COMPARING THE EFFECTIVENESS OF SELF-PACED AND COLLABORATIVE FRAME-OF-REFERENCE TRAINING ON RATER ACCURACY IN A LARGE-SCALE WRITING ASSESSMENT

by

KEVIN ROBERT RACZYNSKI

(Under the Direction of Allan S. Cohen)

ABSTRACT

There exists a large body of research on the effectiveness of rater training methods in the Industrial and Organizational Psychology literature. Far less research has been done on the effectiveness of rater training methods in large-scale writing assessments. The purpose of this dissertation is to compare the effectiveness of two widely-used rater training methods—self-paced and collaborative frame-of-reference training—in the context of a large-scale, statewide writing assessment. Sixty-six raters were randomly assigned to the training methods. After training, all raters scored a common set of fifty representative essays. To determine raters' accuracy on these essays, raters' scores were compared to resolved expert scores and coded accurate (1) when the scores matched and inaccurate (0) otherwise. This approach was taken because over ninety-nine percent of these comparisons aligned either exactly or within one point. A series of logistic mixed models were then fitted to these binary data. Results suggested that the self-paced method was equivalent in effectiveness to the more time-intensive and costly

collaborative method. Implications for large-scale writing assessments and suggestions for further research are discussed.

INDEX WORDS: Rater training, frame-of-reference training, self-paced, collaborative, accuracy, experts, writing assessment, logistic mixed models, validity

COMPARING THE EFFECTIVENESS OF SELF-PACED AND COLLABORATIVE FRAME-OF-REFERENCE TRAINING ON RATER ACCURACY IN A LARGE-SCALE WRITING ASSESSMENT

by

KEVIN R. RACZYNSKI

Bachelor of Arts, The University of Wisconsin-Madison, 2001

Master of Arts, The University of Georgia, 2004

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

COMPARING THE EFFECTIVENESS OF SELF-PACED AND COLLABORATIVE FRAME-OF-REFERENCE TRAINING ON RATER ACCURACY IN A LARGE-SCALE WRITING ASSESSMENT

by

KEVIN ROBERT RACZYNSKI

Major Professor:	Allan Cohen
Committee:	George Engelhard Shawn Glynn Seock-Ho Kim Laura Lu

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
August 2014

ACKNOWLEDGEMENTS

I wish to acknowledge the guidance provided by Allan Cohen, my advisor and the Director of the Georgia Center for Assessment. His support of my academic and professional development has afforded me opportunities I could not have imagined when I began working at the Georgia Center for Assessment. I am similarly indebted to my entire advisory committee, noted here alphabetically: George Engelhard, Shawn Glynn, Seock-Ho Kim, and Laura Lu. It has been an honor learning from and working with such esteemed scholars. George Engelhard's thinking on writing assessment continues to shape much of my own, as evidenced in this dissertation. Shawn Glynn introduced me to the fundamentals of cognition, and scholarship in this discipline has affected the way I think about rater cognition, with specific implications for rater training. Seock-Ho Kim has provided excellent instruction, in the Analysis of Variance and Item Response Theory; he has also given me sound advice throughout my graduate studies, for which I will remain grateful. Laura Lu's instruction on hierarchical models can be found throughout this dissertation. I am indebted to her for her superb guidance. Further, I would like to thank current and retired colleagues at the Georgia Center for Assessment, especially Belita Gordon, Steve Cramer, Jeremy Granade, Candace Langford, and Rick Stoner for their unflagging commitment to best practices in writing assessment. Each of these individuals has influenced my thinking about writing assessment in profound ways. Further, I would like to thank the raters who participated in this research. Finally, I am grateful for Katherine Raczynski, partner in love, life, and the pursuit of careful thinking. It is impossible to imagine reaching this point without you.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

COMPARING THE EFFECTIVENESS OF SELF-PACED AND COLLABORATIVE FRAME-OF-REFERENCE TRAINING ON RATER ACCURACY IN A LARGE-SCALE WRITING ASSESSMENT

by

KEVIN ROBERT RACZYNSKI

CHAPTER 1

INTRODUCTION

Direct writing assessments are used commonly in educational testing, including large-scale state writing assessments, Advanced Placement exams, the SAT, and the Graduate Record Exam. Moreover, direct writing assessments will comprise a sizable portion of the large-scale assessments being developed to align with the Common Core State Standards, a national curriculum in both literacy and mathematics. As several researchers have noted, crafting plausible validity arguments for direct writing assessments and other performance assessments is especially challenging, for at least two reasons: a) most direct writing assessments include a small number of items, and b) raters can be idiosyncratic, introducing unwanted effects (Engelhard, 2002; Kane, Crooks, & Cohen, 1999; Lane & Stone, 2006; Messick, 1995). These issues potentially undermine both structural and scoring reliability—necessary but not sufficient prerequisites for assessment validity. This dissertation focuses on improving the scoring aspect of direct writing assessments.

Most of the research and theoretical work on the scoring aspect of direct writing assessments has focused on identifying, modeling, and adjusting for rater idiosyncrasies. Engelhard (2013) places rater idiosyncrasies into two categories: a) rater errors and systematic biases (hereafter referred to as "rater errors") and b) rater inaccuracy. This distinction is substantive, not merely semantic. Rater errors include severity and failure to use the entire rating scale (Murphy & Cleveland, 1995; Saal, Downey, & Lahey, 1980). In addition to these, there

are halo errors, central tendency errors, and first/last impression errors (Bernardin, 1978; Engelhard, 2002; Latham, Wexley, & Pursell, 1975). *Halo errors* denote a raters' tendency to assign the same score in all domains of the writing sample when the sample is to be rated in multiple domains. That is, the rater fails to differentiate appropriately between, say, how well the essay is developed, how well the essay is organized, and how well the examinee controls grammar. Instead, the rater assigns the same score in all of these domains, typically because the response is especially strong or weak in one of the domains, thereby overwhelming the rater's judgment with respect to the other domains. Raters who demonstrate *central tendency errors* tend to avoid using *both* the lowest and highest categories on the rating scale. As a result, less proficient examinees' scores are inflated and more proficient examinees' scores are artificially low. *First and last impression errors* occur when an examinee's writing sample begins or ends effectively but the remainder is weaker. Raters focus on these strengths and ignore the weaknesses. That is, the rater is not evaluating the writing sample as a whole. All of these errors can bias scores, and, therefore, pose threats to the validity of the scores and interpretations based on them (Lane & Stone 2006; Messick, 1995). Rater inaccuracy, on the other hand, denotes a lack of correspondence between a rater's scores and an expert's (Cronbach, 1955; Engelhard, 1996; Sulsky & Balzer, 1988). Sulsky and Balzer (1988) note some challenges in defining expertise, but assuming that expert raters have a sound understanding of the construct, the writing assessment, the rating scale, and how the rating scale is to be applied properly, expert ratings are likely to bear the closest proximity to true or correct ratings (Engelhard, 1996; Sulsky & Balzer, 1988). Therefore, indices of rater accuracy provide direct evidence of rating quality (Engelhard, 2013; Murphy & Cleveland, 1995). While meaningful, indices of rater errors

provide indirect evidence, mainly because they are established by comparing raters with each other, without an expert, external frame of reference.

Several statistical approaches have been developed to model rater errors and rater inaccuracy. Linacre (1989) provides an overview of mathematical models that have been proposed over the last century; he then describes his many-faceted model, an item response theory (IRT) model that expands the Rasch model to account for a rater's influence on judged performances. Several researchers have used it to model rater errors in direct writing assessments (Engelhard, 1992, 1994; Knoch, Read, & von Randow, 2007; Weigle, 1998; Wolfe & McVay, 2010). In addition to IRT models, researchers have used the analysis of variance (Gyagenda & Engelhard, 2009), hierarchical linear modeling (Leckie & Baird, 2011), and signal detection theory (DeCarlo, 2005) to model rater errors and rater behavior. Comparatively little work has been done on modeling rater inaccuracy (Engelhard, 1996; Leckie & Baird, 2011; Raczynski, Cohen, & Lu, 2013). With few exceptions, this research demonstrates that rater inaccuracy and rater errors are another source of score variation in performance assessments. As a consequence, an examinee's score might depend on the rater assigned to score the performance. Messick (1995) describes such unwanted score variation as construct-irrelevant score variance.

A variety of methods have been proposed to adjust for the undue influence that raters introduce into scoring. Eckes (2009) provides a detailed account of adjusting examinees' scores based on the rater parameter estimates derived from Linacre's many-faceted model. For instance, if an examinee's response was scored by a rater (or raters) shown to be lenient or severe, Eckes (2009) describes methods for adjusting the score to correct for this rater effect. There are difficulties, however, with *post hoc* score adjustments, even if there is acceptable

model-data fit. Although a rater may be systematically lenient or severe, the rater could score any particular essay accurately. Performing a *post-hoc* score adjustment on such an essay, based exclusively on the rater's estimated severity, could, in fact, make the reported score less accurate.

A more conservative approach is to use indices of rater errors to reevaluate examinee responses. For example, say a subset of raters appears to have truncated the rating scale by not using the highest category on it. The responses these raters have scored that are close to a decision boundary can be reviewed by an expert validity committee. This committee can then directly investigate the degree to which any of these raters is truncating the rating scale. Scores can be adjusted as necessary, and these raters can receive feedback.

