0.736)	3.756	(0.485)
	3	–5.901	(0.744)	–2.857	(0.401)	0.019	(0.314)	2.882	(0.398)	6.099	(0.759)	3.810	(0.496)

for raters 4 to 9 shifted 1 up)	4	–6.183	(0.868)	–2.559	(0.400)	0.951	(0.374)	4.157	(0.583)	7.917	(0.998)	4.381	(0.594)
	5	–6.219	(0.878)	–2.595	(0.404)	0.950	(0.375)	4.100	(0.574)	7.834	(0.998)	4.391	(0.601)
	6	–6.210	(0.874)	–2.577	(0.401)	0.906	(0.379)	4.124	(0.584)	7.937	(1.011)	4.407	(0.601)
	7	–6.242	(0.892)	–2.580	(0.407)	0.917	(0.381)	4.183	(0.595)	7.888	(1.009)	4.434	(0.609)
	8	–7.103	(0.984)	–3.035	(0.447)	0.639	(0.419)	4.243	(0.600)	8.358	(1.075)	4.772	(0.650)
	9	–7.045	(0.981)	–3.014	(0.446)	0.734	(0.416)	4.225	(0.599)	8.380	(1.078)	4.769	(0.649)
	10	–8.607	(1.150)	–3.964	(0.576)	–0.065	(0.480)	4.062	(0.563)	8.802	(1.151)	5.241	(0.709)
Both (criteria for raters 1, 4, 5, and 8 shifted up; raters 3, 6, 7, 9, and 10 down)	1	–3.066	(0.374)	–1.015	(0.249)	1.039	(0.249)	3.104	(0.381)	5.313	(0.576)	2.740	(0.338)
	2	–5.830	(0.731)	–2.742	(0.390)	0.035	(0.310)	2.766	(0.393)	5.813	(0.727)	3.715	(0.486)
	3	–6.843	(0.825)	–3.844	(0.527)	–1.002	(0.321)	2.010	(0.343)	4.946	(0.668)	3.741	(0.491)
	4	–6.042	(0.841)	–2.549	(0.398)	0.855	(0.369)	4.073	(0.588)	7.778	(0.993)	4.282	(0.575)
	5	–6.108	(0.858)	–2.559	(0.400)	0.839	(0.369)	4.035	(0.587)	7.675	(0.982)	4.298	(0.584)
	6	–7.513	(0.946)	–4.005	(0.574)	–0.797	(0.368)	2.599	(0.398)	6.108	(0.843)	4.268	(0.574)
	7	–7.742	(0.995)	–4.087	(0.599)	–0.846	(0.380)	2.602	(0.406)	6.233	(0.881)	4.365	(0.598)
	8	–7.130	(0.985)	–3.060	(0.450)	0.727	(0.417)	4.258	(0.630)	8.333	(1.093)	4.721	(0.648)
	9	–8.201	(1.068)	–4.101	(0.599)	–0.665	(0.416)	3.068	(0.452)	7.120	(0.978)	4.686	(0.642)
	10	–8.530	(1.130)	–4.203	(0.615)	–0.493	(0.453)	3.538	(0.502)	7.873	(1.067)	5.008	(0.686)

Note: Simulations ran with 100 replications

Table A2. Fitting the GR model to data generated from latent class SDT model using conditions from the discrimination parameter

Time	Rater	b_1		b_2		b_3		b_4		b_5		a	
1 *d*=norm 2	1	–1.010	(0.165)	–0.524	(0.151)	–0.009	(0.146)	0.487	(0.151)	0.994	(0.164)	0.671	(0.179)
	2	–2.031	(0.252)	–0.999	(0.195)	0.020	(0.173)	1.029	(0.196)	2.061	(0.254)	1.402	(0.250)
	3	–2.031	(0.250)	–1.044	(0.195)	–0.020	(0.172)	0.968	(0.192)	1.981	(0.247)	1.372	(0.247)
	4	–3.876	(0.548)	–1.924	(0.327)	0.000	(0.243)	1.892	(0.325)	3.898	(0.552)	2.629	(0.447)
	5	–3.941	(0.563)	–1.906	(0.329)	–0.018	(0.245)	1.871	(0.326)	3.898	(0.558)	2.663	(0.455)
	6	–3.917	(0.551)	–1.910	(0.326)	0.006	(0.242)	1.899	(0.326)	3.900	(0.549)	2.625	(0.445)
	7	–3.925	(0.559)	–1.894	(0.328)	0.003	(0.245)	1.961	(0.333)	3.950	(0.560)	2.675	(0.456)
	8	–5.124	(0.797)	–2.468	(0.422)	–0.010	(0.305)	2.387	(0.412)	5.122	(0.799)	3.356	(0.580)
	9	–5.136	(0.799)	–2.449	(0.416)	–0.010	(0.304)	2.393	(0.414)	5.144	(0.802)	3.348	(0.580)
	10	–5.814	(0.927)	–2.773	(0.467)	0.026	(0.345)	2.823	(0.475)	5.903	(0.938)	3.745	(0.647)
2 *d*=norm 4	1	–4.134	(0.457)	–2.031	(0.309)	0.046	(0.229)	2.089	(0.314)	4.202	(0.463)	2.841	(0.342)
	2	–5.841	(0.724)	–2.926	(0.469)	0.026	(0.268)	2.944	(0.470)	5.879	(0.728)	3.942	(0.502)
	3	–5.814	(0.716)	–2.875	(0.463)	0.027	(0.268)	2.913	(0.466)	5.857	(0.725)	3.928	(0.500)
	4	–6.914	(0.910)	–3.484	(0.568)	–0.011	(0.287)	3.471	(0.574)	6.938	(0.922)	4.573	(0.596)
	5	–7.009	(0.933)	–3.534	(0.588)	0.047	(0.288)	3.572	(0.587)	7.077	(0.953)	4.674	(0.613)
	6	–7.021	(0.937)	–3.499	(0.579)	–0.031	(0.288)	3.535	(0.586)	7.082	(0.949)	4.653	(0.608)
	7	–7.245	(0.984)	–3.605	(0.605)	0.024	(0.289)	3.631	(0.608)	7.186	(0.974)	4.780	(0.633)
	8	–7.727	(1.070)	–3.889	(0.643)	0.000	(0.301)	3.933	(0.645)	7.793	(1.084)	5.009	(0.658)
	9	–7.570	(1.042)	–3.881	(0.642)	0.026	(0.296)	3.902	(0.647)	7.725	(1.076)	5.007	(0.659)
	10	–8.127	(1.148)	–4.059	(0.678)	0.002	(0.306)	4.091	(0.673)	8.207	(1.150)	5.221	(0.686)

Note: Simulations ran with 100 replications; for time 1, the *d* for rater 1 is 0.5.

Table A3. Fitting data generated from the latent class SDT model: First Scoring Occasion
Normally distributed latent class sizes (BIB design with *d* = norm 4)

Model	Rater	Generating Value						Mean Estimates											
		LC1	LC2	LC3	LC4	LC5	LC6	LC1		LC2		LC3		LC4		LC5		LC6	
		0.08	0.17	0.25	0.25	0.17	0.08	0.089	(0.014)	0.166	(0.019)	0.242	(0.023)	0.244	(0.023)	0.168	(0.020)	0.092	(0.015)
		c_1	c_2	c_3	c_4	c_5	d	c_1		c_2		c_3		c_4		c_5		d	
LC-SDT	1	1.0	3.0	5.0	7.0	9.0	2	0.91	(0.78)	2.88	(0.67)	4.90	(0.58)	6.91	(0.51)	8.91	(0.48)	1.96	(0.21)
	2	1.5	4.5	7.5	10.5	13.5	3	1.35	(1.42)	4.52	(1.19)	7.55	(1.01)	10.70	(0.83)	13.85	(0.74)	3.05	(0.38)
	3	1.5	4.5	7.5	10.5	13.5	3	1.27	(1.39)	4.47	(1.17)	7.61	(0.97)	10.71	(0.82)	13.82	(0.74)	3.03	(0.37)
	4	2.0	6.0	10.0	14.0	18.0	4	1.76	(2.21)	5.96	(1.82)	10.14	(1.49)	14.28	(1.20)	18.45	(1.05)	4.04	(0.58)
	5	2.0	6.0	10.0	14.0	18.0	4	1.71	(2.21)	5.88	(1.83)	10.05	(1.47)	14.12	(1.18)	18.33	(1.01)	4.00	(0.57)
	6	2.0	6.0	10.0	14.0	18.0	4	1.72	(2.20)	5.95	(1.81)	10.18	(1.48)	14.36	(1.21)	18.52	(1.05)	4.04	(0.57)
	7	2.0	6.0	10.0	14.0	18.0	4	1.72	(2.18)	5.88	(1.79)	10.11	(1.44)	14.15	(1.16)	18.41	(0.99)	4.00	(0.56)
	8	2.5	7.5	12.5	17.5	22.5	5	1.97	(3.04)	7.25	(2.46)	12.41	(2.01)	17.47	(1.61)	22.60	(1.38)	4.92	(0.78)
	9	2.5	7.5	12.5	17.5	22.5	5	1.94	(3.00)	7.11	(2.48)	12.43	(1.97)	17.51	(1.62)	22.59	(1.36)	4.93	(0.76)
	10	3.0	9.0	15.0	21.0	27.0	6	2.27	(3.55)	8.13	(2.91)	13.94	(2.35)	19.82	(1.82)	25.76	(1.51)	5.55	(0.89)
								b_1		b_2		b_3		b_4		b_5		a	
GR	1							–4.23	(0.46)	–2.05	(0.29)	–0.02	(0.23)	1.99	(0.29)	4.10	(0.45)	2.67	(0.32)
	2							–6.02	(0.73)	–2.89	(0.40)	–0.04	(0.31)	2.80	(0.39)	5.88	(0.72)	3.74	(0.48)
	3							–6.04	(0.74)	–2.94	(0.41)	–0.03	(0.32)	2.83	(0.40)	6.07	(0.74)	3.76	(0.48)
	4							–7.38	(0.94)	–3.54	(0.49)	–0.08	(0.39)	3.41	(0.47)	7.29	(0.93)	4.49	(0.59)
	5							–7.47	(0.96)	–3.55	(0.50)	–0.08	(0.39)	3.38	(0.47)	7.34	(0.95)	4.50	(0.60)
	6							–7.47	(0.95)	–3.49	(0.49)	–0.01	(0.39)	3.45	(0.48)	7.29	(0.94)	4.48	(0.59)
	7							–7.45	(0.95)	–3.56	(0.50)	–0.08	(0.40)	3.41	(0.48)	7.25	(0.94)	4.52	(0.60)
	8							–8.32	(1.07)	–3.88	(0.55)	–0.03	(0.44)	3.74	(0.52)	8.09	(1.05)	4.93	(0.66)
	9							–8.26	(1.07)	–3.90	(0.55)	–0.01	(0.45)	3.83	(0.54)	8.17	(1.06)	4.96	(0.66)
	10							–8.89	(1.15)	–4.19	(0.59)	–0.09	(0.48)	4.12	(0.59)	8.73	(1.14)	5.24	(0.70)

Note: The table shows mean estimates of data generated using the LC-SDT model (generating values specified). LC-SDT model refers to the latent class SDT model. Data were fit using the LC-SDT and the GR model. Mean estimates were calculated for 100 replications. Values in parenthesis represent standard errors.

Table A4. Fitting data generated from the latent class SDT model: Second Scoring Occasion
Non-normally distributed latent class sizes (BIB design with *d* = norm 4)

Model	Rater	Generating Value						Mean Estimates											
		LC1	LC2	LC3	LC4	LC5	LC6	LC1		LC2		LC3		LC4		LC5		LC6	
		0.03	0.03	0.40	0.40	0.10	0.04	0.034	(0.007)	0.042	(0.012)	0.390	(0.027)	0.374	(0.028)	0.109	(0.019)	0.050	(0.010)
		c_1	c_2	c_3	c_4	c_5	d	c_1		c_2		c_3		c_4		c_5		d	
LC-SDT	1	1.0	3.0	5.0	7.0	9.0	2	0.76	(0.91)	2.75	(0.74)	4.78	(0.65)	6.83	(0.56)	8.86	(0.52)	1.92	(0.26)
	2	1.5	4.5	7.5	10.5	13.5	3	1.06	(1.68)	4.17	(1.22)	7.26	(1.04)	10.41	(0.85)	13.59	(0.72)	2.93	(0.41)
	3	1.5	4.5	7.5	10.5	13.5	3	1.15	(1.69)	4.31	(1.22)	7.38	(1.05)	10.58	(0.86)	13.85	(0.76)	2.97	(0.42)
	4	2.0	6.0	10.0	14.0	18.0	4	1.45	(2.58)	5.55	(1.78)	9.64	(1.51)	13.85	(1.19)	18.02	(0.99)	3.88	(0.60)
	5	2.0	6.0	10.0	14.0	18.0	4	1.55	(2.66)	5.69	(1.82)	9.85	(1.55)	14.17	(1.22)	18.40	(0.97)	3.96	(0.61)
	6	2.0	6.0	10.0	14.0	18.0	4	1.42	(2.70)	5.66	(1.85)	9.85	(1.57)	14.18	(1.23)	18.41	(0.99)	3.97	(0.62)
	7	2.0	6.0	10.0	14.0	18.0	4	1.42	(2.62)	5.51	(1.79)	9.63	(1.50)	13.99	(1.15)	18.14	(0.94)	3.88	(0.60)
	8	2.5	7.5	12.5	17.5	22.5	5	1.89	(3.42)	6.55	(2.35)	11.64	(1.95)	16.91	(1.45)	21.81	(1.19)	4.67	(0.77)
	9	2.5	7.5	12.5	17.5	22.5	5	1.87	(3.52)	6.61	(2.42)	11.73	(2.01)	17.10	(1.49)	22.13	(1.17)	4.74	(0.79)
	10	3.0	9.0	15.0	21.0	27.0	6	2.04	(4.04)	7.14	(2.79)	13.05	(2.30)	19.23	(1.65)	24.78	(1.32)	5.27	(0.90)
								b_1		b_2		b_3		b_4		b_5		a	
GR								–4.33	(0.48)	–2.45	(0.29)	–0.29	(0.19)	1.80	(0.25)	3.75	(0.41)	1.88	(0.27)
								–6.21	(0.81)	–3.68	(0.48)	–0.36	(0.24)	2.83	(0.40)	5.47	(0.69)	2.72	(0.40)
								–6.24	(0.81)	–3.66	(0.47)	–0.34	(0.24)	2.74	(0.39)	5.39	(0.68)	2.70	(0.40)
								–7.59	(1.09)	–4.98	(0.70)	–0.43	(0.28)	3.64	(0.54)	6.72	(0.94)	3.36	(0.52)
								–7.92	(1.17)	–5.18	(0.73)	–0.46	(0.29)	3.80	(0.57)	6.90	(0.98)	3.48	(0.54)
								–7.67	(1.11)	–4.98	(0.71)	–0.46	(0.28)	3.69	(0.55)	6.85	(0.97)	3.46	(0.54)
								–7.65	(1.11)	–5.01	(0.71)	–0.43	(0.28)	3.66	(0.55)	6.79	(0.96)	3.44	(0.54)
								–8.66	(1.32)	–6.05	(0.90)	–0.55	(0.32)	4.40	(0.67)	7.83	(1.16)	3.93	(0.62)
								–8.84	(1.36)	–6.19	(0.92)	–0.53	(0.32)	4.50	(0.69)	8.01	(1.19)	4.02	(0.64)
								–9.53	(1.49)	–6.93	(1.05)	–0.58	(0.34)	4.96	(0.77)	8.66	(1.31)	4.32	(0.69)