Of course, it would be ideal to prevent rater errors and rater inaccuracy from manifesting in the first place. Such prevention is one of the principal aims of rater training. Most research into rater errors and rater inaccuracy has included a training component but not a study of training effects (Eckes, 2009; Engelhard, 1992, 1994; Gyagenda & Engelhard, 2009; Leckie & Barid, 2011). Although it was not the focus of these studies to investigate the effectiveness of training, it seems clear that whatever training the raters received could have been better. In fact, little is known about the effectiveness of training methods in the context of writing assessments. For example, it is not clear whether collaborative or self-paced frame-of-reference methods—two of the most common types of rater training methods—are more effective at making raters accurate. The purpose of this dissertation is to address this gap in the literature. Specifically, this dissertation compares raters randomly assigned to a collaborative or self-paced frame-of-reference training method (described in the Method section). After training, these raters scored a common set of 50 essays that were selected and prescored by an expert validity committee. Finally, a series of logistic mixed models were fitted to the rating data associated with these

essays to model the effects of training on rater accuracy. The central research question of this dissertation is, *does either collaborative or self-paced frame-of-reference training have a more pronounced effect on rater accuracy*? It is hypothesized that collaborative training, sometimes called "classroom training," will make raters more accurate because this method includes dialogue germane to the writing assessment between a group of raters and an expert trainer (Johnson, Penny, & Gordon, 2009). By comparison, self-paced training involves raters working through training at their own pace and conferencing one-on-one with a trainer; other raters are not involved (Johnson, Penny, & Gordon, 2009). Therefore, there are no formal opportunities for raters to learn from one another, as there are in the collaborative training method. Collaborative, discourse-based learning environments have been shown to improve understanding in a variety of subject areas and contexts. Bruning, Shraw, and Norby (2011) review several studies that have shown this to be the case. Bruning, Shraw, and Norby (2011) describe the manner in which expert-facilitated group discussions can help learners enhance and organize their understanding on a particular subject, particularly when the expert serves mostly as a facilitator. That is, the learners engage the content and one another to a large extent. The expert mainly keeps the discussion focused and provides appropriate feedback when misunderstandings arise. For instance, one learner might be able to explain a concept related to the writing assessment with particular clarity, such that it promotes understanding of the concept among the other raters in the group. Conversely, if this explanation strikes the expert as misleading, the expert can offer revisions or ask other raters in the group to do so. In short, the collaborative method offers more opportunities for raters to dialogue with multiple people about their understanding of the writing assessment.

A second research question of this dissertation is, *does either training method have a more pronounced effect on rater accuracy for different essay types*? Research has shown that some essays are more difficult to score accurately than others (Engelhard, 1996). To explore this matter further, the essays used in this dissertation to train raters, and the essays used to determine rater accuracy after training, were independently scored by members of an expert validity committee. The experts' scores on these essays were ultimately resolved and used as the key scores, but the experts' independent scores were retained to investigate the degree to which the experts themselves agreed (Sulsky & Balzer, 1988). It is assumed that as experts disagree to an increasing degree, the essay will be more difficult for raters to score accurately. Accordingly, one of the variables in this dissertation is *essay type*, which takes a numeric value, and is defined as the absolute maximum difference between the scores that the individual members of the expert validity committee assigned to a particular essay. For example, when essay type takes a value of 0, each member of the expert validity committee assigned the same score. It is hypothesized that when essay type takes on higher values, raters' accuracy will decrease. Further, it is hypothesized that this decrease in accuracy will be smaller for raters who received collaborative training.

In summary, this dissertation addresses three research questions:

1) Is there inter-rater variation in accuracy on a common set of 50 essays?
2) How does a rater who received self-paced training compare in accuracy to a rater who received collaborative training?
3) What is the strength of association between accuracy and essay type, and is this slope the same for raters who received self-paced training and raters who received collaborative training?

Results of this dissertation could have practical implications for how writing assessment training is designed and delivered, with the ultimate goal of mitigating rater inaccuracy to the greatest extent possible. The remainder of this dissertation is organized in the following way. The second chapter provides a review of the extant literature on rater training, much of which comes from industrial and organizational psychology research. There are, however, some studies that have been done on the effectiveness of rater training in writing assessments. The third chapter describes in detail the methods of the dissertation. Chapter Four includes the results. Finally, Chapter Five provides a discussion of the results, the limitations of the project, and future directions for research on rater training.

CHAPTER 2

REVIEW OF THE LITERATURE ON RATER TRAINING

Researchers have been exploring the effectiveness of rater training programs for at least four decades. However, much of the research on rater training has not been done in the field of writing assessment but instead in industrial and organizational psychology, with a focus on managers providing ratings of employee performance and students providing ratings of college instructors. This body of research started with a focus on *rater error training*, which stresses particular rating patterns to avoid, such as halo patterns and assigning too many ratings at the lower and upper ends of the rating scale. During the late 1970s and beyond, researchers started realizing that rater error training did not necessarily make raters more accurate. Therefore, researchers shifted their attention to the effectiveness of *frame-of-reference training*, which is a more applied method of rater training. More specifically, it exposes raters to exemplar performances and how a rater might commit an error when scoring them, such as not differentiating appropriately between different aspects of the performance (i.e., halo effects).

This literature review begins with two sections that summarize some of the key research contributions that both reflected and influenced the shift from rater error training to frame-of-reference training. The third section describes the few studies that have been done on the effectiveness of rater training programs in large-scale writing assessments.

Summary of the Research on Rater Error Training

Some of the earliest research on rater training focused little on whether training helped raters score accurately; instead, the focus was on influencing raters' scoring patterns so that the rating data had specific psychometric properties. Bernardin (1978), for instance, explored whether rater error training would help raters avoid overusing certain rating patterns, such as overly lenient, halo, and central tendency patterns. In Bernardin (1978), 80 college students served as raters of college instructors. Students were randomly assigned to one of four groups. One group received one hour of rater error training, a second 5 minutes of rater error training, a third no training, and a fourth either 60 minutes or 5 minutes of rater error training. The training included definitions of scoring patterns that reflected leniency, halo, or central tendency rating patterns, graphic illustrations of these rating patterns, and practice evaluating such rating patterns. After training, raters rated their non-laboratory instructors on three occasions. Results suggested that the more training the raters received, the less likely they were to overuse rating patterns that reflected halo, leniency, or central tendency. Subsequent research, including research conducted by this author, criticized the utility of this approach, particularly the assumption that rating data should look a certain way, such as the absence of too many high scores, too many low scores, and too many homogenous rating patterns across domains of the performance. In fact, subsequent research—some of which is summarized in this literature review—would show that rater error training that stresses particular rating patterns has a detrimental effect on accuracy.

Taking a different approach in exploring the effectiveness of rater error training, Bernardin and Pence (1980) investigated whether rating data with specific psychometric properties accurately reflected the distribution of examinee proficiency. As made clear in

Bernardin (1978), one view of rating data was that it should look a certain way (e.g., distributed roughly normal). Skewed distributions and high correlations of scores across multiple domains of a performance were taken to imply inaccurate ratings. The authors, however, questioned whether such distributional assumptions and implications for accuracy were reasonable. To investigate, the authors randomly assigned 72 undergraduate students to one of three groups. The first group received training on rating patterns exhibiting leniency and halo errors, similar to the training described by Bernardin (1978). The second group received one of the earliest examples of frame-of-reference training, which familiarized the raters with the multiple domains of the performances they would be rating, as well as examples of low, mid, and high proficiency in each of the domains. The third group served as a control and received no training. Subsequently, all raters evaluated the effectiveness of two instructors on 13 domains using a 7-point rating scale. These ratings were compared to the ratings assigned by an independent group of raters, whose scores were assumed to be accurate. (The authors do not elaborate on the credentials of this independent group of raters). Results indicated that group 1 exhibited less leniency and evidence of halo rating patterns, but their ratings were significantly less accurate than the ratings from groups 2 and 3. That is, the raters from group 1 gave ratings that corresponded least well with the ratings assigned by the independent group of raters. These results suggest that rater error training as described in Bernardin (1978) leads to specific rating patterns but does not make raters more accurate. Moreover, the study provides some of the earliest empirical evidence that frame-of-reference training helps raters score more accurately than rater error training.

By the time Borman (1979) conducted one of the earliest studies of training effects on rating accuracy, views on rater training had begun to evolve. Researchers became progressively

less interested in whether training led to rating data that looked a particular way and more interested in whether the training helped raters evaluate performances accurately. In Borman's (1979) study, a sample of 123 raters were randomly assigned to one of ten groups, and each group was assigned one of five rating scales and one of two training treatments: training or no training. The five groups that underwent training received a standardized regimen of rater error training that differed markedly from the type Bernardin (1978) described. It featured a presentation on three rater errors, a viewing of videotaped exemplar interviews between managers and prospective employees, and rating both the manger's and the prospective employee's performances using the assigned rating scale. The participants then discussed their ratings and the degree to which each exemplar might lead a rater to commit any of the three errors. After training, all raters completed two rating sessions. The first occurred one week after training and the second two weeks after the first rating session. Raters scored eight examinees' job performances per rating session, using the rating format to which they were assigned. Accuracy was defined as the correlation between the raters' scores and scores provided by an expert validity committee. It should be noted that the experts assigned their scores using a sixth rating scale distinct from any of the five rating scales that raters used. Results showed that training did not significantly improve raters' accuracy. However, the manner in which the author defined the accuracy measures seems to require further justification. Raters used one of five rating scales, and the experts used an entirely different scale. Ostensibly, it was reasonable to correlate ratings from these different scales, but the author provides little elaboration on whether it was theoretically appropriate to have done so. Therefore, while this study was one of the first to explore the effects of training on accuracy, the methods and measures arguably made the results questionable.

The training protocol in Borman (1979) closely resembled the one described in Latham, Wexley, and Pursell (1975). In fact, Latham, Wexley, and Pursell (1975) developed one of the first frame-of-reference training protocols of record. Instead of discouraging specific rating patterns, this training protocol showed rater trainees examples of performances where specific rating patterns (e.g., a halo pattern) would be *inaccurate*. The authors' training protocol exposed trainees to four rater errors: similarity, halo, contrast, and first impression errors. *Similarity* error occurs when the features of the examinee's performance resonate with the rater's personal biases, often resulting in an unduly high rating. *Halo* error occurs when one aspect of an examinee's performance is so effective (or weak) that it dominates the rater's overall impression. *Contrast* error results when the rater scores one response relative to the quality of other responses instead of rating each performance as an independent event. *First impression* (or last impression) errors occur when a particularly weak or strong beginning (or ending) to a performance dominates the rater's overall impression. Latham, Wexley, and Pursell (1975) were interested in whether training raters to avoid these errors would help them avoid making them when scoring.

Participants in the study were randomly assigned to either frame-of-reference training or no training. Frame-of-reference training involved the following: a presentation of the four rater errors, watching videotaped exemplar interviews between managers and prospective employees, and rating both the manger's and the prospective employee's performances. The participants then discussed their ratings and the degree to which each exemplar might lead a rater to commit any of the four rater errors. Six months after the training, raters were given a test where they rated a candidate's aptitude for a particular job, given specific information about the candidate and the candidate's videotaped interview performance. The test included four candidates, and

each candidate had an essential characteristic that could have resulted in the raters committing one of the four rater errors. Results indicated that the raters who received frame-of-reference training committed significantly fewer errors than the control group.

The manner in which the results were obtained seems questionable, however. The post-training test actually had two forms, and half the raters from both groups took form one while the other half took form two. Both forms involved rating the same four candidates, but some of the characteristics about the candidates changed according to the test form. For instance, in form 1, one of the candidates began the interview strongly and then flagged. In form two, the same candidate began less strongly but improved as the interview progressed. The authors argued that in either case raters should avoid a first (or last) impression error. That is, regardless of the test form, raters should have provided roughly the same score. Mean scores for each of the four candidates were compared for raters who took forms 1 and 2. Mean scores from the raters assigned to frame-of-reference training did not differ significantly, for any of the four candidates. Mean ratings assigned by the control group differed on three of the four. However, similar mean ratings do not necessarily imply accurate ratings. An alternative approach would have been to compare raters' scores to an expert validity committee's scores. Nevertheless, this study was influential, given that researchers in the 1980s began focusing more on the effectiveness of frame-of-reference training rather than rater error training.

A Shift to Research on Frame-of-Reference Training

Though not a formal study, Bernardin and Buckley (1981) make a strong case to abandon rater error training in lieu of expanded frame-of-reference rater training methods. This article is noteworthy because the first author had heretofore been a proponent of rater error training

(Bernardin, 1978). Studies showing how avoidance of particular rating patterns did not necessarily lead to accurate ratings prompted this evolution in thought (Bernardin & Pence, 1980). After establishing this point, the authors offer suggestions for expanding frame-of-reference methods. For instance, the authors recommend making raters familiar with an appropriate rating scale (i.e., rubric), showing raters exemplars, and giving raters the opportunity to practice applying this rubric to additional exemplars. In fact, these activities are staples of most contemporary frame-of-reference training protocols, including those used in direct writing assessments.

Following Bernardin and Buckley (1981), Pulakos (1984) conducted one of the first studies on the effects of expanded frame-of-reference training on rater accuracy. The frame-of-reference training employed in this study progressed as follows: raters became familiar with a rubric, learned how to use the rubric to guide their observations of examinee performances, and then discussed how exemplar performances reflected specific aspects of the rubric.

Pulakos (1984) used an experimental design, randomly assigning 108 undergraduate students to one of four groups: frame-of-reference training, rater error training similar to the model Bernardin (1978) described, a combination of both, or no training. Each training session lasted 75 minutes. After training, raters evaluated five video recordings of managers interacting with employees. The participants rated the managers' performance in five domains on a scale of 1-7. To establish a measure of their accuracy, the participants' ratings were correlated with the ratings of an expert validity committee. Results showed that raters who received only rater error training gave ratings that were most variable across domains but least accurate among any of the other three groups. Raters who received only frame-of-reference training, by contrast, scored most accurately. Given that this study was one of the first to investigate the effect of expanded

frame-of- reference training on rater accuracy, the contribution to the literature is especially meaningful. It provided some of the first empirical evidence that frame-of-reference training could help raters score more like an expert validity committee.

Woehr and Huffcutt's (1994) comprehensive overview of research on rater training makes a clear case that any rater training model should have a frame-of-reference component. The authors conducted a meta-analysis of 29 studies that explored the effects of rater training on indices of rater proficiency, such as halo, leniency, and accuracy. It should be noted that the majority of these studies came from industrial and organizational psychology research where raters evaluated job performances. Results from the meta-analysis showed that frame-of-reference training, in particular, had a large positive effect on accuracy. Suggestions for further research included a focus on the method of training delivery.

Rater Training in Direct Writing Assessments

Direct writing assessments became common in the late 1980s, and most descriptions of rater training in the writing assessment literature reflect frame-of- reference training. This is hardly surprising, given that by the early 1980s the research on rater training clearly suggested that frame-of-reference training was more effective than rater error training at helping raters score accurately.

Weigle (1998) conducted one of the first studies of rater training in the context of writing assessment and explored an important question: does rater training succeed in making raters more proficient? She used a pre/post design; the sixteen raters in the study scored a set of essays both prior to training and after training. In the interim, the raters in the study received frame-of-reference training that was collaborative in nature. That is, raters studied benchmark essays assembled and scored by experts and then engaged in group discussions about any benchmarks

that presented challenges. Rating data from both of the pre and post scoring sessions were fitted to Linacre's (1989) many-faceted Rasch model. Results from these analyses show that raters differed in severity both prior to training and after training. However, after training, raters became more internally consistent. Therefore, training had some effect on rater proficiency.

Additional research on rater training in direct writing assessments has focused on how frame-of-reference training is *delivered*, as suggested by Woehr and Huffcutt (1994), but the research also features two of the most common types of frame-of-reference training: self-paced and collaborative. Self-paced and collaborative methods include identical training materials, such as rubrics and exemplars (i.e., benchmarks), but there are key differences. In the self-paced approach, raters work through each segment of training independently but meet with trainers at specific points to discuss their progress. In a collaborative format, trainers and raters progress through training as a group, often referred to as a cohort.

Knoch, Read, and von Randow (2007) conducted one of the few studies comparing self-paced and collaborative frame-of-reference training methods associated with a direct writing assessment. It should be noted, however, that the authors focused less on the training methods themselves and more on how they were delivered: online or face-to-face. Raters assigned to the self-paced method completed training online, whereas raters assigned to the collaborative method completed training at a scoring center. The authors explored the effectiveness of these two training methods and delivery systems at improving internal consistency and reducing leniency, severity, central tendency, and halo effects. The study included 16 raters, 8 assigned to the self-paced group and 8 to the collaborative group. Random assignment was not utilized. Prior to training, all raters scored a common set of 70 essays to establish preliminary measures of rater proficiency. The self-paced training involved raters studying the rubrics, scoring

benchmark essays, writing a rationale for their ratings, and comparing their ratings and rationales to the benchmark ratings and rationales. Training for the collaborative group was identical in content, but raters and trainers discussed the benchmark essays, including scores and rationales, as a group. After training, all raters rescored the common set of 70 essays. The authors argued that because this set contained a large number of essays, raters were not likely to have remembered them. The rating data were analyzed using the many-faceted Rasch model, developed by Linacre (1989). Results were mixed. Both training methods made raters more internally consistent; self-paced training led to greater reductions in severity and leniency; collaborative training led to greater reductions in halo effects. Limitations to the study include lack of random assignment and lack of comparison to an expert validity committee. That is, the analysis centered on what Engelhard (2013) refers to as rater errors/systematic biases, which are indirect measures of rating accuracy. Comparing raters' scores to those assigned by an expert validity committee would have provided direct measures of raters' accuracy.

Wolfe and McVay (2010) also investigated whether raters assigned to three different training methods and delivery systems differed on multiple measures of proficiency. These three training methods and delivery systems were: a) collaborative frame-of-reference training delivered at a scoring center; b) self-paced frame-of-reference training raters completed online at a scoring center; and c) self-paced frame-of-reference training raters completed online at a remote location, such as a residence. The content of the training was identical for all three groups, and after training, all raters scored a common set of 400 responses prescored by a group of expert raters. The authors analyzed the rating data using Linacre's (1989) many-faceted model and explored the effects of training on raters' accuracy as well as rater errors/systematic biases, including centrality, severity, and halo scoring patterns. Results showed that raters

assigned to the online training methods were more accurate and showed less centrality than raters assigned to the collaborative method. It should be noted that raters were not randomly assigned to the three training methods, though the authors suggest that the three groups were roughly equivalent on a variety of demographic characteristics.

This dissertation extends this research on the effectiveness of self-paced and collaborative frame-of-reference training methods in direct writing assessments. Key differences between this dissertation and previous research include: a) random assignment of raters to either self-paced or collaborative training methods, b) having all raters complete training and scoring at a scoring facility rather than having some train/score online and others at a scoring facility, and c) establishing direct measures of rater accuracy by comparing raters' scores to those assigned by an expert validity committee.

CHAPTER 3

METHOD

Instrument: The Georgia Supplemental Writing Assessment

This dissertation presents a study comparing the effects of self-paced and collaborative frame-of-reference training on rater accuracy in the context of the Georgia Supplemental Writing Assessment (GSWA). The GSWA is a practice writing test designed to give diagnostic feedback on examinees' writing achievement. It is administered at three grade levels—4, 7, and 10. In this dissertation, the focus is on data from the Grade 10 administration from the autumn of 2013. Examinees were assigned a persuasive writing topic: should school systems offer online high school courses? Further, examinees were allotted 100 minutes to complete a multi-paragraph essay in response to the topic. Raters scored each essay on a scale of 1-5 in four domains of writing: ideas, organization, style, and conventions. Of particular interest were raters' scores on a set of 50 essays that all raters scored immediately after receiving training. These essays were selected and pre-scored by an expert validity committee.