Note: The table shows mean estimates of data generated using the LC-SDT model (generating values specified). LC-SDT model refers to the latent class SDT model. Data were fit using the LC-SDT and the GR model. Mean estimates were calculated for 100 replications. Values in parenthesis represent standard errors.

Table A5. Fitting data generated from the latent class SDT model: First Scoring Occasion
Non-normally distributed latent class sizes (BIB design with *d* = norm 4)

Model	Rater	Generating Value				Mean Estimates			
		LC1	LC2	LC3	LC4	LC1	LC2	LC3	LC4
		0.17	0.33	0.33	0.17	0.180 (0.021)	0.320 (0.026)	0.316 (0.027)	0.183 (0.022)
		c_1	c_2	c_3	d	c_1	c_2	c_3	d
LC-SDT	1	1.0	3.0	5.0	2	0.97 (0.48)	3.03 (0.45)	5.12 (0.46)	2.02 (0.27)
	2	1.5	4.5	7.5	3	1.46 (0.74)	4.53 (0.67)	7.63 (0.67)	2.99 (0.42)
	3	1.5	4.5	7.5	3	1.44 (0.73)	4.52 (0.67)	7.64 (0.67)	3.01 (0.42)
	4	2.0	6.0	10.0	4	1.82 (1.07)	6.10 (0.98)	10.33 (1.01)	4.04 (0.62)
	5	2.0	6.0	10.0	4	1.90 (1.07)	6.10 (0.99)	10.30 (1.03)	4.05 (0.64)
	6	2.0	6.0	10.0	4	1.88 (1.07)	6.01 (0.97)	10.21 (0.97)	4.03 (0.62)
	7	2.0	6.0	10.0	4	1.95 (1.12)	6.18 (0.99)	10.46 (1.04)	4.11 (0.65)
	8	2.5	7.5	12.5	5	2.26 (1.52)	7.61 (1.33)	13.01 (1.38)	5.08 (0.90)
	9	2.5	7.5	12.5	5	2.47 (1.62)	7.72 (1.43)	13.27 (1.37)	5.14 (0.95)
	10	3.0	9.0	15.0	6	2.38 (1.81)	8.52 (1.62)	14.71 (1.66)	5.69 (1.04)
						b_1	b_2	b_3	a
GR						–2.08 (0.28)	0.02 (0.20)	2.08 (0.28)	2.04 (0.31)
						–2.97 (0.43)	0.02 (0.25)	3.00 (0.43)	2.95 (0.46)
						–3.00 (0.43)	0.01 (0.25)	3.00 (0.43)	2.93 (0.46)
						–3.88 (0.59)	0.00 (0.30)	3.90 (0.60)	3.74 (0.60)
						–3.83 (0.58)	–0.04 (0.29)	3.79 (0.58)	3.65 (0.58)
						–3.91 (0.60)	–0.01 (0.30)	3.90 (0.60)	3.78 (0.60)
						–3.84 (0.58)	0.01 (0.29)	3.86 (0.59)	3.67 (0.59)
						–4.46 (0.70)	–0.03 (0.33)	4.45 (0.70)	4.20 (0.67)
						–4.55 (0.72)	–0.01 (0.33)	4.49 (0.71)	4.31 (0.69)
						–5.02 (0.80)	0.01 (0.35)	5.05 (0.80)	4.64 (0.74)

Note: The table shows mean estimates of data generated using the LC-SDT model (generating values specified). LC-SDT model refers to the latent class SDT model. Data were fit using the LC-SDT and the GR model. Mean estimates were calculated for 100 replications. Values in parenthesis represent standard errors.

Table A6. Fitting data generated from the latent class SDT model: Second Scoring Occasion
Non-normally distributed latent class sizes (BIB design with *d* = norm 4)

Model	Rater	Generating Value				Mean Estimates			
		LC1	LC2	LC3	LC4	LC1	LC2	LC3	LC4
		0.07	0.43	0.43	0.07	0.106 (0.022)	0.397 (0.031)	0.393 (0.030)	0.104 (0.021)
		c_1	c_2	c_3	d	c_1	c_2	c_3	d
LC-SDT	1	1.0	3.0	5.0	2	0.63 (0.50)	2.67 (0.49)	4.68 (0.52)	1.84 (0.31)
	2	1.5	4.5	7.5	3	0.98 (0.73)	4.10 (0.71)	7.22 (0.75)	2.79 (0.45)
	3	1.5	4.5	7.5	3	1.07 (0.74)	4.29 (0.74)	7.44 (0.79)	2.80 (0.46)
	4	2.0	6.0	10.0	4	1.14 (0.96)	5.42 (0.98)	9.63 (1.07)	3.82 (0.67)
	5	2.0	6.0	10.0	4	1.23 (1.01)	5.46 (0.99)	9.80 (1.06)	3.65 (0.62)
	6	2.0	6.0	10.0	4	1.27 (0.99)	5.59 (1.00)	9.93 (1.12)	3.76 (0.64)
	7	2.0	6.0	10.0	4	1.27 (1.06)	5.65 (1.03)	10.04 (1.09)	3.72 (0.64)
	8	2.5	7.5	12.5	5	1.40 (1.49)	7.19 (1.40)	12.93 (1.51)	4.79 (0.92)
	9	2.5	7.5	12.5	5	1.41 (1.38)	6.92 (1.33)	12.46 (1.45)	4.84 (0.92)
	10	3.0	9.0	15.0	6	1.26 (1.71)	8.18 (1.71)	14.94 (1.92)	5.50 (1.11)
						b_1	b_2	b_3	a
GR						–2.02 (0.26)	0.00 (0.17)	2.00 (0.25)	1.40 (0.26)
						–3.10 (0.42)	0.05 (0.21)	3.13 (0.43)	2.18 (0.39)
						–3.16 (0.43)	0.00 (0.22)	3.15 (0.43)	2.21 (0.39)
						–4.06 (0.60)	0.03 (0.25)	4.10 (0.60)	2.75 (0.49)
						–4.02 (0.59)	0.02 (0.24)	4.03 (0.59)	2.68 (0.48)
						–4.17 (0.62)	0.05 (0.25)	4.16 (0.62)	2.77 (0.50)
						–4.10 (0.61)	0.01 (0.25)	4.12 (0.61)	2.76 (0.49)
						–4.90 (0.75)	0.01 (0.27)	4.92 (0.75)	3.13 (0.55)
						–4.98 (0.76)	0.06 (0.28)	4.98 (0.77)	3.19 (0.57)
						–5.63 (0.88)	0.01 (0.30)	5.71 (0.89)	3.49 (0.61)

Note: The table shows mean estimates of data generated using the LC-SDT model (generating values specified). LC-SDT model refers to the latent class SDT model. Data were fit using the LC-SDT and the GR model. Mean estimates were calculated for 100 replications. Values in parenthesis represent standard errors.

Table A7. Fitting data generated from the latent class SDT model: First Scoring Occasion Shift in latent class sizes (BIB design with *d* = norm 4)

Model	Rater	Generating Value				Mean Estimates			
		LC1	LC2	LC3	LC4	LC1	LC2	LC3	LC4
		0.07	0.50	0.40	0.03	0.109 (0.024)	0.446 (0.033)	0.340 (0.038)	0.106 (0.029)
		c_1	c_2	c_3	d	c_1	c_2	c_3	d
LC-SDT	1	1.0	3.0	5.0	2	0.61 (0.48)	2.65 (0.50)	4.71 (0.56)	1.72 (0.32)
	2	1.5	4.5	7.5	3	0.90 (0.66)	3.99 (0.70)	7.15 (0.80)	2.55 (0.45)
	3	1.5	4.5	7.5	3	0.93 (0.67)	4.05 (0.72)	7.17 (0.81)	2.58 (0.46)
	4	2.0	6.0	10.0	4	1.24 (0.97)	5.60 (1.05)	9.92 (1.18)	3.57 (0.67)
	5	2.0	6.0	10.0	4	1.23 (0.98)	5.54 (1.07)	10.05 (1.22)	3.58 (0.68)
	6	2.0	6.0	10.0	4	1.37 (1.03)	5.75 (1.11)	10.41 (1.22)	3.69 (0.71)
	7	2.0	6.0	10.0	4	1.18 (0.98)	5.56 (1.02)	9.99 (1.19)	3.56 (0.66)
	8	2.5	7.5	12.5	5	1.41 (1.35)	7.09 (1.44)	12.93 (1.58)	4.51 (0.92)
	9	2.5	7.5	12.5	5	1.33 (1.25)	6.90 (1.39)	12.62 (1.65)	4.43 (0.89)
	10	3.0	9.0	15.0	6	1.32 (1.64)	8.25 (1.76)	15.24 (2.01)	5.25 (1.13)
						b_1	b_2	b_3	a
GR						–1.80 (0.24)	0.20 (0.17)	2.25 (0.27)	1.30 (0.26)
						–2.75 (0.37)	0.31 (0.20)	3.37 (0.44)	1.89 (0.36)
						–2.72 (0.38)	0.35 (0.21)	3.41 (0.45)	1.94 (0.37)
						–3.73 (0.55)	0.46 (0.24)	4.64 (0.67)	2.48 (0.47)
						–3.71 (0.55)	0.48 (0.24)	4.53 (0.66)	2.46 (0.47)
						–3.78 (0.57)	0.45 (0.25)	4.69 (0.70)	2.57 (0.49)
						–3.66 (0.55)	0.48 (0.25)	4.58 (0.67)	2.49 (0.47)
						–4.64 (0.74)	0.59 (0.28)	5.78 (0.91)	3.00 (0.57)
						–4.47 (0.70)	0.60 (0.28)	5.64 (0.88)	2.91 (0.56)
						–5.23 (0.85)	0.63 (0.30)	6.55 (1.06)	3.23 (0.61)

Note: The table shows mean estimates of data generated using the LC-SDT model (generating values specified). LC-SDT model refers to the latent class SDT model. Data were fit using the LC-SDT and the GR model. Mean estimates were calculated for 100 replications. Values in parenthesis represent standard errors.