Participants

Sixty-six raters out of a pool of approximately 90 raters were available for the 2013 GSWA project. Prior to participating in the GSWA, all raters had completed at least one high school scoring project at the Georgia Center for Assessment, a test development, scoring and reporting center at the University of Georgia. The 66 raters were randomly assigned to one of six cohorts

of raters. Each cohort started with eleven raters, led by a trainer. Cohorts 1-3 received collaborative training, and cohorts 4-6 received self-paced training. Raters were assigned to smaller cohorts for two reasons: a) to avoid the potential confounding of training method effects and trainer effects, and b) to observe standard protocol associated with the Georgia Supplemental Writing Assessment program, where raters work in small cohorts led by a trainer. Due to personal reasons, four raters dropped out of the project. Three of these raters were from the cohorts that received collaborative training, and one came from the cohorts that received self-paced training. Therefore, a total of 30 raters received collaborative training, and 32 raters received self-paced training. This resulted in a slight imbalance in the rating data described and analyzed in subsequent sections of this dissertation. However, these data were analyzed by fitting logistic mixed models, which are robust to imbalanced data (Hox, 2010). Finally, Table 1 reflects the mean overall accuracy values, for each of the six cohorts, from the 2013 Georgia High School Writing Test (GHSWT), the scoring project that preceded the 2013 Georgia Supplemental Writing Assessment. Overall accuracy values denote the percentage of time raters agreed exactly with expert scores on check essays used to monitor raters' accuracy during the 2013 GHSWT. Each rater scored a minimum of 480 check essays during the 2013 GHSWT.

Table 1

Mean Overall Accuracy Values from the 2013 GHSWT, by Cohort

Cohort	Training Method	Mean Overall Accuracy	Grand Mean (by training method)
1	Collaborative	64.9	
2	Collaborative	63.8	
3	Collaborative	65.3	64.7
4	Self-Paced	64.4	
5	Self-Paced	64.9	
6	Self-Paced	64.7	64.7

The information in Table 1 demonstrates that the random assignment resulted in essentially equivalent cohorts with respect to overall accuracy on a previous high school scoring project. In fact, the grand means for the previous scoring project were identical for both training conditions.

Selection of Training Materials and Common Sets

The training materials used in both training methods were identical. This section focuses on the process used to assemble not only the training materials but also the 50 common set essays that all raters scored after training to determine their scoring accuracy. This process will be described in some detail because some essays have been shown to be more difficult to score accurately than others (Engelhard, 1996; Raczynski, 2012). Therefore, to get a clearer understanding of the relative effects of both training conditions on rater accuracy, it was important to choose essays carefully. To this end, a group of four experts on the Georgia Supplemental Writing Assessment independently scored a set of approximately 175 essays available for use as training benchmarks and common set essays. These individuals are considered experts because they helped develop the Georgia Supplemental Writing Assessment and its rubrics. Moreover, they have considerable experience applying these rubrics to essays. An assumption was made that when either all four experts, or three of the four experts, independently assigned the same score to a given essay that this score was the most accurate score. If, on the other hand, the experts were evenly split in the scores they assigned to an essay, it was assumed that the essay exhibited characteristics of two adjacent score points on the rubric, meaning that it would be too difficult to determine one accurate score for the essay. Such essays were not used, either as training benchmarks or common set essays.

After all four experts scored the approximately 175 essays, the experts' scores were compiled, and selections were made for the training benchmarks and the common set essays, using the criteria noted above. Ultimately, four sets of training benchmarks were selected, one for each domain. Each set of training benchmarks included 18 essays, which reflected the entire range of the rubric, from score point one to score point five. For most of the training benchmarks, there was unanimous agreement on the scores assigned by each expert. There were, however, a few benchmarks in each domain where only three experts assigned the same score and one of the experts assigned the adjacent score. For instance, if three experts assigned a score of 2 and one expert a score of 3 in a particular domain, the essay was resolved as a 2+ in that domain. Such examples were included to help raters see the range of each score point on the rubric. Please see the tables in Appendix A for a full accounting of the set of training benchmarks compiled for each domain. These tables include both the resolved score for each training benchmark, in addition to the independent score each expert assigned.

In addition to the four sets of training benchmarks, an additional set of ten practice benchmarks was selected for practice scoring. Further, selections were made for the two common sets, each of which included 25 essays. Selections for the ten practice benchmarks and the two common sets were made in the same way as the training benchmarks. Please see the tables in Appendix A for a full accounting of the practice benchmarks and the common set essays.

Training Methods

This study began with training. Raters assigned to either training condition studied the same sets of training benchmarks, as reflected in Appendix A. The only difference between the

two groups was the training method. Self-paced and collaborative frame-of-reference training methods are two of the most common types of rater training used in large-scale writing assessments (Johnson, Penny, & Gordon, 2009). Because the training materials in both methods were assembled by experts, it follows that the objective of these training methods was to help raters score like experts on the writing assessment. These methods have a great deal in common, but there is also a key difference. Beginning with the commonalities, the sequence of activities in both training methods was identical. Raters first received training in the ideas domain, followed by the organization, style, and conventions domains. This training consisted of first studying training benchmarks in the domain in question. Associated with each benchmark was a brief annotation explaining why the benchmark illustrated a particular score point on the rubric. After studying the benchmarks in a particular domain, raters scored a set of ten practice benchmarks in that domain. The key difference between the self-paced and collaborative methods was the frequency and format of the discussions between raters and their trainer about the training and practice benchmarks. Figures 1 and 2 reflect the complete sequence of activities for the self-paced and collaborative training methods, respectively.