Table A8. Fitting data generated from the latent class SDT model: Second Scoring Occasion Shift in latent class sizes (BIB design with *d* = norm 4)

Model	Rater	Generating Value				Mean Estimates			
		LC1	LC2	LC3	LC4	LC1	LC2	LC3	LC4
		0.03	0.40	0.50	0.07	0.096 (0.027)	0.346 (0.036)	0.446 (0.033)	0.111 (0.025)
		c_1	c_2	c_3	d	c_1	c_2	c_3	d
LC-SDT	1	1.0	3.0	5.0	2	0.48 (0.59)	2.57 (0.51)	4.59 (0.47)	1.74 (0.33)
	2	1.5	4.5	7.5	3	0.66 (0.86)	3.72 (0.71)	6.79 (0.62)	2.57 (0.46)
	3	1.5	4.5	7.5	3	0.64 (0.86)	3.77 (0.70)	6.91 (0.62)	2.60 (0.45)
	4	2.0	6.0	10.0	4	0.86 (1.36)	5.34 (1.10)	9.82 (0.96)	3.69 (0.71)
	5	2.0	6.0	10.0	4	0.74 (1.25)	5.08 (1.00)	9.42 (0.87)	3.50 (0.65)
	6	2.0	6.0	10.0	4	0.78 (1.30)	5.28 (1.03)	9.59 (0.92)	3.61 (0.67)
	7	2.0	6.0	10.0	4	0.72 (1.36)	5.28 (1.05)	9.74 (0.90)	3.64 (0.68)
	8	2.5	7.5	12.5	5	0.72 (1.80)	6.57 (1.42)	12.18 (1.19)	4.49 (0.91)
	9	2.5	7.5	12.5	5	0.61 (1.82)	6.44 (1.35)	11.99 (1.11)	4.40 (0.88)
	10	3.0	9.0	15.0	6	0.67 (2.44)	7.94 (1.82)	15.02 (1.46)	5.48 (1.16)
						b_1	b_2	b_3	a
GR						–2.23 (0.27)	–0.23 (0.17)	1.83 (0.24)	1.31 (0.26)
						–3.51 (0.47)	–0.37 (0.21)	2.80 (0.39)	2.02 (0.38)
						–3.40 (0.44)	–0.35 (0.20)	2.67 (0.36)	1.87 (0.35)
						–4.69 (0.69)	–0.49 (0.25)	3.71 (0.56)	2.55 (0.49)
						–4.58 (0.67)	–0.48 (0.25)	3.68 (0.55)	2.49 (0.47)
						–4.56 (0.66)	–0.44 (0.24)	3.66 (0.55)	2.49 (0.47)
						–4.53 (0.66)	–0.48 (0.24)	3.77 (0.56)	2.49 (0.47)
						–5.60 (0.87)	–0.55 (0.28)	4.49 (0.70)	2.90 (0.55)
						–5.60 (0.87)	–0.57 (0.28)	4.55 (0.72)	2.95 (0.56)
						–6.62 (1.08)	–0.68 (0.31)	5.30 (0.86)	3.32 (0.62)

Note: The table shows mean estimates of data generated using the LC-SDT model (generating values specified). LC-SDT model refers to the latent class SDT model. Data were fit using the LC-SDT and the GR model. Mean estimates were calculated for 100 replications. Values in parenthesis represent standard errors.

Appendix B

Parameter Estimates, Bias, Percent Bias, and MSE

Table B1. Intersection Point Criteria, *d* = Normal 4, BIB, 6 Categories with Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
216	d_1	2.0	1.956	–0.0438	2.190	0.041
216	d_2	3.0	3.045	0.0454	1.513	0.153
216	d_3	3.0	3.029	0.0294	0.980	0.157
216	d_4	4.0	4.040	0.0398	0.995	0.301
216	d_5	4.0	3.995	–0.0046	0.115	0.250
216	d_6	4.0	4.043	0.0429	1.073	0.224
216	d_7	4.0	3.997	–0.0027	0.068	0.269
216	d_8	5.0	4.916	–0.0843	1.686	0.406
216	d_9	5.0	4.928	–0.0725	1.450	0.377
216	d_{10}	6.0	5.555	–0.4453	7.422	0.707
	c_{11}	1.0	0.910	–0.0899	8.990	0.119
	c_{12}	3.0	2.884	–0.1159	3.863	0.155
	c_{13}	5.0	4.901	–0.0994	1.988	0.302
	c_{14}	7.0	6.914	–0.0865	1.236	0.475
	c_{15}	9.0	8.905	–0.0946	1.051	0.686
	c_{21}	1.5	1.346	–0.1540	10.267	0.281
	c_{22}	4.5	4.524	0.0236	0.524	0.528
	c_{23}	7.5	7.555	0.0546	0.728	0.884
	c_{24}	10.5	10.698	0.1975	1.881	1.794
	c_{25}	13.5	13.850	0.3497	2.590	3.024
	c_{31}	1.5	1.274	–0.2265	15.100	0.268
	c_{32}	4.5	4.467	–0.0332	0.738	0.495
	c_{33}	7.5	7.607	0.1074	1.432	1.019
	c_{34}	10.5	10.711	0.2112	2.011	1.865
	c_{35}	13.5	13.822	0.3221	2.386	2.947
	c_{41}	2.0	1.758	–0.2420	12.100	0.519
	c_{42}	6.0	5.959	–0.0409	0.682	1.062
	c_{43}	10.0	10.145	0.1446	1.446	1.993
	c_{44}	14.0	14.282	0.2815	2.011	3.890
	c_{45}	18.0	18.450	0.4498	2.499	6.712
	c_{51}	2.0	1.711	–0.2891	14.455	0.561
	c_{52}	6.0	5.880	–0.1198	1.997	1.064
	c_{53}	10.0	10.053	0.0525	0.525	1.808
	c_{54}	14.0	14.122	0.1222	0.873	2.985
	c_{55}	18.0	18.328	0.3284	1.824	5.130

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{61}	2.0	1.723	–0.2766	13.830	0.542
	c_{62}	6.0	5.950	–0.0503	0.838	0.818
	c_{63}	10.0	10.180	0.1795	1.795	1.760
	c_{64}	14.0	14.360	0.3602	2.573	3.074
	c_{65}	18.0	18.516	0.5159	2.866	4.681
	c_{71}	2.0	1.715	–0.2850	14.250	0.553
	c_{72}	6.0	5.881	–0.1190	1.983	0.893
	c_{73}	10.0	10.106	0.1064	1.064	2.025
	c_{74}	14.0	14.149	0.1494	1.067	3.704
	c_{75}	18.0	18.407	0.4068	2.260	5.771
	c_{81}	2.5	1.973	–0.5274	21.095	1.008
	c_{82}	7.5	7.247	–0.2533	3.378	1.530
	c_{83}	12.5	12.410	–0.0905	0.724	2.539
	c_{84}	17.5	17.467	–0.0335	0.191	4.862
	c_{85}	22.5	22.601	0.1005	0.447	7.933
	c_{91}	2.5	1.939	–0.5611	22.443	0.959
	c_{92}	7.5	7.113	–0.3874	5.165	1.250
	c_{93}	12.5	12.431	–0.0690	0.552	2.635
	c_{94}	17.5	17.506	0.0062	0.035	5.037
	c_{95}	22.5	22.586	0.0857	0.381	8.017
	c_{101}	3.0	2.270	–0.7299	24.331	1.351
	c_{102}	9.0	8.131	–0.8690	9.656	2.494
	c_{103}	15.0	13.935	–1.0646	7.097	4.480
	c_{104}	21.0	19.818	–1.1821	5.629	7.953
	c_{105}	27.0	25.757	–1.2426	4.602	12.395
			Latent class sizes			
	Class 1	0.080	0.089	0.0090	11.250	
	Class 2	0.170	0.166	–0.0040	2.353	
	Class 3	0.250	0.242	–0.0080	3.200	
	Class 4	0.250	0.244	–0.0060	2.400	
	Class 5	0.170	0.168	–0.0020	1.176	
	Class 6	0.080	0.092	0.0120	15.000	

Table B2. Criteria Shifted Up (Strict Raters), *d* = Normal 4, BIB, 6 Categories with Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
216	d_1	2.0	1.966	–0.0340	1.700	0.051
216	d_2	3.0	3.006	0.0057	0.190	0.141
216	d_3	3.0	3.020	0.0200	0.667	0.159
216	d_4	4.0	3.815	–0.1851	4.628	0.292
216	d_5	4.0	3.828	–0.1719	4.298	0.280
216	d_6	4.0	3.760	–0.2397	5.993	0.340
216	d_7	4.0	3.842	–0.1585	3.963	0.310
216	d_8	5.0	4.429	–0.5706	11.412	0.680
216	d_9	5.0	4.524	–0.4762	9.524	0.633
216	d_{10}	6.0	5.380	–0.6205	10.342	0.743
	c_{11}	1.0	0.646	–0.3537	35.370	0.325
	c_{12}	3.0	2.709	–0.2913	9.710	0.340
	c_{13}	5.0	4.761	–0.2392	4.784	0.445
	c_{14}	7.0	6.804	–0.1960	2.800	0.608
	c_{15}	9.0	8.810	–0.1902	2.113	0.954
	c_{21}	1.5	1.041	–0.4594	30.627	0.524
	c_{22}	4.5	4.189	–0.3111	6.913	0.620
	c_{23}	7.5	7.321	–0.1793	2.391	1.043
	c_{24}	10.5	10.349	–0.1514	1.442	1.797
	c_{25}	13.5	13.539	0.0385	0.285	2.539
	c_{31}	1.5	1.036	–0.4642	30.947	0.456
	c_{32}	4.5	4.230	–0.2701	6.002	0.668
	c_{33}	7.5	7.335	–0.1647	2.196	1.129
	c_{34}	10.5	10.457	–0.0435	0.414	1.746
	c_{35}	13.5	13.619	0.1193	0.884	2.844
	c_{41}	3.0	2.200	–0.8005	26.683	1.346
	c_{42}	7.0	6.285	–0.7147	10.210	1.747
	c_{43}	11.0	10.216	–0.7842	7.129	2.901
	c_{44}	15.0	14.150	–0.8496	5.664	4.479
	c_{45}	19.0	18.165	–0.8352	4.396	5.998
	c_{51}	3.0	2.115	–0.8846	29.487	1.392
	c_{52}	7.0	6.326	–0.6737	9.624	1.487
	c_{53}	11.0	10.188	–0.8125	7.386	2.555
	c_{54}	15.0	14.173	–0.8268	5.512	4.370
	c_{55}	19.0	18.202	–0.7982	4.201	5.830
	c_{61}	3.0	2.134	–0.8665	28.883	1.349
	c_{62}	7.0	6.188	–0.8116	11.594	2.058
	c_{63}	11.0	9.990	–1.0100	9.182	3.487
	c_{64}	15.0	13.935	–1.0650	7.100	5.135

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{65}	19.0	17.917	–1.0832	5.701	7.160
	c_{71}	3.0	2.177	–0.8232	27.440	1.440
	c_{72}	7.0	6.330	–0.6705	9.579	1.853
	c_{73}	11.0	10.315	–0.6846	6.224	2.960
	c_{74}	15.0	14.187	–0.8133	5.422	4.517
	c_{75}	19.0	18.229	–0.7706	4.056	7.220
	c_{81}	3.5	2.060	–1.4401	41.145	2.840
	c_{82}	8.5	6.868	–1.6317	19.196	3.864
	c_{83}	13.5	11.469	–2.0314	15.048	6.617
	c_{84}	18.5	16.140	–2.3602	12.758	10.410
	c_{85}	23.5	20.885	–2.6150	11.128	14.825
	c_{91}	3.5	2.187	–1.3135	37.528	2.689
	c_{92}	8.5	7.123	–1.3766	16.195	3.521
	c_{93}	13.5	11.754	–1.7457	12.931	6.201
	c_{94}	18.5	16.385	–2.1148	11.431	9.724
	c_{95}	23.5	21.194	–2.3065	9.815	13.547
	c_{101}	3.0	1.295	–1.7046	56.819	3.728
	c_{102}	9.0	7.226	–1.7745	19.716	4.568
	c_{103}	15.0	12.975	–2.0254	13.502	6.827
	c_{104}	21.0	18.656	–2.3440	11.162	10.350
	c_{105}	27.0	24.468	–2.5317	9.377	14.092
			Latent class sizes			
	Class 1	0.080	0.108	0.0280	35.000	
	Class 2	0.170	0.163	–0.0070	4.118	
	Class 3	0.250	0.248	–0.0020	0.800	
	Class 4	0.250	0.240	–0.0100	4.000	
	Class 5	0.170	0.159	–0.0110	6.471	
	Class 6	0.080	0.083	0.0030	3.750	

Table B3. Criteria Shifted Down (Lenient Raters), *d* = Normal 4, BIB, 6 Categories with Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
216	d_1	2.0	1.910	–0.0901	4.505	0.049
216	d_2	3.0	2.949	–0.0508	1.693	0.130
216	d_3	3.0	3.007	0.0067	0.223	0.122
216	d_4	4.0	3.831	–0.1689	4.223	0.309
216	d_5	4.0	3.678	–0.3219	8.048	0.319
216	d_6	4.0	3.783	–0.2167	5.418	0.267
216	d_7	4.0	3.784	–0.2161	5.403	0.295
216	d_8	5.0	4.470	–0.5305	10.610	0.636
216	d_9	5.0	4.520	–0.4800	9.600	0.563
216	d_{10}	6.0	5.540	–0.4604	7.673	0.563
	c_{11}	1.0	0.885	–0.1152	11.520	0.204
	c_{12}	3.0	2.942	–0.0580	1.933	0.235
	c_{13}	5.0	4.880	–0.1202	2.404	0.347
	c_{14}	7.0	6.956	–0.0443	0.633	0.506
	c_{15}	9.0	8.946	–0.0542	0.602	0.730
	c_{21}	1.5	1.445	–0.0552	3.680	0.286
	c_{22}	4.5	4.550	0.0503	1.118	0.556
	c_{23}	7.5	7.628	0.1277	1.703	0.952
	c_{24}	10.5	10.702	0.2022	1.926	1.572
	c_{25}	13.5	13.741	0.2408	1.784	2.308
	c_{31}	1.5	1.411	–0.0894	5.960	0.251
	c_{32}	4.5	4.619	0.1186	2.636	0.394
	c_{33}	7.5	7.735	0.2347	3.129	0.923
	c_{34}	10.5	10.884	0.3842	3.659	1.627
	c_{35}	13.5	14.034	0.5343	3.958	2.712
	c_{41}	1.0	0.966	–0.0344	3.440	0.325
	c_{42}	5.0	4.944	–0.0563	1.126	0.578
	c_{43}	9.0	8.966	–0.0345	0.383	1.485
	c_{44}	13.0	12.901	–0.0995	0.765	2.887
	c_{45}	17.0	16.941	–0.0588	0.346	4.727
	c_{51}	1.0	0.792	–0.2085	20.850	0.324
	c_{52}	5.0	4.682	–0.3184	6.368	0.592
	c_{53}	9.0	8.617	–0.3833	4.259	1.415
	c_{54}	13.0	12.431	–0.5690	4.377	2.642
	c_{55}	17.0	16.267	–0.7334	4.314	4.471
	c_{61}	1.0	0.928	–0.0723	7.230	0.315
	c_{62}	5.0	4.893	–0.1070	2.140	0.639
	c_{63}	9.0	8.799	–0.2008	2.231	1.182
	c_{64}	13.0	12.758	–0.2418	1.860	2.451

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{65}	17.0	16.780	–0.2202	1.295	4.290
	c_{71}	1.0	0.895	–0.1049	10.490	0.286
	c_{72}	5.0	4.863	–0.1368	2.736	0.534
	c_{73}	9.0	8.826	–0.1741	1.935	1.407
	c_{74}	13.0	12.816	–0.1836	1.412	2.307
	c_{75}	17.0	16.804	–0.1965	1.156	4.436
	c_{81}	1.5	1.281	–0.2193	14.619	0.463
	c_{82}	6.5	6.053	–0.4475	6.884	1.102
	c_{83}	11.5	10.778	–0.7223	6.281	2.910
	c_{84}	16.5	15.409	–1.0909	6.631	5.439
	c_{85}	21.5	20.224	–1.2762	5.936	8.360
	c_{91}	1.5	1.294	–0.2062	13.748	0.493
	c_{92}	6.5	6.112	–0.3880	5.933	0.972
	c_{93}	11.5	10.938	–0.5620	4.887	2.729
	c_{94}	16.5	15.677	–0.8227	4.986	5.033
	c_{95}	21.5	20.521	–0.9795	4.556	8.085
	c_{101}	3.0	2.403	–0.5975	19.917	0.979
	c_{102}	9.0	8.531	–0.4690	5.211	1.382
	c_{103}	15.0	14.472	–0.5282	3.522	2.538
	c_{104}	21.0	20.421	–0.5791	2.757	5.071
	c_{105}	27.0	26.297	–0.7027	2.603	7.515
	Latent class sizes					
	Class 1	0.080	0.084	0.0040	5.000	
	Class 2	0.170	0.159	–0.0110	6.471	
	Class 3	0.250	0.235	–0.0150	6.000	
	Class 4	0.250	0.246	–0.0040	1.600	
	Class 5	0.170	0.168	–0.0020	1.176	
	Class 6	0.080	0.109	0.0290	36.250	

Table B4. Criteria Shifted Up and Down (Strict and Lenient Raters), *d* =Normal 4, BIB, 6 Categories with Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
216	d_1	2.0	1.965	–0.0345	1.725	0.054
216	d_2	3.0	3.043	0.0434	1.447	0.197
216	d_3	3.0	3.007	0.0074	0.247	0.131
216	d_4	4.0	3.859	–0.1405	3.513	0.296
216	d_5	4.0	3.828	–0.1721	4.303	0.295
216	d_6	4.0	3.892	–0.1079	2.698	0.289
216	d_7	4.0	3.840	–0.1605	4.013	0.330
216	d_8	5.0	4.728	–0.2722	5.444	0.356
216	d_9	5.0	4.495	–0.5048	10.096	0.632
216	d_{10}	6.0	5.151	–0.8494	14.157	1.063
	c_{11}	2.0	1.926	–0.0739	3.695	0.193
	c_{12}	4.0	3.944	–0.0564	1.410	0.289
	c_{13}	6.0	5.953	–0.0472	0.787	0.517
	c_{14}	8.0	7.936	–0.0637	0.796	0.736
	c_{15}	10.0	10.007	0.0070	0.070	1.117
	c_{21}	1.5	1.412	–0.0881	5.873	0.318
	c_{22}	4.5	4.564	0.0641	1.424	0.773
	c_{23}	7.5	7.662	0.1620	2.160	1.245
	c_{24}	10.5	10.821	0.3207	3.054	2.550
	c_{25}	13.5	13.976	0.4756	3.523	3.440
	c_{31}	0.5	0.363	–0.1372	27.440	0.274
	c_{32}	3.5	3.536	0.0355	1.014	0.454
	c_{33}	6.5	6.589	0.0886	1.363	0.762
	c_{34}	9.5	9.608	0.1078	1.135	1.289
	c_{35}	12.5	12.756	0.2556	2.045	2.158
	c_{41}	3.0	2.781	–0.2190	7.300	0.550
	c_{42}	7.0	6.750	–0.2501	3.573	1.234
	c_{43}	11.0	10.677	–0.3230	2.936	2.269
	c_{44}	15.0	14.692	–0.3079	2.053	3.969
	c_{45}	19.0	18.590	–0.4103	2.159	6.154
	c_{51}	3.0	2.671	–0.3294	10.980	0.546
	c_{52}	7.0	6.633	–0.3672	5.246	1.191
	c_{53}	11.0	10.588	–0.4119	3.745	2.304
	c_{54}	15.0	14.585	–0.4152	2.768	3.748
	c_{55}	19.0	18.469	–0.5311	2.795	5.197
	c_{61}	1.0	0.884	–0.1157	11.570	0.433
	c_{62}	5.0	4.887	–0.1134	2.268	0.798
	c_{63}	9.0	8.870	–0.1302	1.447	1.420
	c_{64}	13.0	12.888	–0.1125	0.865	2.864

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{65}	17.0	16.884	–0.1158	0.681	4.886
	c_{71}	1.0	0.822	–0.1776	17.760	0.307
	c_{72}	5.0	4.724	–0.2763	5.526	0.764
	c_{73}	9.0	8.754	–0.2461	2.734	1.591
	c_{74}	13.0	12.702	–0.2982	2.293	3.138
	c_{75}	17.0	16.697	–0.3034	1.784	5.708
	c_{81}	3.5	2.938	–0.5625	16.073	0.869
	c_{82}	8.5	7.967	–0.5327	6.267	1.473
	c_{83}	13.5	12.798	–0.7022	5.201	2.663
	c_{84}	18.5	17.612	–0.8878	4.799	4.645
	c_{85}	23.5	22.576	–0.9238	3.931	7.417
	c_{91}	1.5	1.243	–0.2575	17.167	0.565
	c_{92}	6.5	5.978	–0.5217	8.026	1.360
	c_{93}	11.5	10.472	–1.0277	8.937	3.484
	c_{94}	16.5	15.193	–1.3074	7.924	5.901
	c_{95}	21.5	19.883	–1.6167	7.520	10.145
	c_{101}	2.0	1.557	–0.4432	22.161	0.876
	c_{102}	8.0	6.961	–1.0388	12.985	2.187
	c_{103}	14.0	12.394	–1.6059	11.471	4.526
	c_{104}	20.0	17.667	–2.3328	11.664	9.488
	c_{105}	26.0	23.123	–2.8766	11.064	14.970
			Latent class sizes			
	Class 1	0.080	0.085	0.0050	6.250	
	Class 2	0.170	0.165	–0.0050	2.941	
	Class 3	0.250	0.245	–0.0050	2.000	
	Class 4	0.250	0.246	–0.0040	1.600	
	Class 5	0.170	0.168	–0.0020	1.176	
	Class 6	0.080	0.091	0.0110	13.750	

Table B5. Intersection Point Criteria, *d* = Normal 4, Unbalanced, 6 Categories with Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
370	d_1	2.0	2.081	0.0812	4.060	0.043
50	d_2	3.0	2.741	–0.2591	8.637	0.321
200	d_3	3.0	3.054	0.0540	1.800	0.144
140	d_4	4.0	3.710	–0.2903	7.258	0.336
60	d_5	4.0	3.431	–0.5686	14.215	0.555
120	d_6	4.0	4.097	0.0969	2.423	0.299
280	d_7	4.0	4.532	0.5321	13.303	0.511
400	d_8	5.0	4.396	–0.6042	12.084	0.608
230	d_9	5.0	5.094	0.0937	1.874	0.271
310	d_{10}	6.0	4.977	–1.0226	17.043	1.376
	c_{11}	1.0	0.978	–0.0219	2.190	0.098
	c_{12}	3.0	3.078	0.0776	2.587	0.195
	c_{13}	5.0	5.202	0.2018	4.036	0.308
	c_{14}	7.0	7.348	0.3479	4.970	0.519
	c_{15}	9.0	9.446	0.4459	4.954	0.823
	c_{21}	1.5	1.319	–0.1808	12.053	0.865
	c_{22}	4.5	4.062	–0.4379	9.731	1.407
	c_{23}	7.5	6.899	–0.6007	8.009	2.276
	c_{24}	10.5	9.799	–0.7007	6.673	3.481
	c_{25}	13.5	12.574	–0.9260	6.859	5.530
	c_{31}	1.5	1.358	–0.1418	9.453	0.328
	c_{32}	4.5	4.524	0.0242	0.538	0.602
	c_{33}	7.5	7.624	0.1244	1.659	0.953
	c_{34}	10.5	10.809	0.3093	2.946	1.724
	c_{35}	13.5	13.990	0.4903	3.632	2.892
	c_{41}	2.0	1.452	–0.5484	27.420	0.797
	c_{42}	6.0	5.403	–0.5967	9.945	1.464
	c_{43}	10.0	9.334	–0.6665	6.665	2.430
	c_{44}	14.0	13.177	–0.8230	5.879	3.858
	c_{45}	18.0	17.089	–0.9113	5.063	5.952
	c_{51}	2.0	1.430	–0.5696	28.480	1.137
	c_{52}	6.0	4.945	–1.0547	17.578	2.102
	c_{53}	10.0	8.662	–1.3377	13.377	3.655
	c_{54}	14.0	12.209	–1.7915	12.796	6.229
	c_{55}	18.0	15.964	–2.0361	11.312	9.623
	c_{61}	2.0	1.664	–0.3364	16.820	0.662
	c_{62}	6.0	5.977	–0.0229	0.382	1.270
	c_{63}	10.0	10.309	0.3086	3.086	2.399
	c_{64}	14.0	14.546	0.5462	3.901	4.252

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{65}	18.0	18.765	0.7650	4.250	6.402
	c_{71}	2.0	2.084	0.0842	4.210	0.517
	c_{72}	6.0	6.713	0.7132	11.887	1.261
	c_{73}	10.0	11.389	1.3893	13.893	3.895
	c_{74}	14.0	16.071	2.0706	14.790	7.315
	c_{75}	18.0	20.689	2.6892	14.940	12.081
	c_{81}	2.5	1.748	–0.7525	30.101	1.038
	c_{82}	7.5	6.379	–1.1211	14.948	2.029
	c_{83}	12.5	10.976	–1.5237	12.189	4.065
	c_{84}	17.5	15.638	–1.8623	10.642	6.669
	c_{85}	22.5	20.327	–2.1734	9.660	9.710
	c_{91}	2.5	1.995	–0.5052	20.208	0.723
	c_{92}	7.5	7.465	–0.0350	0.467	0.843
	c_{93}	12.5	12.805	0.3045	2.436	2.418
	c_{94}	17.5	18.125	0.6253	3.573	4.416
	c_{95}	22.5	23.380	0.8798	3.910	6.412
	c_{101}	3.0	1.854	–1.1457	38.188	1.874
	c_{102}	9.0	7.309	–1.6912	18.791	4.092
	c_{103}	15.0	12.473	–2.5270	16.847	8.913
	c_{104}	21.0	17.773	–3.2274	15.369	14.923
	c_{105}	27.0	23.025	–3.9753	14.723	22.024
			Latent class sizes			
	Class 1	0.080	0.108	0.0280	35.000	
	Class 2	0.170	0.163	–0.0070	4.118	
	Class 3	0.250	0.248	–0.0020	0.800	
	Class 4	0.250	0.240	–0.0100	4.000	
	Class 5	0.170	0.159	–0.0110	6.471	
	Class 6	0.080	0.083	0.0030	3.750	

Table B6. Criteria Shifted Up (Strict Raters), *d* = Normal 4, Unbalanced, 6 Categories with Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
370	d_1	2.0	2.070	0.0696	3.480	0.032
50	d_2	3.0	2.589	–0.4114	13.713	0.393
200	d_3	3.0	2.936	–0.0642	2.140	0.168
140	d_4	4.0	3.532	–0.4680	11.700	0.451
60	d_5	4.0	3.043	–0.9565	23.913	1.124
120	d_6	4.0	3.659	–0.3406	8.515	0.292
280	d_7	4.0	4.153	0.1531	3.828	0.207
400	d_8	5.0	4.074	–0.9265	18.530	1.106
230	d_9	5.0	4.910	–0.0896	1.792	0.224
310	d_{10}	6.0	4.928	–1.0725	17.875	1.407
	c_{11}	1.0	0.697	–0.3026	30.260	0.196
	c_{12}	3.0	2.834	–0.1664	5.547	0.142
	c_{13}	5.0	4.954	–0.0462	0.924	0.191
	c_{14}	7.0	7.096	0.0961	1.373	0.282
	c_{15}	9.0	9.240	0.2395	2.661	0.438
	c_{21}	1.5	0.629	–0.8710	58.067	1.546
	c_{22}	4.5	3.369	–1.1307	25.127	2.180
	c_{23}	7.5	6.230	–1.2702	16.936	3.252
	c_{24}	10.5	8.986	–1.5139	14.418	5.133
	c_{25}	13.5	11.936	–1.5641	11.586	6.822
	c_{31}	1.5	0.804	–0.6965	46.433	0.855
	c_{32}	4.5	3.977	–0.5231	11.624	1.156
	c_{33}	7.5	7.065	–0.4355	5.807	1.466
	c_{34}	10.5	10.129	–0.3706	3.530	2.310
	c_{35}	13.5	13.332	–0.1676	1.241	3.233
	c_{41}	3.0	1.858	–1.1417	38.057	2.060
	c_{42}	7.0	5.620	–1.3803	19.719	2.969
	c_{43}	11.0	9.319	–1.6812	15.284	4.743
	c_{44}	15.0	13.098	–1.9023	12.682	7.023
	c_{45}	19.0	16.861	–2.1390	11.258	9.307
	c_{51}	3.0	1.433	–1.5672	52.240	3.180
	c_{52}	7.0	4.730	–2.2705	32.436	6.240
	c_{53}	11.0	8.021	–2.9788	27.080	10.537
	c_{54}	15.0	11.461	–3.5392	23.595	15.438
	c_{55}	19.0	14.904	–4.0962	21.559	21.402
	c_{61}	3.0	1.862	–1.1379	37.930	2.020
	c_{62}	7.0	5.706	–1.2941	18.487	2.631
	c_{63}	11.0	9.640	–1.3600	12.364	3.344
	c_{64}	15.0	13.539	–1.4611	9.741	4.321