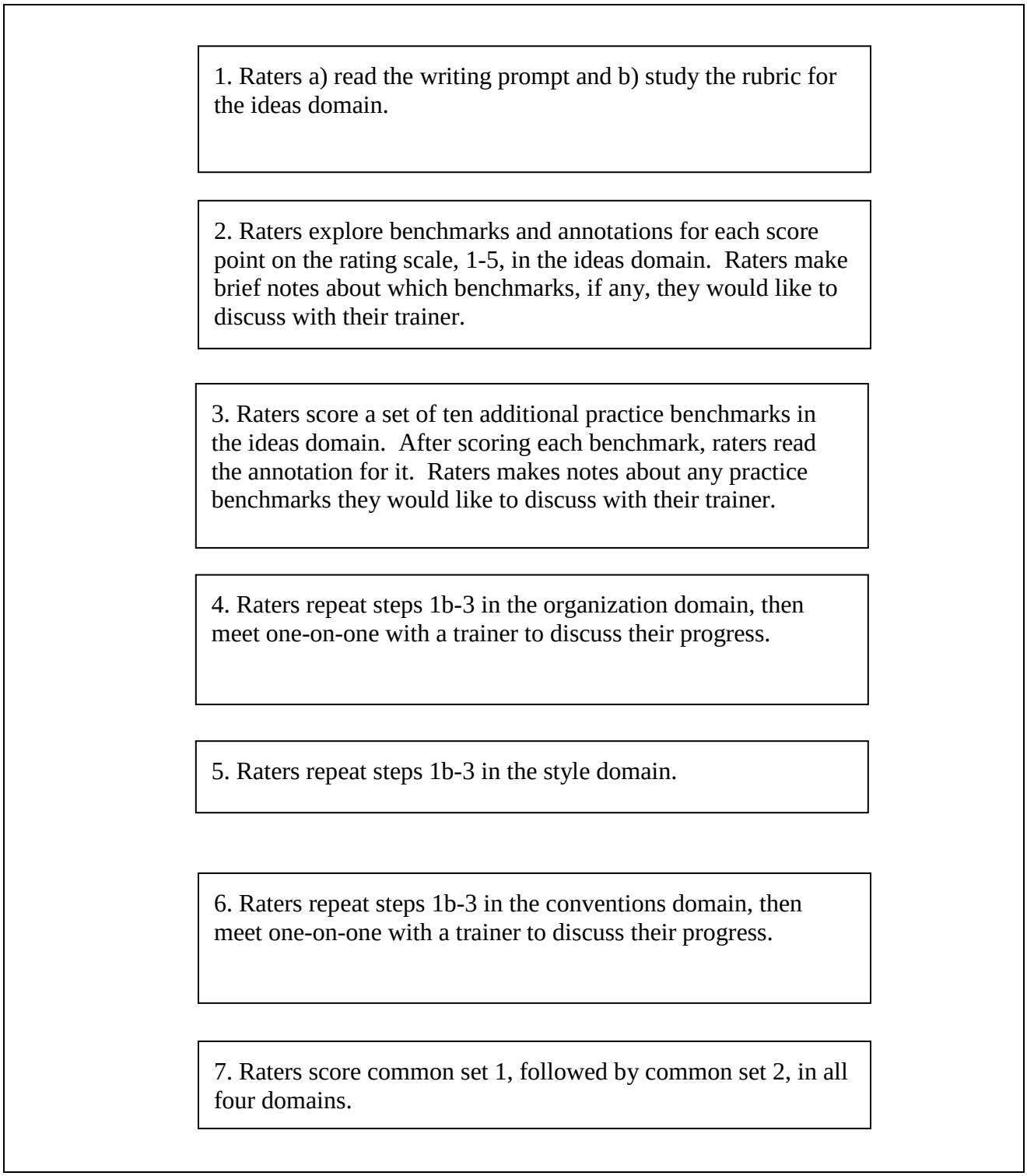


Figure 1. Sequence of self-paced training

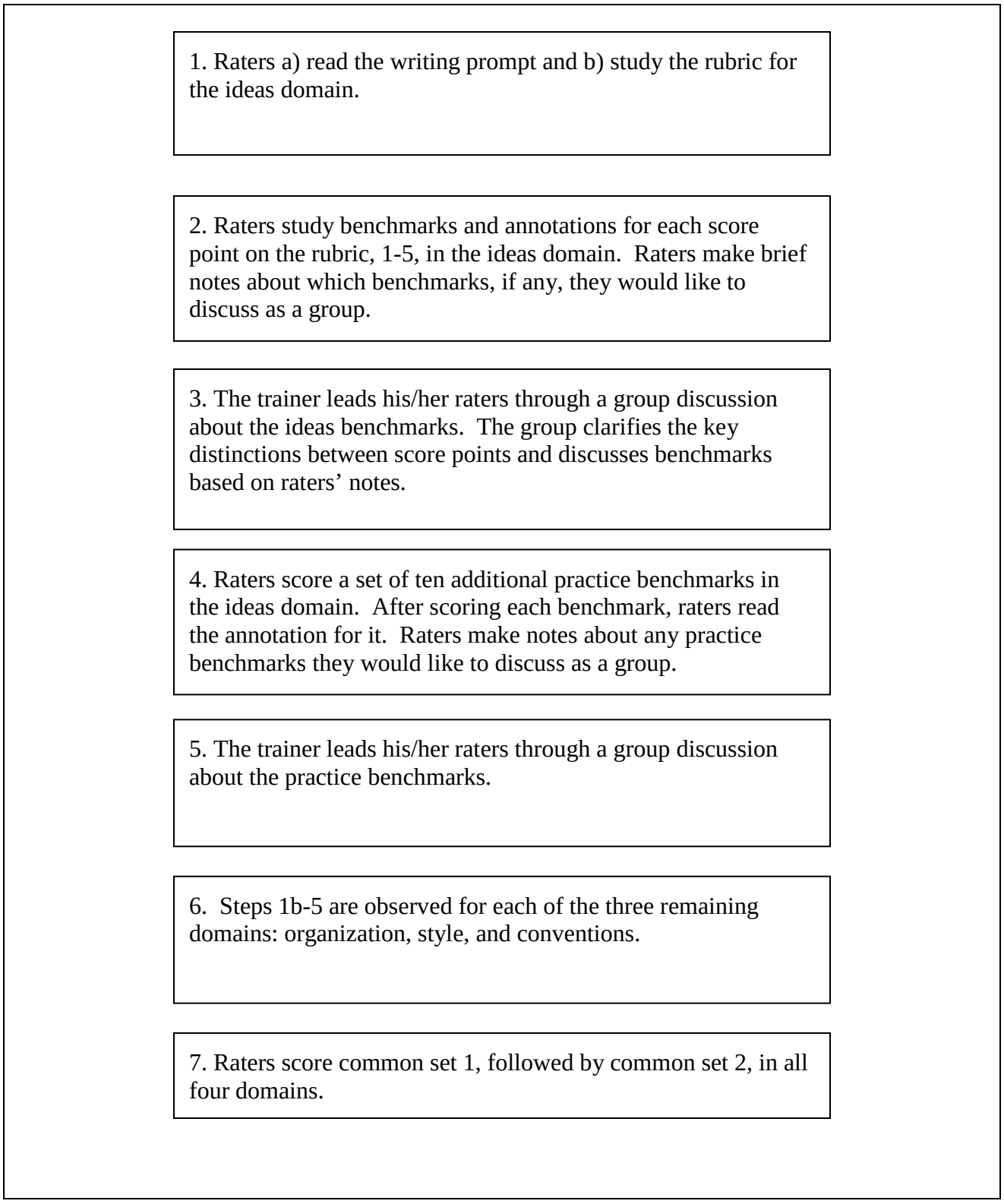


Figure 2. Sequence of collaborative training

In the self-paced method, as Figure 1 reflects, raters studied the benchmarks and scored practice benchmarks on their own, then met with a trainer, one-on-one, at two designated times to discuss their progress. To keep these meetings focused, raters completed a notes table in which they catalogued any benchmarks they wanted to discuss with the trainer. This notes table is included in Appendix B. In the collaborative method, by contrast, the entire cohort of raters met at designated times, as noted in Figure 2, to discuss the benchmarks they had studied and the practice benchmarks they had scored. In preparation for these discussions, raters in the collaborative method also completed the notes table found in Appendix B. During these discussions, facilitated by the cohort's trainer, raters could ask questions and make comments, and the entire cohort was encouraged to participate. Therefore, unlike raters who received self-paced training, raters who received collaborative training had the opportunity to learn from both their trainer and the other raters in the cohort. The trainer sought to keep the discussions constructive and concise and clarified misconceptions about rubric application, as necessary. After training, raters from both training methods scored common set 1, followed by common set 2. Both common sets had 25 essays. Rating data from these common sets were used to determine rater accuracy.

Scoring and Data Collection

After training, all raters proceeded to scoring. As noted, the scoring window began with all raters scoring common set 1, followed by common set 2. Because there were 25 essays in each common set, all raters scored 50 common essays. The rating data of interest are raters' scores on these 50 essays. The resulting data set is fully crossed: all raters scored all essays of interest. Upon completing the two common sets, raters scored the remaining essays associated with the Georgia Supplemental Writing Assessment. Scoring lasted approximately 1.5 days.

Data Set and Dependent Variable

Because raters scored each essay in four domains (i.e., aspects) of writing, the data file contains four scores that all 62 raters assigned to each of the 50 common set essays. These raw scores in the data set were converted to accuracy scores by comparing the rater's scores to the resolved expert scores. In this dissertation, accuracy is defined as it is described in Engelhard (1996) and Sulsky and Balzer (1988): the degree to which raters and experts agree exactly when making a judgment, such as rating an essay. Defined in this way, accuracy is arguably the most direct measure of rater proficiency because experts are assumed to have a sound theoretical understanding of the construct and considerable experience applying the rubrics to essays, both of which hold true for the Georgia Supplemental Writing Assessment experts who scored and selected the training benchmarks and common set essays. Yet even when accuracy is defined in this way, converting raw scores to measures of accuracy is not a trivial matter. In fact, Sulsky and Balzer (1988) describe several ways that rater accuracy measures could be calculated, such as subtracting a rater's score from an expert's, resulting in what McIntrye, Smith, and Hassett (1984) call a *leniency* index for rater *k*, which may be expressed,

$$Leniency_k = \frac{\sum_{j=1}^{n} \frac{\left(\sum_{i=1}^{d}\left(t_{ij} - r_{ijk}\right)\right)}{d}}{n}, \quad (1)$$

where,

j = the number of essays, $j = 1,\ldots, J$;

d = the number of domains;

r_{ijk} = the score assigned in domain i on essay j by rater k;

t_{ij} = the resolved expert's score in domain i on essay j;

n = the number of examinees, $n = 1,\ldots, N$.

An index of zero denotes perfect accuracy, meaning that the rater and expert assigned the same score, on average. Positive values indicate that a rater is severe relative to the expert; negative values indicate leniency. A potential issue with the leniency index, however, is its sensitivity to raters whose severity fluctuates. Table 2 illustrates such fluctuation.

Table 2

Resolved Expert Scores Versus Rater Scores on Four Essays

Essay	Expert scores	Rater scores
1	2222	3333
2	3333	2222
3	4445	3334
4	4333	5444

On half the essays the rater was severe and on the other half lenient. Using Equation 1, the leniency index for this rater would be 0; in other words, this rater would appear to be perfectly accurate, and neither severe nor lenient. Visual inspection of the rater's scores in Table 2 suggests otherwise. A solution to this issue is to take the absolute value of the difference between the expert's scores and rater's scores, which yields a measure that McIntyre, Smith, and Hassett (1984) refer to as *distance accuracy* and may be expressed,

$$Distance\ Accuracy_k = \frac{\sum_{j=1}^{N} \frac{\left(\sum_{i=1}^{d} \left| t_{ij} - r_{ijk} \right|\right)}{d}}{n}, \quad (2)$$

where *k*, *j*, *n*, *d*, *i*, *t*, and *r* are defined as in Equation 1. The distance accuracy index does not illustrate whether a rater is severe or lenient relative to the expert, but neither is it subject to the concern reflected in Table 2. Therefore, the accuracy measure used in this dissertation is a

variation on the distance accuracy measure reflected in Equation 2. Instead of calculating an overall index, however, distance accuracy measures were calculated at the domain level (*i*), using the following equation:

$$\textit{Distance Accuracy}_i = \left|t_{ij} - r_{ijk}\right|, \quad (3)$$

where,

t_{ij} = the resolved expert score assigned in domain *i* on essay *j*;

r_{ijk} = the score assigned in domain *i* on essay *j* by rater *k*.

Because each essay was scored in four domains, the data file contains four distance accuracy scores for each rater on each of the 50 common set essays. A distance accuracy value of 0 indicates perfect accuracy, and higher values denote greater inaccuracy. It is worth elaborating on why domain-level distance accuracy scores, rather than an overall accuracy index, were calculated. An overall index can mask potentially important aspects of a rater's accuracy or lack thereof (Cronbach, 1955; Engelhard, 1996; Sulsky & Balzer, 1988). For instance, if a rater had a relatively low overall accuracy index, it would not be clear from the overall index whether the rater was equally inaccurate across all domains or accurate in some while inaccurate in others.

Upon conversion from raw scores at the domain level to distance accuracy scores at the domain level, the data set for each of the 62 raters contained 200 distance accuracy scores: 50 essays × 4 distance accuracy scores per essay. For each of the four domains, therefore, the data file contains 3,100 distance accuracy scores. Figures 3-6 show the distribution of the distance accuracy scores in the ideas, organization, style, and conventions domains, respectively. These are labeled AccIdeas, AccOrg, AccSty and AccConv, respectively.

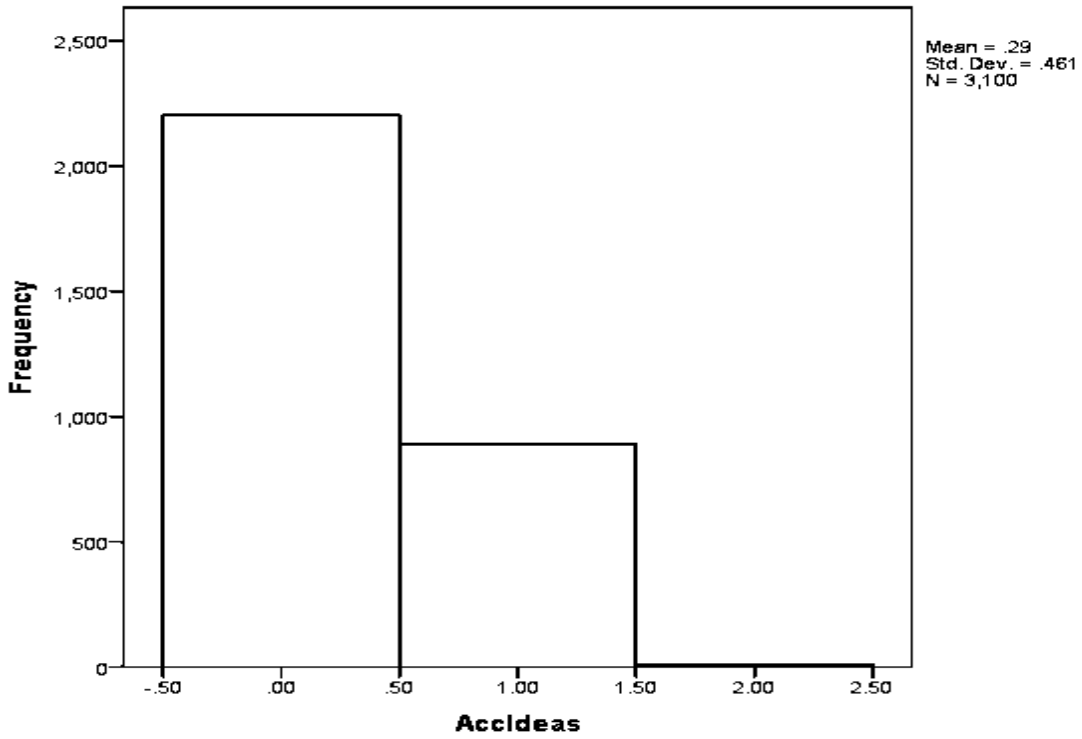


Figure 3. Distribution of distance accuracy scores in the ideas domain.

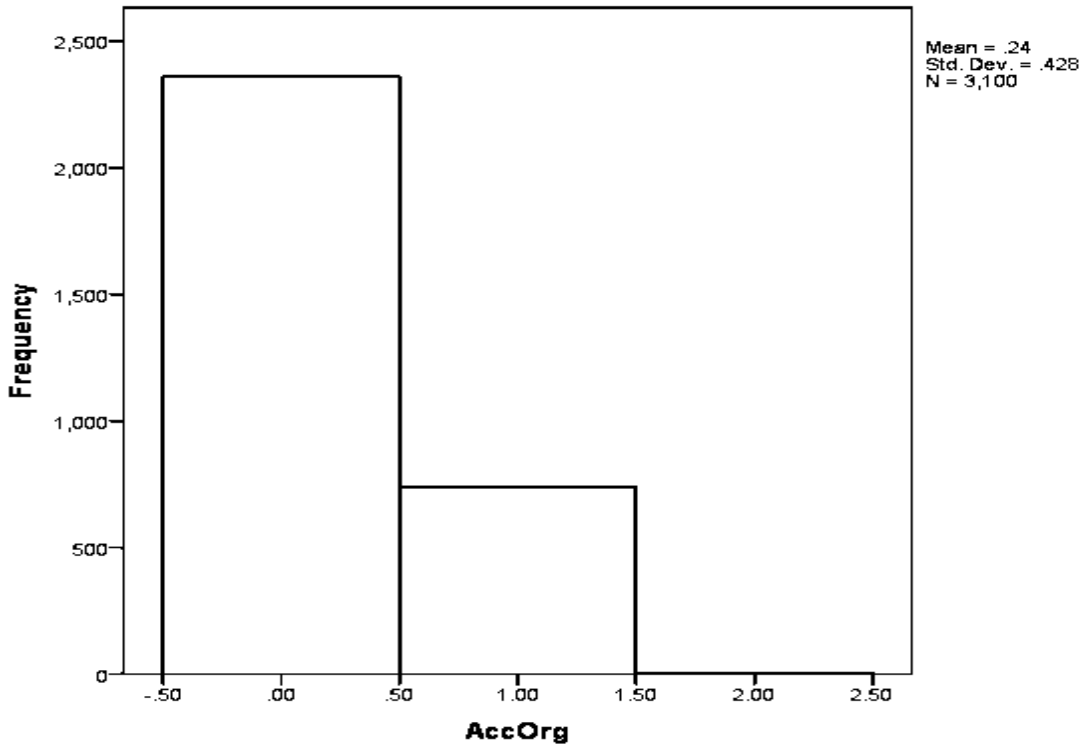


Figure 4. Distribution of distance accuracy scores in the organization domain.

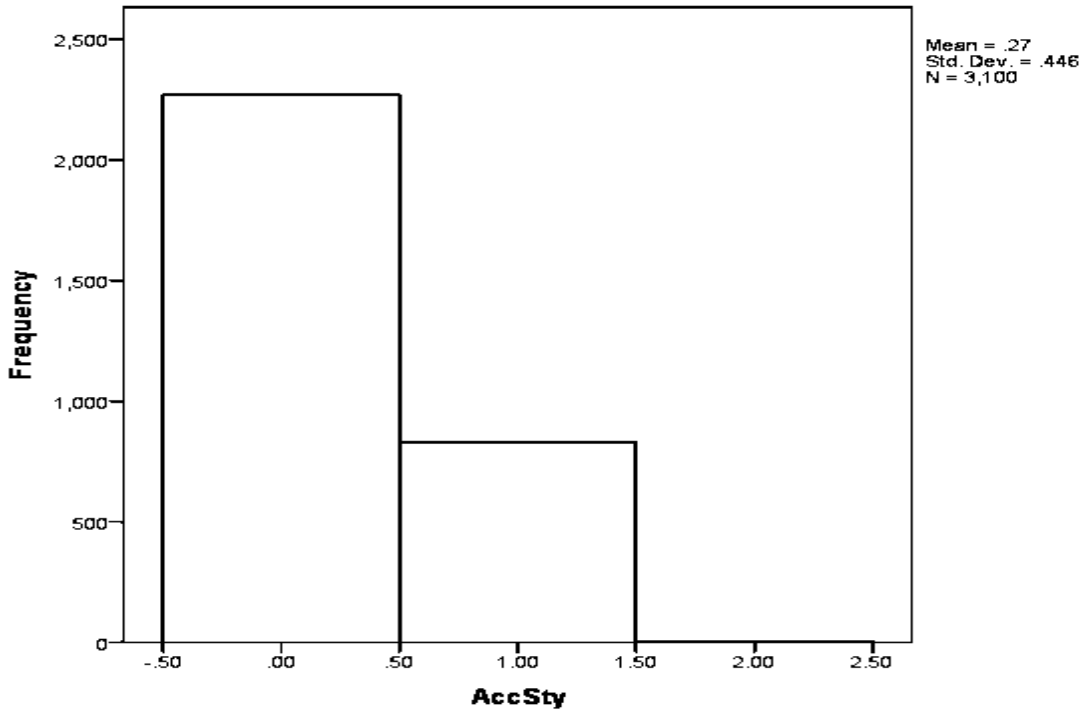


Figure 5. Distribution of distance accuracy scores in the style domain.

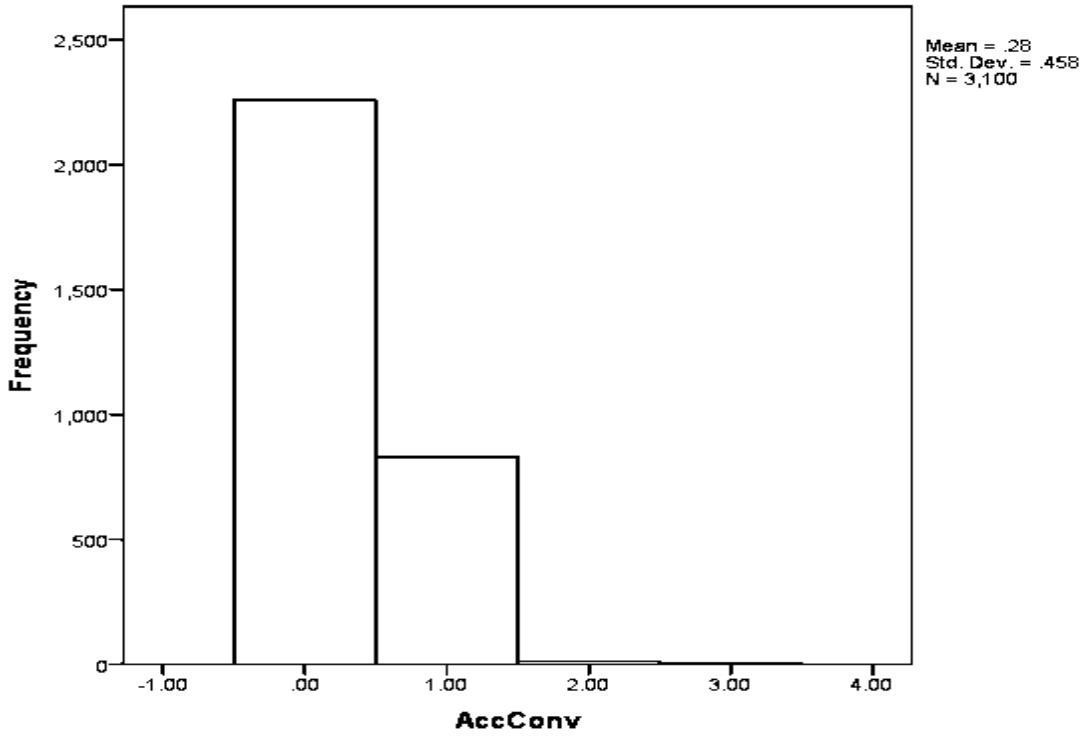


Figure 6. Distribution of distance accuracy scores in the conventions domain.

From the histograms in Figures 3-6, it is clear that the data follow what are, essentially, binomial distributions. Most of the distance accuracy scores in each domain take a value of zero, denoting no difference between the rater's score and the resolved expert score. A smaller proportion of the observations take a value of 1. Less than one-half of one percent of the observations, in any of the four domains, take a value of 2. Therefore, the distance accuracy scores were treated as binary, and the observations were coded as follows: when the distance accuracy score was zero, the observation was coded as 1. Otherwise, the observation was coded 0. In other words, a value of 1 denotes an accurate rating, whereas a value of 0 signifies an inaccurate rating. Engelhard (1996) took a similar approach. These accuracy measures constitute the dependent variable in this dissertation.