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{65}	19.0	17.491	–1.5092	7.943	5.933
	c_{71}	3.0	2.114	–0.8865	29.550	1.649
	c_{72}	7.0	6.560	–0.4398	6.283	1.281
	c_{73}	11.0	10.969	–0.0306	0.278	1.756
	c_{74}	15.0	15.289	0.2885	1.923	2.882
	c_{75}	19.0	19.765	0.7650	4.026	4.597
	c_{81}	3.5	1.672	–1.8277	52.220	4.234
	c_{82}	8.5	6.099	–2.4012	28.249	7.364
	c_{83}	13.5	10.395	–3.1048	22.999	11.936
	c_{84}	18.5	14.779	–3.7215	20.116	17.783
	c_{85}	23.5	19.176	–4.3237	18.399	24.450
	c_{91}	3.5	2.073	–1.4273	40.780	2.904
	c_{92}	8.5	7.585	–0.9150	10.765	2.180
	c_{93}	13.5	12.682	–0.8176	6.056	2.680
	c_{94}	18.5	17.905	–0.5955	3.219	3.788
	c_{95}	23.5	23.161	–0.3395	1.445	5.153
	c_{101}	3.0	1.127	–1.8734	62.447	4.245
	c_{102}	9.0	6.527	–2.4731	27.479	7.237
	c_{103}	15.0	11.742	–3.2577	21.718	12.395
	c_{104}	21.0	16.967	–4.0334	19.207	19.638
	c_{105}	27.0	22.436	–4.5639	16.903	26.442
			Latent class sizes			
	Class 1	0.080	0.112	0.0320	40.000	
	Class 2	0.170	0.166	–0.0040	2.352	
	Class 3	0.250	0.249	–0.0010	0.400	
	Class 4	0.250	0.235	–0.0150	6.000	
	Class 5	0.170	0.154	–0.0160	9.411	
	Class 6	0.080	0.083	0.0030	3.750	

Table B7. Criteria Shifted Down (Lenient Raters), *d* =Normal 4, Unbalanced, 6 Categories with Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
370	d_1	2.0	2.076	0.0762	3.810	0.038
50	d_2	3.0	2.727	–0.2730	9.100	0.387
200	d_3	3.0	2.942	–0.0583	1.943	0.127
140	d_4	4.0	3.416	–0.5842	14.605	0.658
60	d_5	4.0	3.127	–0.8728	21.820	1.025
120	d_6	4.0	3.521	–0.4793	11.983	0.455
280	d_7	4.0	4.023	0.0232	0.580	0.200
400	d_8	5.0	4.004	–0.9958	19.916	1.250
230	d_9	5.0	4.988	–0.0119	0.238	0.288
310	d_{10}	6.0	4.923	–1.0773	17.955	1.452
	c_{11}	1.0	1.135	0.1347	13.470	0.160
	c_{12}	3.0	3.262	0.2621	8.737	0.243
	c_{13}	5.0	5.397	0.3968	7.936	0.437
	c_{14}	7.0	7.560	0.5599	7.999	0.749
	c_{15}	9.0	9.701	0.7010	7.789	1.102
	c_{21}	1.5	1.175	–0.3254	21.693	1.348
	c_{22}	4.5	4.186	–0.3139	6.976	1.397
	c_{23}	7.5	7.093	–0.4070	5.427	2.220
	c_{24}	10.5	10.053	–0.4466	4.253	4.479
	c_{25}	13.5	13.112	–0.3884	2.877	6.658
	c_{31}	1.5	1.470	–0.0298	1.987	0.342
	c_{32}	4.5	4.547	0.0472	1.049	0.536
	c_{33}	7.5	7.605	0.1046	1.395	0.967
	c_{34}	10.5	10.693	0.1927	1.835	1.722
	c_{35}	13.5	13.822	0.3218	2.384	2.686
	c_{41}	1.0	0.832	–0.1683	16.830	0.587
	c_{42}	5.0	4.450	–0.5502	11.004	1.291
	c_{43}	9.0	8.044	–0.9564	10.627	2.689
	c_{44}	13.0	11.689	–1.3110	10.085	4.835
	c_{45}	17.0	15.559	–1.4413	8.478	8.177
	c_{51}	1.0	0.458	–0.5416	54.160	1.189
	c_{52}	5.0	3.906	–1.0939	21.878	2.321
	c_{53}	9.0	7.370	–1.6303	18.114	4.596
	c_{54}	13.0	10.938	–2.0621	15.862	7.661
	c_{55}	17.0	14.343	–2.6567	15.628	12.610
	c_{61}	1.0	0.651	–0.3495	34.950	0.640
	c_{62}	5.0	4.515	–0.4853	9.706	1.190
	c_{63}	9.0	8.272	–0.7285	8.094	1.971
	c_{64}	13.0	12.075	–0.9252	7.117	3.397

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{65}	17.0	16.037	–0.9626	5.662	5.690
	c_{71}	1.0	1.037	0.0374	3.740	0.305
	c_{72}	5.0	5.302	0.3023	6.046	0.859
	c_{73}	9.0	9.534	0.5336	5.929	2.026
	c_{74}	13.0	13.695	0.6947	5.344	3.735
	c_{75}	17.0	18.080	1.0804	6.355	6.005
	c_{81}	1.5	1.263	–0.2369	15.792	0.354
	c_{82}	6.5	5.517	–0.9827	15.118	1.597
	c_{83}	11.5	9.811	–1.6887	14.684	4.467
	c_{84}	16.5	14.001	–2.4987	15.190	8.803
	c_{85}	21.5	18.371	–3.1286	14.552	13.892
	c_{91}	1.5	1.480	–0.0202	1.347	0.537
	c_{92}	6.5	6.938	0.4380	6.697	1.472
	c_{93}	11.5	12.034	0.5344	4.647	2.331
	c_{94}	16.5	17.231	0.7306	4.428	4.401
	c_{95}	21.5	22.743	1.2428	5.780	7.603
	c_{101}	3.0	2.257	–0.7428	24.760	1.087
	c_{102}	9.0	7.614	–1.3861	15.401	3.045
	c_{103}	15.0	12.952	–2.0480	13.653	6.134
	c_{104}	21.0	18.158	–2.8417	13.532	11.917
	c_{105}	27.0	23.476	–3.5245	13.054	18.341
			Latent class sizes			
	Class 1	0.080	0.080	0.0000	0.000	
	Class 2	0.170	0.156	–0.0140	8.235	
	Class 3	0.250	0.234	–0.0160	6.400	
	Class 4	0.250	0.248	–0.0020	0.800	
	Class 5	0.170	0.165	–0.0050	2.941	
	Class 6	0.080	0.116	0.0360	45.000	

Table B8. Criteria Shifted Up and Down (Strict and Lenient Raters), *d* =Normal 4, Unbalanced, 6 Categories with Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
370	d_1	2.0	2.092	0.0918	4.590	0.049
50	d_2	3.0	2.603	–0.3966	13.220	0.434
200	d_3	3.0	2.987	–0.0135	0.450	0.159
140	d_4	4.0	3.579	–0.4213	10.533	0.473
60	d_5	4.0	3.181	–0.8188	20.470	0.922
120	d_6	4.0	3.970	–0.0296	0.740	0.278
280	d_7	4.0	4.322	0.3220	8.050	0.285
400	d_8	5.0	4.224	–0.7756	15.512	0.766
230	d_9	5.0	4.795	–0.2047	4.094	0.219
310	d_{10}	6.0	4.635	–1.3655	22.758	2.085
	c_{11}	2.0	2.159	0.1591	7.955	0.162
	c_{12}	4.0	4.282	0.2822	7.055	0.315
	c_{13}	6.0	6.400	0.4002	6.670	0.547
	c_{14}	8.0	8.522	0.5223	6.529	0.838
	c_{15}	10.0	10.636	0.6356	6.356	1.176
	c_{21}	1.5	1.257	–0.2426	16.173	0.757
	c_{22}	4.5	3.932	–0.5682	12.627	1.457
	c_{23}	7.5	6.722	–0.7780	10.373	2.924
	c_{24}	10.5	9.441	–1.0593	10.089	4.520
	c_{25}	13.5	12.295	–1.2049	8.925	7.109
	c_{31}	0.5	0.456	–0.0437	8.740	0.238
	c_{32}	3.5	3.548	0.0476	1.360	0.352
	c_{33}	6.5	6.622	0.1218	1.874	0.777
	c_{34}	9.5	9.684	0.1835	1.932	1.563
	c_{35}	12.5	12.795	0.2953	2.362	2.598
	c_{41}	3.0	2.587	–0.4127	13.757	0.700
	c_{42}	7.0	6.292	–0.7084	10.120	1.730
	c_{43}	11.0	9.986	–1.0141	9.219	3.413
	c_{44}	15.0	13.717	–1.2827	8.551	5.282
	c_{45}	19.0	17.561	–1.4390	7.574	7.894
	c_{51}	3.0	2.030	–0.9700	32.333	1.814
	c_{52}	7.0	5.461	–1.5392	21.989	3.337
	c_{53}	11.0	8.889	–2.1114	19.195	6.596
	c_{54}	15.0	12.212	–2.7876	18.584	11.243
	c_{55}	19.0	15.682	–3.3180	17.463	15.766
	c_{61}	1.0	0.878	–0.1216	12.160	0.755
	c_{62}	5.0	5.060	0.0602	1.204	1.033
	c_{63}	9.0	9.109	0.1088	1.209	2.142
	c_{64}	13.0	13.160	0.1603	1.233	3.327

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{65}	17.0	17.503	0.5027	2.957	6.266
	c_{71}	1.0	1.101	0.1006	10.060	0.547
	c_{72}	5.0	5.482	0.4820	9.640	0.766
	c_{73}	9.0	9.986	0.9863	10.959	2.128
	c_{74}	13.0	14.345	1.3453	10.348	3.859
	c_{75}	17.0	18.944	1.9442	11.436	7.524
	c_{81}	3.5	2.741	–0.7592	21.691	0.969
	c_{82}	8.5	7.144	–1.3563	15.956	2.448
	c_{83}	13.5	11.555	–1.9450	14.407	5.154
	c_{84}	18.5	15.965	–2.5346	13.700	8.703
	c_{85}	23.5	20.407	–3.0930	13.162	13.098
	c_{91}	1.5	1.489	–0.0115	0.769	0.478
	c_{92}	6.5	6.555	0.0551	0.848	0.735
	c_{93}	11.5	11.563	0.0633	0.551	1.411
	c_{94}	16.5	16.302	–0.1981	1.201	2.075
	c_{95}	21.5	21.517	0.0174	0.081	3.927
	c_{101}	2.0	1.496	–0.5039	25.197	0.673
	c_{102}	8.0	6.521	–1.4792	18.490	3.009
	c_{103}	14.0	11.411	–2.5894	18.496	8.324
	c_{104}	20.0	16.186	–3.8137	19.068	17.402
	c_{105}	26.0	21.230	–4.7696	18.345	28.056
			Latent class sizes			
	Class 1	0.080	0.083	0.0030	3.750	
	Class 2	0.170	0.161	–0.0090	5.294	
	Class 3	0.250	0.243	–0.0070	2.800	
	Class 4	0.250	0.248	–0.0020	0.800	
	Class 5	0.170	0.167	–0.0030	1.765	
	Class 6	0.080	0.099	0.0190	23.750	