Covariates and Research Questions

Two covariates were included to predict rater accuracy: the training condition the rater received and a variable called essay type, which warrants further explanation. As reflected in Appendix A, the experts did not always agree unanimously on the essays selected for the training benchmarks or the common sets. For an essay to be selected as a training benchmark or a common set essay, at least three of the four experts needed to agree on the score, and this score constituted the resolved expert score. In this dissertation, essay type denotes the maximum absolute difference between the scores that each of the four experts *independently* assigned to the essay in question. Therefore, essay type took a value of 0 when this difference was zero. If the maximum absolute difference was 1, essay type took a value of 1, and so on. This approach allows for exploration of the possibility that some essays may be more difficult to score accurately than others (Engelhard, 1996). That is, as experts' independent scores differ to an increasing degree, it may be inferred that the essay will be more difficult for *raters* to score accurately. Given these covariates and the dependent variable, three research questions were investigated:

1) Is there inter-rater variation in accuracy on a common set of 50 essays?
2) How does a rater who received self-paced training compare in accuracy to a rater who received collaborative training?
3) What is the strength of association between accuracy and essay type, and is this slope the same for raters who received self-paced training and raters who received collaborative training?

The same three research questions were explored independently for each of the four domains.

Power Analysis

To explore these research questions, a series of logistic mixed models were fitted to the accuracy scores at the domain level. Before describing the models, it is appropriate to discuss power, given that the results of a power analysis directly affect the modeling approach. Power is the probability of rejecting a null hypothesis that is, in fact, false (Hox, 2010). Said differently, it is the probability of detecting an effect of a particular size. An *a priori* power analysis in the context of a hierarchical model, such as a logistic mixed model, is complex because power depends not only on effect size and significance level but also on whether the explanatory variables subjected to hypothesis testing occur at level 1, level 2, or higher levels within the hierarchical model (Snijders, 2005). Further, the power analysis depends on whether the effects of interest are fixed or random (Snijders, 2005). As will be discussed during the presentation of the models, of principal interest in this dissertation were the fixed effects of training condition, a level-2 predictor, and essay type, a level-1 predictor, on rater accuracy. Given that 62 raters were available as research participants, a logical next step was to investigate the statistical power of detecting fixed effects of a particular size in a multi-level design with only 62 raters available. The Optimal Design software (Raudenbush, Spybrook, Congdon, Liu, Martinez, Bloom, & Hill, 2011) was used for this purpose. First, a design where accuracy scores are nested within raters was explored. Figure 7 shows the power curves for design 1. As reflected in the legend in the top right corner of Figure 7, the significance level (α) was set to 0.1, and three different effect sizes (δ) were selected: 0.2, 0.5, and 0.8.

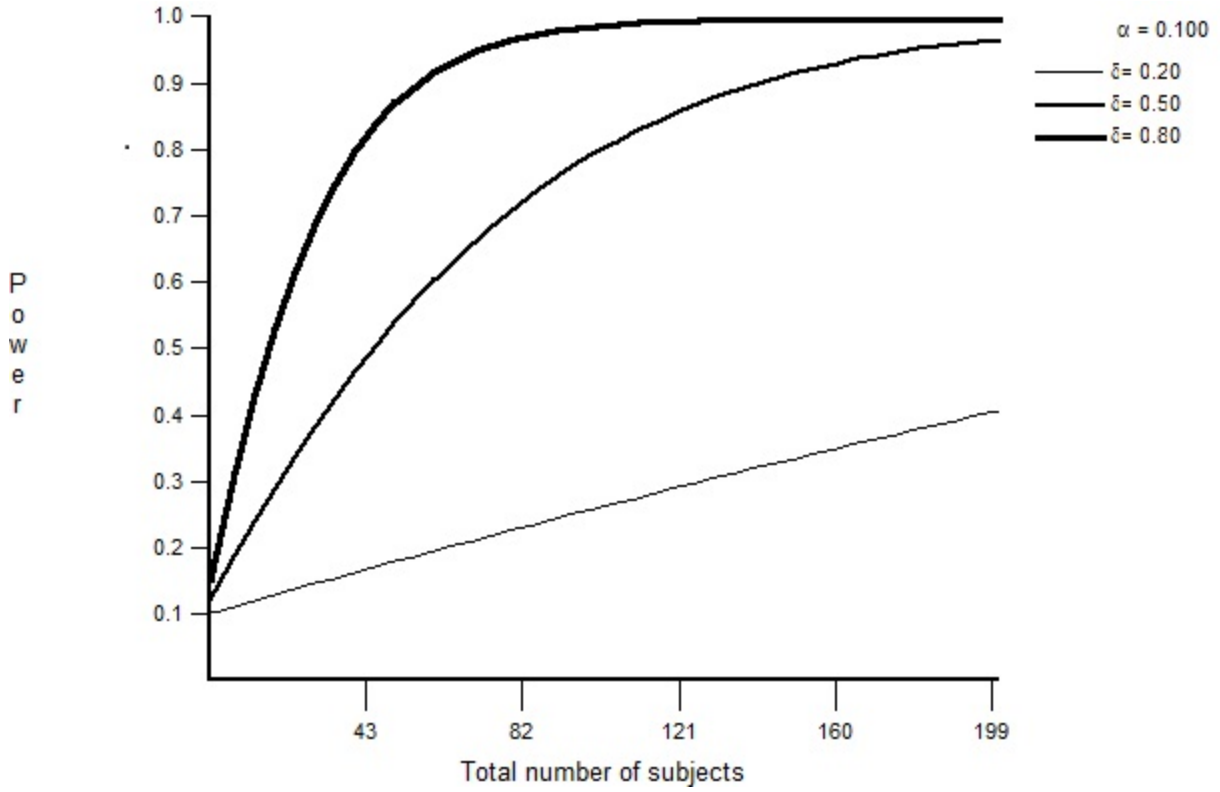


Figure 7. Power curves for design 1: Accuracy scores nested within raters.

t-tests are used in this dissertation for comparing the mean accuracy (i.e., the intercept) and change in accuracy based on essay type (i.e., the slope) for raters in self-paced versus collaborative training methods. The power curves yield information on how much power the *t*-tests would need to have to detect effects of different sizes, given the sample size. The bottom-most trajectory in Figure 7 shows that with only 62 raters, the *t*-tests are underpowered ($1 - \beta = .2$) to detect a small effect, where the standardized mean difference in intercepts and slopes between training methods is 0.2 or less (Cohen, 1988). The middle trajectory reveals that power improves to 0.66 for detecting a medium effect of .5 (Cohen, 1988). Finally, the top-most trajectory reflects very high power ($1 - \beta = .94$) for detecting a large effect of 0.8 (Cohen, 1988).

In this design, there is sufficient power to detect medium and high standardized differences in intercepts and slopes.

However, design 1 ignores a key fact of the Georgia Supplemental Writing Assessment protocol: raters are further divided into cohorts, led by a trainer. Therefore, there is a possibility of a cohort effect on a rater's estimated accuracy. To account for this possibility, a second design is one with three levels: accuracy scores nested within raters nested within cohorts. The power curves for this cluster design (design 2) were created using the Optimal Design program (Raudenbush et al., 2011) and are shown in Figure 8. As reflected in the legend in the top right corner of Figure 8, the significance level (α) was set to 0.1. Three different effect sizes (δ) were selected: 0.2, 0.5, and 0.8. By default, the Optimal Design program sets to 0.05 and 0.1 the intra-class correlation (ρ), which, in the present context, denotes the correlation between the accuracy measures of raters within a cohort. The sample size was set to eleven raters per cohort.

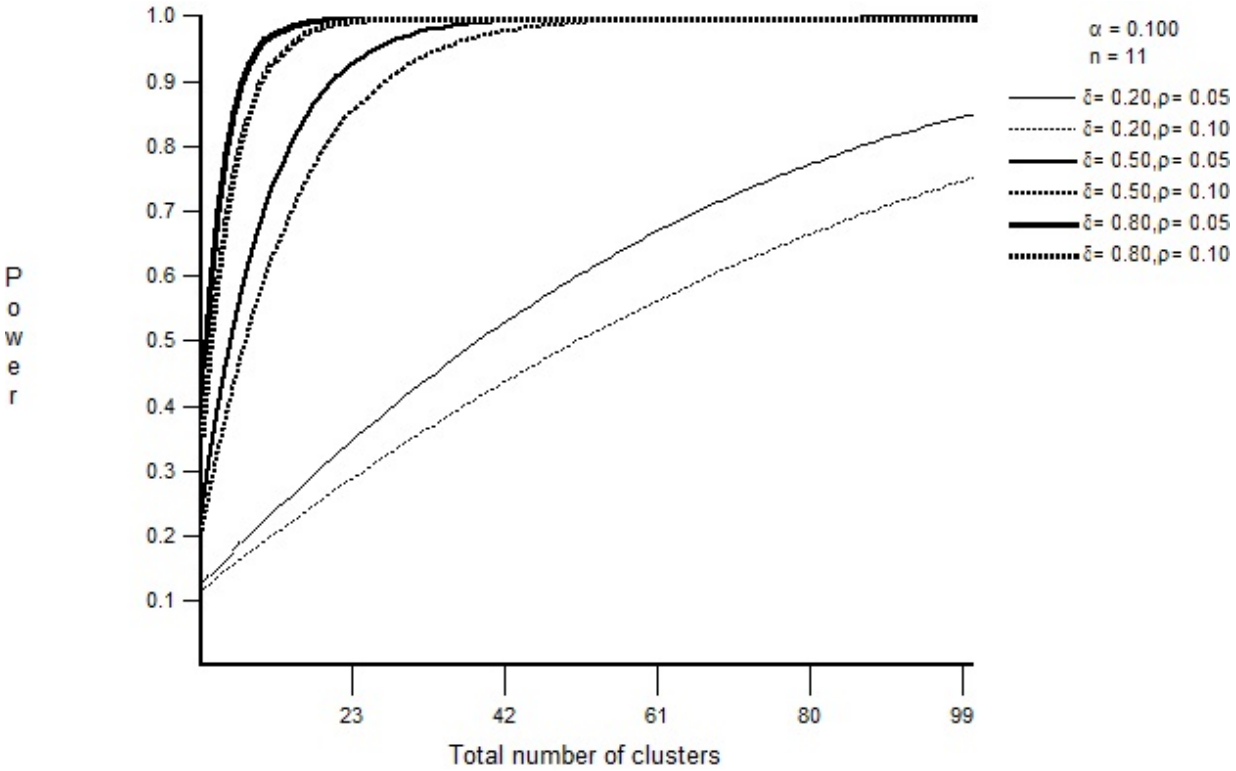


Figure 8. Power curves for design 2: The cluster design.