Table B9. Intersection Point Criteria, *d* = Normal 2, BIB, 6 Categories with Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
216	d_1	0.50	0.456	–0.0440	8.800	0.011
216	d_2	1.00	0.936	–0.0639	6.390	0.026
216	d_3	1.00	0.915	–0.0847	8.470	0.029
216	d_4	2.00	1.911	–0.0890	4.450	0.157
216	d_5	2.00	1.859	–0.1410	7.050	0.122
216	d_6	2.00	1.839	–0.1609	8.045	0.134
216	d_7	2.00	1.896	–0.1036	5.180	0.162
216	d_8	3.00	2.671	–0.3286	10.953	0.379
216	d_9	3.00	2.777	–0.2235	7.450	0.327
216	d_{10}	4.00	3.305	–0.6945	17.363	0.836
	c_{11}	0.25	0.153	–0.0972	38.880	0.095
	c_{12}	0.75	0.637	–0.1126	15.013	0.101
	c_{13}	1.25	1.131	–0.1189	9.512	0.113
	c_{14}	1.75	1.641	–0.1095	6.257	0.107
	c_{15}	2.25	2.158	–0.0923	4.102	0.106
	c_{21}	0.50	0.321	–0.1789	35.780	0.171
	c_{22}	1.50	1.334	–0.1659	11.060	0.186
	c_{23}	2.50	2.349	–0.1511	6.044	0.210
	c_{24}	3.50	3.368	–0.1324	3.783	0.258
	c_{25}	4.50	4.385	–0.1153	2.562	0.281
	c_{31}	0.50	0.300	–0.1999	39.980	0.206
	c_{32}	1.50	1.310	–0.1898	12.653	0.200
	c_{33}	2.50	2.302	–0.1983	7.932	0.234
	c_{34}	3.50	3.327	–0.1726	4.931	0.259
	c_{35}	4.50	4.319	–0.1814	4.031	0.328
	c_{41}	1.00	0.578	–0.4221	42.210	0.432
	c_{42}	3.00	2.718	–0.2823	9.410	0.772
	c_{43}	5.00	4.870	–0.1301	2.602	1.366
	c_{44}	7.00	6.968	–0.0319	0.456	1.974
	c_{45}	9.00	9.002	0.0018	0.020	2.740
	c_{51}	1.00	0.551	–0.4491	44.910	0.505
	c_{52}	3.00	2.612	–0.3883	12.943	0.679
	c_{53}	5.00	4.675	–0.3252	6.504	0.860
	c_{54}	7.00	6.739	–0.2607	3.724	1.241
	c_{55}	9.00	8.786	–0.2144	2.382	1.758
	c_{61}	1.00	0.523	–0.4768	47.680	0.435
	c_{62}	3.00	2.585	–0.4151	13.837	0.592
	c_{63}	5.00	4.635	–0.3650	7.300	0.884
	c_{64}	7.00	6.660	–0.3398	4.854	1.387

c_{65}	9.00	8.680	–0.3198	3.553	1.942
c_{71}	1.00	0.597	–0.4026	40.260	0.557
c_{72}	3.00	2.688	–0.3116	10.387	0.886
c_{73}	5.00	4.774	–0.2261	4.522	1.239
c_{74}	7.00	6.869	–0.1308	1.869	1.734
c_{75}	9.00	8.920	–0.0800	0.888	2.647
c_{81}	1.50	0.705	–0.7950	53.001	1.005
c_{82}	4.50	3.697	–0.8027	17.838	1.638
c_{83}	7.50	6.707	–0.7932	10.575	2.520
c_{84}	10.50	9.711	–0.7888	7.512	3.951
c_{85}	13.50	12.734	–0.7657	5.672	6.188
c_{91}	1.50	0.751	–0.7495	49.965	1.057
c_{92}	4.50	3.852	–0.6480	14.400	1.485
c_{93}	7.50	6.977	–0.5232	6.976	2.288
c_{94}	10.50	10.039	–0.4615	4.395	3.711
c_{95}	13.50	13.162	–0.3382	2.505	5.938
c_{101}	2.00	0.841	–1.1594	57.970	1.843
c_{102}	6.00	4.572	–1.4284	23.806	3.410
c_{103}	10.00	8.255	–1.7452	17.452	5.727
c_{104}	14.00	12.074	–1.9264	13.760	8.593
c_{105}	18.00	15.784	–2.2165	12.314	11.845

Latent class sizes

Class 1	0.080	0.122	0.0420	52.500
Class 2	0.170	0.144	–0.0260	15.294
Class 3	0.250	0.230	–0.0200	8.000
Class 4	0.250	0.232	–0.0180	7.200
Class 5	0.170	0.148	–0.0220	12.941
Class 6	0.080	0.123	0.0430	53.750

Table B10. Intersection Point Criteria, *d* = Normal 2, Unbalanced, 6 Categories with Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
370	d_1	0.50	0.483	–0.0167	3.340	0.009
50	d_2	1.00	0.972	–0.0275	2.750	0.117
200	d_3	1.00	0.915	–0.0855	8.550	0.027
140	d_4	2.00	1.624	–0.3763	18.815	0.233
60	d_5	2.00	1.552	–0.4480	22.400	0.343
120	d_6	2.00	1.968	–0.0316	1.580	0.108
280	d_7	2.00	2.215	0.2149	10.745	0.145
400	d_8	3.00	2.173	–0.8266	27.553	0.794
230	d_9	3.00	3.080	0.0795	2.650	0.212
310	d_{10}	4.00	3.085	–0.9147	22.868	1.063
	c_{11}	0.25	0.197	–0.0532	21.280	0.055
	c_{12}	0.75	0.711	–0.0385	5.133	0.056
	c_{13}	1.25	1.223	–0.0270	2.160	0.058
	c_{14}	1.75	1.731	–0.0187	1.069	0.066
	c_{15}	2.25	2.233	–0.0166	0.738	0.068
	c_{21}	0.50	0.306	–0.1936	38.720	0.468
	c_{22}	1.50	1.441	–0.0592	3.947	0.710
	c_{23}	2.50	2.469	–0.0314	1.256	0.907
	c_{24}	3.50	3.532	0.0320	0.914	1.309
	c_{25}	4.50	4.598	0.0977	2.171	1.727
	c_{31}	0.50	0.274	–0.2264	45.280	0.162
	c_{32}	1.50	1.254	–0.2457	16.380	0.191
	c_{33}	2.50	2.261	–0.2391	9.564	0.235
	c_{34}	3.50	3.299	–0.2010	5.743	0.273
	c_{35}	4.50	4.326	–0.1736	3.858	0.321
	c_{41}	1.00	0.336	–0.6641	66.410	0.816
	c_{42}	3.00	2.149	–0.8506	28.353	1.248
	c_{43}	5.00	4.068	–0.9322	18.644	1.597
	c_{44}	7.00	5.917	–1.0835	15.479	2.324
	c_{45}	9.00	7.771	–1.2293	13.659	3.043
	c_{51}	1.00	0.259	–0.7407	74.070	1.111
	c_{52}	3.00	2.108	–0.8923	29.743	1.551
	c_{53}	5.00	3.910	–1.0904	21.808	2.287
	c_{54}	7.00	5.739	–1.2607	18.010	3.268
	c_{55}	9.00	7.596	–1.4045	15.606	4.638
	c_{61}	1.00	0.580	–0.4200	42.000	0.625
	c_{62}	3.00	2.791	–0.2086	6.953	0.641
	c_{63}	5.00	4.908	–0.0921	1.842	0.818
	c_{64}	7.00	7.097	0.0967	1.381	1.265

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{65}	9.00	9.302	0.3023	3.359	2.063
	c_{71}	1.00	0.792	–0.2082	20.820	0.303
	c_{72}	3.00	3.195	0.1945	6.483	0.517
	c_{73}	5.00	5.535	0.5347	10.694	1.101
	c_{74}	7.00	7.939	0.9393	13.418	2.177
	c_{75}	9.00	10.284	1.2836	14.262	3.566
	c_{81}	1.50	0.411	–1.0890	72.601	1.404
	c_{82}	4.50	2.933	–1.5672	34.826	2.874
	c_{83}	7.50	5.432	–2.0677	27.569	4.994
	c_{84}	10.50	7.905	–2.5952	24.717	7.920
	c_{85}	13.50	10.445	–3.0550	22.629	11.413
	c_{91}	1.50	0.850	–0.6498	43.322	0.877
	c_{92}	4.50	4.365	–0.1349	2.999	0.918
	c_{93}	7.50	7.712	0.2124	2.832	1.588
	c_{94}	10.50	11.146	0.6460	6.152	2.924
	c_{95}	13.50	14.509	1.0087	7.472	4.758
	c_{101}	2.00	0.693	–1.3074	65.370	2.132
	c_{102}	6.00	4.258	–1.7419	29.031	4.097
	c_{103}	10.00	7.722	–2.2782	22.782	6.768
	c_{104}	14.00	11.193	–2.8068	20.048	10.737
	c_{105}	18.00	14.785	–3.2146	17.859	14.745
			Latent class sizes			
	Class 1	0.080	0.123	0.0430	53.750	
	Class 2	0.170	0.139	–0.0310	18.235	
	Class 3	0.250	0.237	–0.0130	5.200	
	Class 4	0.250	0.235	–0.0150	6.000	
	Class 5	0.170	0.141	–0.0290	17.059	
	Class 6	0.080	0.124	0.0440	55.000	

Table B11. Intersection Point Criteria, *d* = Normal 4, BIB, 6 Categories with Non-normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
216	d_1	2.0	1.918	–0.0820	4.100	0.069
216	d_2	3.0	2.934	–0.0662	2.207	0.176
216	d_3	3.0	2.974	–0.0261	0.870	0.137
216	d_4	4.0	3.876	–0.1245	3.113	0.363
216	d_5	4.0	3.964	–0.0365	0.913	0.358
216	d_6	4.0	3.967	–0.0330	0.825	0.327
216	d_7	4.0	3.881	–0.1193	2.983	0.305
216	d_8	5.0	4.674	–0.3256	6.512	0.568
216	d_9	5.0	4.739	–0.2613	5.226	0.548
216	d_{10}	6.0	5.272	–0.7276	12.127	0.911
	c_{11}	1.0	0.757	–0.2434	24.340	0.507
	c_{12}	3.0	2.746	–0.2542	8.473	0.476
	c_{13}	5.0	4.776	–0.2239	4.478	0.554
	c_{14}	7.0	6.835	–0.1655	2.364	0.665
	c_{15}	9.0	8.860	–0.1404	1.560	1.022
	c_{21}	1.5	1.064	–0.4361	29.073	0.860
	c_{22}	4.5	4.174	–0.3258	7.240	1.033
	c_{23}	7.5	7.261	–0.2388	3.184	1.357
	c_{24}	10.5	10.412	–0.0881	0.839	1.890
	c_{25}	13.5	13.589	0.0893	0.661	3.035
	c_{31}	1.5	1.147	–0.3533	23.553	0.924
	c_{32}	4.5	4.311	–0.1895	4.211	0.994
	c_{33}	7.5	7.381	–0.1190	1.587	1.112
	c_{34}	10.5	10.584	0.0842	0.802	1.635
	c_{35}	13.5	13.849	0.3488	2.584	2.726
	c_{41}	2.0	1.453	–0.5466	27.330	1.293
	c_{42}	6.0	5.546	–0.4542	7.570	2.303
	c_{43}	10.0	9.637	–0.3630	3.630	2.627
	c_{44}	14.0	13.849	–0.1514	1.081	3.933
	c_{45}	18.0	18.019	0.0189	0.105	6.741
	c_{51}	2.0	1.547	–0.4534	22.670	1.374
	c_{52}	6.0	5.695	–0.3052	5.087	1.929
	c_{53}	10.0	9.850	–0.1500	1.500	2.739
	c_{54}	14.0	14.170	0.1701	1.215	4.230
	c_{55}	18.0	18.402	0.4022	2.234	7.631
	c_{61}	2.0	1.422	–0.5783	28.915	1.547
	c_{62}	6.0	5.661	–0.3387	5.645	2.076
	c_{63}	10.0	9.846	–0.1537	1.537	2.606
	c_{64}	14.0	14.177	0.1771	1.265	4.118

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{65}	18.0	18.408	0.4081	2.267	6.568
	c_{71}	2.0	1.419	–0.5814	29.070	1.493
	c_{72}	6.0	5.515	–0.4854	8.090	2.044
	c_{73}	10.0	9.633	–0.3668	3.668	2.533
	c_{74}	14.0	13.993	–0.0074	0.053	3.913
	c_{75}	18.0	18.144	0.1438	0.799	5.577
	c_{81}	2.5	1.889	–0.6107	24.429	1.721
	c_{82}	7.5	6.549	–0.9510	12.679	3.670
	c_{83}	12.5	11.639	–0.8612	6.890	4.308
	c_{84}	17.5	16.912	–0.5882	3.361	5.923
	c_{85}	22.5	21.807	–0.6934	3.082	9.906
	c_{91}	2.5	1.869	–0.6311	25.244	1.894
	c_{92}	7.5	6.610	–0.8896	11.861	3.702
	c_{93}	12.5	11.725	–0.7747	6.197	4.187
	c_{94}	17.5	17.104	–0.3960	2.263	6.890
	c_{95}	22.5	22.126	–0.3741	1.663	10.884
	c_{101}	3.0	2.036	–0.9642	32.141	2.816
	c_{102}	9.0	7.140	–1.8602	20.669	5.992
	c_{103}	15.0	13.053	–1.9473	12.982	6.939
	c_{104}	21.0	19.232	–1.7680	8.419	8.725
	c_{105}	27.0	24.785	–2.2151	8.204	14.222
			Latent class sizes			
	Class 1	0.030	0.034	0.0040	13.333	
	Class 2	0.030	0.042	0.0120	40.000	
	Class 3	0.400	0.390	–0.0100	2.500	
	Class 4	0.400	0.374	–0.0260	6.500	
	Class 5	0.100	0.109	0.0090	9.000	
	Class 6	0.040	0.050	0.0100	25.000	

Table B12. Intersection Point Criteria, *d* = Normal 4, BIB, 4 Categories with Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
216	d_1	2.0	2.016	0.0164	0.820	0.083
216	d_2	3.0	2.991	–0.0093	0.310	0.173
216	d_3	3.0	3.012	0.0121	0.403	0.206
216	d_4	4.0	4.044	0.0442	1.105	0.313
216	d_5	4.0	4.051	0.0514	1.285	0.369
216	d_6	4.0	4.026	0.0262	0.655	0.386
216	d_7	4.0	4.106	0.1061	2.653	0.385
216	d_8	5.0	5.083	0.0833	1.666	0.592
216	d_9	5.0	5.140	0.1396	2.792	0.641
216	d_{10}	6.0	5.685	–0.3148	5.247	0.703
	c_{11}	1.0	0.974	–0.0264	2.640	0.126
	c_{12}	3.0	3.033	0.0330	1.100	0.194
	c_{13}	5.0	5.119	0.1191	2.382	0.414
	c_{21}	1.5	1.455	–0.0448	2.987	0.262
	c_{22}	4.5	4.532	0.0320	0.711	0.575
	c_{23}	7.5	7.632	0.1319	1.759	0.977
	c_{31}	1.5	1.443	–0.0572	3.813	0.220
	c_{32}	4.5	4.519	0.0192	0.427	0.570
	c_{33}	7.5	7.641	0.1410	1.880	1.067
	c_{41}	2.0	1.820	–0.1804	9.020	0.399
	c_{42}	6.0	6.101	0.1006	1.677	0.840
	c_{43}	10.0	10.326	0.3257	3.257	2.265
	c_{51}	2.0	1.904	–0.0962	4.810	0.399
	c_{52}	6.0	6.096	0.0955	1.592	0.948
	c_{53}	10.0	10.300	0.3000	3.000	2.199
	c_{61}	2.0	1.883	–0.1171	5.855	0.364
	c_{62}	6.0	6.011	0.0110	0.183	0.996
	c_{63}	10.0	10.205	0.2050	2.050	2.550
	c_{71}	2.0	1.947	–0.0534	2.668	0.388
	c_{72}	6.0	6.182	0.1816	3.026	0.970
	c_{73}	10.0	10.463	0.4629	4.629	2.334
	c_{81}	2.5	2.262	–0.2376	9.503	0.750
	c_{82}	7.5	7.613	0.1134	1.512	1.687
	c_{83}	12.5	13.010	0.5097	4.078	4.129
	c_{91}	2.5	2.471	–0.0289	1.156	0.927
	c_{92}	7.5	7.718	0.2175	2.900	1.870
	c_{93}	12.5	13.275	0.7748	6.198	5.297
	c_{101}	3.0	2.378	–0.6224	20.746	1.010
	c_{102}	9.0	8.521	–0.4794	5.327	1.772

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{103}	15.0	14.707	–0.2927	1.951	4.155
			Latent class sizes			
	Class 1	0.170	0.180	0.0100	5.882	
	Class 2	0.330	0.320	–0.0100	3.030	
	Class 3	0.330	0.316	–0.0140	4.242	
	Class 4	0.170	0.183	0.0130	7.647	

Table B13. Intersection Point Criteria, *d* = Normal 4, BIB, 4 Categories with Non-Normal Class Sizes

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
216	d_1	2.0	1.835	–0.1649	8.245	0.141
216	d_2	3.0	2.786	–0.2141	7.137	0.217
216	d_3	3.0	2.801	–0.1993	6.643	0.291
216	d_4	4.0	3.818	–0.1818	4.545	0.529
216	d_5	4.0	3.647	–0.3530	8.825	0.450
216	d_6	4.0	3.757	–0.2426	6.065	0.547
216	d_7	4.0	3.716	–0.2839	7.098	0.491
216	d_8	5.0	4.791	–0.2093	4.186	0.801
216	d_9	5.0	4.838	–0.1623	3.246	0.589
216	d_{10}	6.0	5.505	–0.4952	8.253	1.049
	c_{11}	1.0	0.694	–0.3062	30.620	0.306
	c_{12}	3.0	2.760	–0.2404	8.013	0.332
	c_{13}	5.0	4.827	–0.1728	3.456	0.441
	c_{21}	1.5	0.973	–0.5273	35.153	0.585
	c_{22}	4.5	4.191	–0.3089	6.864	0.578
	c_{23}	7.5	7.363	–0.1368	1.824	0.951
	c_{31}	1.5	1.032	–0.4678	31.187	0.600
	c_{32}	4.5	4.227	–0.2727	6.060	0.728
	c_{33}	7.5	7.388	–0.1116	1.488	1.443
	c_{41}	2.0	1.295	–0.7052	35.260	1.136
	c_{42}	6.0	5.716	–0.2836	4.727	1.256
	c_{43}	10.0	10.101	0.1006	1.006	2.427
	c_{51}	2.0	1.199	–0.8008	40.040	1.134
	c_{52}	6.0	5.423	–0.5773	9.622	1.223
	c_{53}	10.0	9.703	–0.2975	2.975	2.003
	c_{61}	2.0	1.312	–0.6879	34.395	1.124
	c_{62}	6.0	5.642	–0.3578	5.963	1.350
	c_{63}	10.0	9.964	–0.0357	0.357	2.786
	c_{71}	2.0	1.291	–0.7092	35.459	1.048
	c_{72}	6.0	5.584	–0.4157	6.928	1.271
	c_{73}	10.0	9.909	–0.0915	0.915	2.595
	c_{81}	2.5	1.482	–1.0180	40.721	2.078
	c_{82}	7.5	7.136	–0.3645	4.860	1.985
	c_{83}	12.5	12.850	0.3495	2.796	5.194
	c_{91}	2.5	1.515	–0.9846	39.382	1.882
	c_{92}	7.5	7.209	–0.2913	3.884	1.454
	c_{93}	12.5	12.984	0.4838	3.870	3.931
	c_{101}	3.0	1.429	–1.5707	52.357	3.672
	c_{102}	9.0	8.209	–0.7906	8.784	2.704

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{103}	15.0	15.128	0.1278	0.852	5.738
			Latent class sizes			
	Class 1	0.070	0.106	0.0360	51.429	
	Class 2	0.430	0.397	–0.0330	7.674	
	Class 3	0.430	0.393	–0.0370	8.605	
	Class 4	0.070	0.104	0.0340	48.571	

Table B14. Intersection Point Criteria, *d* = Normal 4, BIB, 4 Categories with Non-Normal Class Sizes (Shift in Density, First Scoring Occasion)

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
216	d_1	2.0	1.717	–0.2832	14.160	0.254
216	d_2	3.0	2.553	–0.4467	14.890	0.468
216	d_3	3.0	2.577	–0.4226	14.087	0.533
216	d_4	4.0	3.571	–0.4287	10.718	0.825
216	d_5	4.0	3.579	–0.4211	10.528	0.708
216	d_6	4.0	3.686	–0.3144	7.860	0.732
216	d_7	4.0	3.558	–0.4421	11.053	0.779
216	d_8	5.0	4.514	–0.4857	9.714	1.042
216	d_9	5.0	4.430	–0.5701	11.402	1.172
216	d_{10}	6.0	5.246	–0.7542	12.570	1.329
	c_{11}	1.0	0.605	–0.3948	39.480	0.429
	c_{12}	3.0	2.653	–0.3470	11.567	0.464
	c_{13}	5.0	4.712	–0.2881	5.762	0.559
	c_{21}	1.5	0.898	–0.6016	40.107	0.730
	c_{22}	4.5	3.987	–0.5131	11.402	0.792
	c_{23}	7.5	7.146	–0.3542	4.723	0.957
	c_{31}	1.5	0.925	–0.5748	38.320	0.711
	c_{32}	4.5	4.050	–0.4503	10.007	0.914
	c_{33}	7.5	7.175	–0.3254	4.339	1.388
	c_{41}	2.0	1.240	–0.7601	38.005	1.375
	c_{42}	6.0	5.600	–0.4002	6.670	1.573
	c_{43}	10.0	9.917	–0.0832	0.832	3.025
	c_{51}	2.0	1.232	–0.7677	38.385	1.208
	c_{52}	6.0	5.545	–0.4551	7.585	1.346
	c_{53}	10.0	10.049	0.0485	0.485	2.546
	c_{61}	2.0	1.373	–0.6267	31.335	1.135
	c_{62}	6.0	5.748	–0.2521	4.202	1.301
	c_{63}	10.0	10.412	0.4123	4.123	3.712
	c_{71}	2.0	1.181	–0.8189	40.945	1.368
	c_{72}	6.0	5.556	–0.4439	7.398	1.382
	c_{73}	10.0	9.989	–0.0115	0.115	2.666
	c_{81}	2.5	1.406	–1.0937	43.747	2.057
	c_{82}	7.5	7.085	–0.4148	5.531	1.989
	c_{83}	12.5	12.935	0.4347	3.478	5.555
	c_{91}	2.5	1.334	–1.1663	46.652	2.352
	c_{92}	7.5	6.899	–0.6014	8.018	2.155
	c_{93}	12.5	12.617	0.1167	0.934	4.728
	c_{101}	3.0	1.322	–1.6778	55.925	3.806
	c_{102}	9.0	8.247	–0.7532	8.369	2.489

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{103}	15.0	15.236	0.2364	1.576	5.776
			Latent class sizes			
	Class 1	0.070	0.109	0.0390	55.714	
	Class 2	0.500	0.446	–0.0540	10.800	
	Class 3	0.400	0.340	–0.0600	15.000	
	Class 4	0.030	0.106	0.0760	253.333	

Table B15. Intersection Point Criteria, *d* = Normal 4, BIB, 4 Categories with Non-Normal Class Sizes (Shift in Density, Second Scoring Occasion)

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
			Rater parameters			
216	d_1	2.0	1.737	–0.2628	13.140	0.266
216	d_2	3.0	2.568	–0.4317	14.390	0.530
216	d_3	3.0	2.601	–0.3987	13.290	0.456
216	d_4	4.0	3.687	–0.3132	7.830	0.696
216	d_5	4.0	3.503	–0.4971	12.428	0.881
216	d_6	4.0	3.609	–0.3914	9.785	0.601
216	d_7	4.0	3.638	–0.3616	9.040	0.718
216	d_8	5.0	4.488	–0.5115	10.230	1.233
216	d_9	5.0	4.405	–0.5951	11.902	0.881
216	d_{10}	6.0	5.479	–0.5213	8.688	1.501
	c_{11}	1.0	0.480	–0.5203	52.030	0.742
	c_{12}	3.0	2.573	–0.4265	14.217	0.753
	c_{13}	5.0	4.585	–0.4149	8.298	0.984
	c_{21}	1.5	0.662	–0.8379	55.860	1.485
	c_{22}	4.5	3.717	–0.7831	17.402	1.547
	c_{23}	7.5	6.789	–0.7110	9.480	2.019
	c_{31}	1.5	0.636	–0.8638	57.587	1.437
	c_{32}	4.5	3.774	–0.7255	16.122	1.363
	c_{33}	7.5	6.915	–0.5855	7.807	1.671
	c_{41}	2.0	0.863	–1.1374	56.870	2.845
	c_{42}	6.0	5.339	–0.6611	11.018	2.043
	c_{43}	10.0	9.825	–0.1751	1.751	3.096
	c_{51}	2.0	0.736	–1.2645	63.225	2.795
	c_{52}	6.0	5.080	–0.9199	15.332	2.522
	c_{53}	10.0	9.421	–0.5792	5.792	3.452
	c_{61}	2.0	0.781	–1.2191	60.955	2.873
	c_{62}	6.0	5.277	–0.7231	12.052	1.773
	c_{63}	10.0	9.593	–0.4065	4.065	2.325
	c_{71}	2.0	0.719	–1.2807	64.035	2.845
	c_{72}	6.0	5.278	–0.7225	12.042	2.109
	c_{73}	10.0	9.736	–0.2643	2.643	3.647
	c_{81}	2.5	0.724	–1.7758	71.032	5.271
	c_{82}	7.5	6.568	–0.9323	12.431	3.357
	c_{83}	12.5	12.177	–0.3235	2.588	5.785
	c_{91}	2.5	0.611	–1.8889	75.557	5.112
	c_{92}	7.5	6.443	–1.0568	14.090	2.820
	c_{93}	12.5	11.992	–0.5080	4.064	4.386
	c_{101}	3.0	0.667	–2.3330	77.767	7.808
	c_{102}	9.0	7.944	–1.0562	11.736	4.395

Size	Parameter	Value	Estimate	Bias	%Bias	MSE
	c_{103}	15.0	15.020	0.0202	0.134	8.344
			Latent class sizes			
	Class 1	0.030	0.096	0.0660	220.000	
	Class 2	0.400	0.346	–0.0540	13.500	
	Class 3	0.500	0.446	–0.0540	10.800	
	Class 4	0.070	0.111	0.0410	58.571	

Appendix C

Evaluation of the Estimated Standard Errors for *d* and the Latent Class Sizes

Table C1. Intersection Point Criteria, *d* = Normal 4, BIB, 6 Categories with Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
216	d_1	0.198	0.214	0.016	8.026
216	d_2	0.391	0.379	–0.012	3.069
216	d_3	0.397	0.370	–0.027	6.871
216	d_4	0.550	0.576	0.026	4.784
216	d_5	0.503	0.568	0.065	12.947
216	d_6	0.473	0.573	0.099	21.006
216	d_7	0.521	0.561	0.040	7.601
216	d_8	0.635	0.776	0.140	22.107
216	d_9	0.613	0.763	0.150	24.445
216	d_{10}	0.717	0.892	0.175	24.456
	Class Size 1	0.012	0.014	0.002	12.903
	Class Size 2	0.019	0.019	0.000	1.604
	Class Size 3	0.021	0.023	0.002	10.577
	Class Size 4	0.020	0.023	0.003	15.578
	Class Size 5	0.017	0.020	0.003	14.943
	Class Size 6	0.012	0.015	0.003	21.951

Table C2. Criteria Shifted Up (Strict Raters), *d* = Normal 4, BIB, 6 Categories with Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
216	d_1	0.224	0.225	0.001	0.491
216	d_2	0.377	0.391	0.014	3.604
216	d_3	0.400	0.390	–0.010	2.597
216	d_4	0.510	0.601	0.091	17.797
216	d_5	0.503	0.592	0.089	17.670
216	d_6	0.534	0.582	0.048	8.916
216	d_7	0.537	0.599	0.063	11.687
216	d_8	0.598	0.715	0.117	19.612
216	d_9	0.641	0.739	0.098	15.298
216	d_{10}	0.602	0.895	0.294	48.828
	Class Size 1	0.022	0.018	–0.004	18.919
	Class Size 2	0.024	0.021	–0.003	11.017
	Class Size 3	0.025	0.023	–0.002	6.504
	Class Size 4	0.025	0.023	–0.002	8.730
	Class Size 5	0.019	0.020	0.001	6.383
	Class Size 6	0.012	0.014	0.002	16.667

Table C3. Criteria Shifted Down (Lenient Raters), *d* = Normal 4, BIB, 6 Categories with Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
216	d_1	0.204	0.217	0.013	6.399
216	d_2	0.359	0.381	0.022	6.113
216	d_3	0.351	0.388	0.037	10.491
216	d_4	0.533	0.591	0.058	10.967
216	d_5	0.466	0.561	0.095	20.335
216	d_6	0.471	0.588	0.116	24.661
216	d_7	0.501	0.586	0.085	16.997
216	d_8	0.599	0.728	0.129	21.543
216	d_9	0.580	0.740	0.160	27.694
216	d_{10}	0.596	0.928	0.332	55.816
	Class Size 1	0.013	0.014	0.001	6.383
	Class Size 2	0.018	0.020	0.002	9.170
	Class Size 3	0.023	0.023	0.000	0.966
	Class Size 4	0.023	0.023	0.000	1.288
	Class Size 5	0.020	0.022	0.002	7.949
	Class Size 6	0.020	0.019	–0.001	3.700

Table C4. Criteria Shifted Up and Down (Strict and Lenient Raters), *d* =Normal 4, BIB, 6 Categories with Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
216	d_1	0.232	0.232	0.000	0.129
216	d_2	0.444	0.406	–0.038	8.476
216	d_3	0.364	0.424	0.060	16.452
216	d_4	0.529	0.626	0.097	18.426
216	d_5	0.517	0.620	0.103	19.907
216	d_6	0.529	0.625	0.095	17.986
216	d_7	0.555	0.607	0.052	9.450
216	d_8	0.534	0.796	0.263	49.176
216	d_9	0.618	0.727	0.110	17.798
216	d_{10}	0.587	0.858	0.271	46.175
	Class Size 1	0.012	0.015	0.003	21.951
	Class Size 2	0.018	0.019	0.001	5.556
	Class Size 3	0.021	0.022	0.002	7.317
	Class Size 4	0.022	0.022	0.000	1.852
	Class Size 5	0.021	0.019	–0.002	8.654
	Class Size 6	0.015	0.015	0.000	1.316

Table C5. Intersection Point Criteria, *d* = Normal 4, Unbalanced, 6 Categories with Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
370	d_1	0.191	0.189	–0.002	1.151
50	d_2	0.507	0.645	0.138	27.269
200	d_3	0.378	0.421	0.043	11.435
140	d_4	0.504	0.715	0.211	41.809
60	d_5	0.484	0.824	0.340	70.219
120	d_6	0.541	0.854	0.313	57.973
280	d_7	0.479	0.805	0.326	67.974
400	d_8	0.495	0.759	0.264	53.354
230	d_9	0.515	0.916	0.401	77.963
310	d_{10}	0.577	0.887	0.309	53.560
	Class Size 1	0.014	0.014	0.000	0.000
	Class Size 2	0.020	0.020	0.000	0.503
	Class Size 3	0.026	0.023	–0.003	10.506
	Class Size 4	0.021	0.023	0.002	9.005
	Class Size 5	0.021	0.020	–0.001	2.913
	Class Size 6	0.012	0.014	0.002	17.647

Table C6. Criteria Shifted Up (Strict Raters), *d* = Normal 4, Unbalanced, 6 Categories with Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
370	d_1	0.166	0.196	0.030	17.930
50	d_2	0.475	0.629	0.154	32.421
200	d_3	0.407	0.413	0.006	1.499
140	d_4	0.484	0.716	0.232	47.812
60	d_5	0.460	0.753	0.294	63.896
120	d_6	0.421	0.791	0.370	87.773
280	d_7	0.431	0.777	0.347	80.516
400	d_8	0.500	0.737	0.238	47.519
230	d_9	0.467	0.895	0.428	91.524
310	d_{10}	0.509	0.900	0.391	76.817
	Class Size 1	0.021	0.017	–0.004	20.188
	Class Size 2	0.030	0.023	–0.007	23.333
	Class Size 3	0.028	0.025	–0.003	11.661
	Class Size 4	0.025	0.024	–0.001	4.762
	Class Size 5	0.022	0.020	–0.002	9.502
	Class Size 6	0.013	0.015	0.002	16.279

Table C7. Criteria Shifted Down (Lenient Raters), *d* =Normal 4, Unbalanced, 6 Categories with Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
370	d_1	0.181	0.197	0.016	8.900
50	d_2	0.562	0.681	0.119	21.218
200	d_3	0.353	0.415	0.062	17.597
140	d_4	0.566	0.704	0.138	24.426
60	d_5	0.516	0.781	0.265	51.484
120	d_6	0.477	0.759	0.282	59.057
280	d_7	0.449	0.740	0.291	64.943
400	d_8	0.511	0.739	0.228	44.494
230	d_9	0.539	0.917	0.379	70.236
310	d_{10}	0.543	0.916	0.373	68.680
	Class Size 1	0.011	0.014	0.004	33.333
	Class Size 2	0.017	0.020	0.004	21.212
	Class Size 3	0.028	0.024	–0.004	14.591
	Class Size 4	0.026	0.025	–0.001	2.724
	Class Size 5	0.028	0.023	–0.005	16.968
	Class Size 6	0.022	0.018	–0.004	18.552

Table C8. Criteria Shifted Up and Down (Strict and Lenient Raters), *d* =Normal 4, Unbalanced, 6 Categories with Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
370	d_1	0.201	0.206	0.005	2.335
50	d_2	0.528	0.632	0.104	19.606
200	d_3	0.401	0.447	0.046	11.527
140	d_4	0.546	0.725	0.179	32.687
60	d_5	0.504	0.792	0.288	57.112
120	d_6	0.529	0.840	0.310	58.655
280	d_7	0.428	0.762	0.335	78.172
400	d_8	0.407	0.749	0.341	83.796
230	d_9	0.423	0.896	0.473	111.797
310	d_{10}	0.472	0.870	0.398	84.496
	Class Size 1	0.014	0.015	0.001	10.294
	Class Size 2	0.019	0.020	0.001	3.627
	Class Size 3	0.023	0.023	0.000	1.709
	Class Size 4	0.019	0.023	0.004	19.171
	Class Size 5	0.024	0.021	–0.003	12.863
	Class Size 6	0.019	0.016	–0.003	15.344

Table C9. Intersection Point Criteria, *d* = Normal 2, BIB, 6 Categories with Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
216	d_1	0.098	0.116	0.018	18.367
216	d_2	0.148	0.161	0.013	9.005
216	d_3	0.148	0.159	0.011	7.215
216	d_4	0.388	0.355	–0.033	8.411
216	d_5	0.322	0.336	0.014	4.344
216	d_6	0.331	0.332	0.001	0.320
216	d_7	0.391	0.353	–0.038	9.716
216	d_8	0.524	0.586	0.062	11.816
216	d_9	0.529	0.613	0.084	15.822
216	d_{10}	0.597	0.777	0.180	30.100
	Class Size 1	0.031	0.036	0.005	15.756
	Class Size 2	0.046	0.045	–0.001	2.808
	Class Size 3	0.052	0.051	–0.001	1.163
	Class Size 4	0.056	0.051	–0.005	8.602
	Class Size 5	0.048	0.044	–0.004	7.757
	Class Size 6	0.029	0.035	0.006	21.107

Table C10. Intersection Point Criteria, *d* = Normal 2, Unbalanced, 6 Categories with Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
370	d_1	0.093	0.085	–0.008	8.700
50	d_2	0.343	0.337	–0.006	1.663
200	d_3	0.142	0.153	0.011	7.444
140	d_4	0.305	0.455	0.150	49.376
60	d_5	0.379	0.516	0.136	35.878
120	d_6	0.329	0.617	0.288	87.591
280	d_7	0.316	0.604	0.288	91.347
400	d_8	0.335	0.597	0.262	78.174
230	d_9	0.456	0.771	0.315	68.987
310	d_{10}	0.478	0.784	0.306	64.066
	Class Size 1	0.030	0.036	0.006	20.000
	Class Size 2	0.044	0.046	0.002	4.072
	Class Size 3	0.045	0.054	0.009	20.805
	Class Size 4	0.048	0.054	0.006	12.735
	Class Size 5	0.044	0.046	0.002	4.545
	Class Size 6	0.033	0.036	0.003	8.434

Table C11. Intersection Point Criteria, *d* = Normal 4, BIB, 6 Categories with Non-normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
216	d_1	0.252	0.255	0.003	1.351
216	d_2	0.416	0.410	–0.006	1.442
216	d_3	0.372	0.415	0.043	11.649
216	d_4	0.593	0.600	0.007	1.232
216	d_5	0.600	0.614	0.013	2.199
216	d_6	0.574	0.618	0.044	7.671
216	d_7	0.542	0.596	0.054	9.869
216	d_8	0.683	0.770	0.086	12.648
216	d_9	0.696	0.785	0.089	12.753
216	d_{10}	0.621	0.904	0.284	45.714
	Class Size 1	0.007	0.007	0.000	1.408
	Class Size 2	0.061	0.012	–0.049	80.263
	Class Size 3	0.029	0.027	–0.002	7.216
	Class Size 4	0.047	0.028	–0.019	40.803
	Class Size 5	0.028	0.019	–0.009	33.099
	Class Size 6	0.009	0.010	0.001	9.890

Table C12. Intersection Point Criteria, *d* = Normal 4, BIB, 4 Categories with Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
216	d_1	0.290	0.272	–0.018	6.132
216	d_2	0.417	0.420	0.003	0.637
216	d_3	0.456	0.420	–0.036	7.885
216	d_4	0.560	0.624	0.064	11.389
216	d_5	0.608	0.643	0.035	5.758
216	d_6	0.624	0.624	0.000	0.059
216	d_7	0.615	0.651	0.036	5.834
216	d_8	0.769	0.900	0.131	17.016
216	d_9	0.792	0.946	0.154	19.438
216	d_{10}	0.781	1.044	0.263	33.631
	Class Size 1	0.018	0.021	0.003	19.048
	Class Size 2	0.025	0.026	0.001	3.462
	Class Size 3	0.026	0.027	0.001	2.779
	Class Size 4	0.020	0.022	0.002	10.943

Table C13. Intersection Point Criteria, *d* = Normal 4, BIB, 4 Categories with Non-Normal Class Sizes

Size	Parameter	SD	Mean SE	Bias	% Bias
216	d_1	0.338	0.310	–0.028	8.403
216	d_2	0.416	0.446	0.030	7.312
216	d_3	0.504	0.462	–0.042	8.361
216	d_4	0.708	0.671	–0.037	5.193
216	d_5	0.574	0.620	0.047	8.120
216	d_6	0.702	0.643	–0.059	8.385
216	d_7	0.644	0.641	–0.003	0.520
216	d_8	0.874	0.918	0.043	4.973
216	d_9	0.754	0.925	0.171	22.660
216	d_{10}	0.901	1.114	0.212	23.545
	Class Size 1	0.031	0.022	–0.009	29.283
	Class Size 2	0.033	0.031	–0.002	5.257
	Class Size 3	0.036	0.030	–0.006	16.037
	Class Size 4	0.031	0.021	–0.010	33.078

Table C14. Intersection Point Criteria, *d* = Normal 4, BIB, 4 Categories with Non-Normal Class Sizes (Shift in Density, First Scoring Occasion)

Size	Parameter	SD	Mean SE	Bias	% Bias
216	d_1	0.420	0.321	–0.099	23.484
216	d_2	0.521	0.453	–0.068	13.079
216	d_3	0.598	0.458	–0.140	23.465
216	d_4	0.805	0.672	–0.133	16.539
216	d_5	0.732	0.681	–0.051	7.009
216	d_6	0.800	0.709	–0.090	11.292
216	d_7	0.768	0.664	–0.104	13.499
216	d_8	0.902	0.924	0.021	2.361
216	d_9	0.925	0.891	–0.034	3.690
216	d_{10}	0.876	1.130	0.253	28.924
	Class Size 1	0.033	0.024	–0.009	28.337
	Class Size 2	0.042	0.033	–0.009	21.485
	Class Size 3	0.082	0.038	–0.044	53.771
	Class Size 4	0.094	0.029	–0.065	69.007

Table C15. Intersection Point Criteria, *d* = Normal 4, BIB, 4 Categories with Non-Normal Class Sizes (Shift in Density, Second Scoring Occasion)

Size	Parameter	SD	Mean SE	Bias	% Bias
216	d_1	0.446	0.325	–0.121	27.050
216	d_2	0.589	0.458	–0.131	22.289
216	d_3	0.548	0.453	–0.095	17.312
216	d_4	0.777	0.709	–0.068	8.769
216	d_5	0.800	0.650	–0.150	18.784
216	d_6	0.673	0.675	0.002	0.279
216	d_7	0.770	0.682	–0.088	11.408
216	d_8	0.991	0.912	–0.079	7.979
216	d_9	0.730	0.883	0.153	20.948
216	d_{10}	1.114	1.165	0.051	4.556
	Class Size 1	0.077	0.027	–0.050	65.152
	Class Size 2	0.073	0.036	–0.037	50.488
	Class Size 3	0.040	0.033	–0.007	18.012
	Class Size 4	0.033	0.025	–0.008	24.562