Again, of interest is the use of *t*-tests for comparing the intercepts and slopes of *cohorts* of raters that received self-paced versus collaborative training. The two bottom-most trajectories in Figure 8 show that with 11 raters per cohort (i.e., cluster) and a less conservative significance level of .1, the *t*-tests have very low power ($1 - \beta = .14$) to detect a small effect, where the standardized mean differences in intercepts and slopes between training methods is .2 or less (Cohen, 1988). As the pairs of middle and top-most trajectories illustrate, power improves to .35 and .6 when the effect sizes increase to .5 and .8, which, respectively, are medium and large effects (Cohen, 1988). This design is not optimal, given the number of raters available. It would be important to investigate potential cohort effects on rater accuracy, but the available sample size is too small to detect these potentially important effects.

Measurement Models: Logistic Mixed Models

Given the results of the power analysis, a series of 2-level logistic mixed models were fitted to the accuracy scores at the domain level. These data have the following hierarchical structure: accuracy scores nested within raters. Logistic mixed models can be fitted to hierarchical data where the dependent variable is not distributed as continuous normal. Rather, the dependent variable is categorical in nature. Attempting to fit *general* linear mixed models to account for the clustering would be problematic for several reasons, as outlined in Raudenbush and Bryk (2002). These reasons include the fact that the level-1 residuals cannot be normally distributed—an assumption that general linear mixed models make—because of the restricted range of the observed data. For instance, in a binary case, the observed data take one of only two values, while the estimated data are unbounded in the context of general linear mixed models. A more appropriate model for predicting the outcome would thus limit the predicted values to a (0,

1) interval (Raudenbush & Bryk, 2002). Logistic mixed models, alternatively called generalized linear mixed models, do exactly this.

A level-one logistic mixed model is made up of three parts: a sampling model, a link function, and a structural model (Raudenbush & Bryk, 2002). The sampling model for binary data like the outcomes in this dissertation is a binomial sampling model, which may be expressed,

$$Y_{ij}|\phi_{ij} \sim B(m_{ij}, \phi_{ij}). \quad (4.1)$$

Relative to Equations 1-3, the subscripts take on different meanings in Equation 4.1. The level-one outcome is assumed to follow a binomial distribution (*B*), where m_{ij} indexes the number of trials (i.e., essays) *i*=1,…,*N*, scored by rater *j*=1,…,*N*, and ϕ_{ij}indexes the probability of success (i.e., accuracy) on trial *i* for rater *j* (Raudenbush & Bryk, 2002). The second part of the level-one model, the link function, ensures that predicted values are constrained to a given interval (Raudenbush & Bryk, 2002). When this interval is (0, 1), as is the case for the outcomes in this dissertation, the appropriate link function is the logit link function, which may be expressed,

$$\eta_{ij} = \log\left(\frac{\phi_{ij}}{1-\phi_{ij}}\right), \quad (4.2)$$

where,

η_{ij}= the log-odds of success (i.e., accuracy) on essay *i* for rater *j*.

By expressing the *probability* of success as the *log-odds* of success, the outcomes have a theoretical range of $(-\infty, \infty)$ and can now be related to predictors through a linear structural model, which is the final component of the level-one model.

The structural model may be expressed,

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \cdots + \beta_{Qj}X_{Qij}, \quad (4.3)$$

where,

$\beta_{0j} =$ the intercept;

$\beta_{qj}(q = 0,1, \ldots, Q) =$ the level-1 coefficients;

$X_{qij}(q = 0,1, \ldots, Q) =$ the level-1 predictor q for essay i scored by rater j.

The structural model in 4.3 follows the familiar form of general linear models and linear mixed models, with the exception that the predicted outcome has been transformed. It should be noted that the level-two logistic mixed model is expressed in the same manner as the level-two general linear mixed model. The specific logistic mixed models used in this dissertation will be presented in turn.

To address research question 1 about whether there is inter-rater variation in accuracy, an unconditional 2-level logistic mixed model was used. At level-1, the log-odds of rater *j* scoring essay *i* accurately is a function of the intercept: the overall log-odds of scoring essays accurately (i.e., the grand mean). This model is expressed,

$$\eta_{ij} = \beta_{0j}. \quad (5.1)$$

It should be noted that the level-one error variance in a logistic mixed model is not estimated but is assumed to follow a logistic distribution with a mean of zero and a known variance of 3.29 (Raudenbush & Bryk, 2002; Templin, 2012). Therefore, the level-one random effect is not included in the model. At level-2, β_{0j} is extended to include a random effect for rater j. This model is expressed,

$$\beta_{0j} = \gamma_{00} + u_{0j}. \quad (5.2)$$

Substituting Equation 5.2 into 5.1 yields the composite model, which is expressed as,

$$\eta_{ij} = \gamma_{00} + u_{0j}, \tag{5.3}$$

$$u_{0j} \sim N(0, \tau_{00})$$

where,

η_{ij}= the log-odds that rater *j* will score essay *i* accurately;

γ_{00}= the overall log-odds of accuracy for a rater for whom u_{0j}= 0;

u_{0j} = the random effect of rater *j*.

The random effect u_{0j} provides an estimate of how raters vary in overall log-odds of accuracy, thereby addressing research question 1.

To address research question 2 about differences in rater accuracy as a function of training method, a categorical covariate (W_j) is added to the unconditional model at level-2. The covariate W_j denotes the training condition and takes a value of 0 for the self-paced training condition and a value of 1 for the collaborative training condition. The level-1 model is identical to Equation 5.1. At level-2, $\beta_{0j}$is extended to include W_j and a random effect for rater *j*. The model is expressed,

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}. \tag{6.1}$$

Substituting Equation 6.1 into 5.1, yields the composite model, which is expressed as

$$\eta_{ij} = \gamma_{00} + \gamma_{01} W_j + u_{0j}, \tag{6.2}$$

$$u_{0j} \sim N(0, \tau_{00})$$

where,

η_{ij}= the log-odds that rater *j* will score essay *i* accurately;

γ_{00}= the overall log-odds of accuracy for a rater who received self-paced training;

γ_{01}= the overall log-odds of accuracy for a rater who received collaborative training ($W_j = 1$), relative to self-paced training ($W_j = 0$);

u_{0j} = the random effect of rater *j*, conditioning on W_j.

The fixed effects in 6.2 (γ_{00} and γ_{01}) provide evidence pertinent to the second research question: How does a rater who received self-paced training compare in the log-odds of accuracy to a rater who received collaborative training? The null hypothesis related to this research question is one of no difference in the log-odds of accuracy. If the estimate for $\gamma_{01}$is significant, there would be evidence to reject the null hypothesis.

The third research question about change in rater accuracy as a function of essay type is addressed by adding a slope variable Z_{ij} to the model in Equation 6.2. This variable is the independent variable for essay type, which allows for the modeling of change in the log-odds of accuracy for each unit increase in essay type. At level-1, the predicted log-odds of accuracy is a function of the intercept and the change in the log-odds of scoring essay *j* accurately for each unit increase in Z_{ij}. The level-1 model is expressed,

$$\eta_{ij} = \beta_{0j} + \beta_{1j} Z_{ij}. \tag{7.1}$$

At level-2, $\beta_{0j}$is extended to include W_j and a random effect for rater *j*. Further, $\beta_{1j}$is extended to include an effect for the interaction between the type of training the rater received and the type of essay being scored. The level-2 model is expressed,

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \tag{7.2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j.$$

Substituting Equation 7.2 into equation 7.1 yields the composite model, which is expressed as,

$$\eta_{ij} = \gamma_{00} + \gamma_{10}Z_{ij} + \gamma_{01}W_j + \gamma_{11}W_jZ_{ij} + u_{0j} \tag{7.3}$$

$$u_{0j} \sim N(0, \tau_{00})$$

where,

η_{ij}= the log-odds that rater *j* will score essay *i* accurately;

γ_{00}= the overall log-odds of accuracy for a rater who received self-paced training;

γ_{01}= the overall log-odds of accuracy for a rater who received collaborative training ($W_j = 1$), relative to self-paced training ($W_j = 0$);

γ_{10}= the change in log-odds of accuracy for each unit increase in Z_{ij} for a rater who received self-paced training;

γ_{11}= the change in log-odds of accuracy for each unit increase in Z_{ij} for a rater who received collaborative training ($W_j = 1$) relative to self-paced training ($W_j = 0$);

u_{0j} = the random effect of rater *j*, conditioning on W_j.

Of particular interest is the fixed effect γ_{11}. Hypothesis testing related to this fixed effect provides information pertinent to research question 3: What is the strength of association between the log-odds of accuracy and essay type (i.e., the slope), and is this slope the same for

raters who received self-paced training relative to collaborative training? The null hypothesis related to this research question is one of no difference in the slope for a rater who received either self-paced or collaborative training. If the estimate for $\gamma_{11}$is significant, there would be evidence to reject the null hypothesis. It should be noted that a random effect for the slope will be added to 7.3 only if it improves fit.

CHAPTER 4

RESULTS

Data Analysis

Estimation of the fixed and random effects in Equations 5.3, 6.2, and 7.3 was done using the HLM software, version 7 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011). Because these models differ in fixed effects, full maximum likelihood estimation (ML) was specified for all analyses. ML allows for model-fit comparisons when the models differ in fixed effects (Templin, 2012). The objective of ML is to use an iterative process to obtain estimates of the fixed and random effects that make the observed data most likely (Fitzmaurice, Laird, & Ware, 2011). ML estimation of logistic mixed-models, which have discrete rather than continuous outcomes, is challenging because the level-1 error variance is not assumed to be normally distributed. Therefore, an approximation of ML is used (Raudenbush & Bryk, 2002). There are several approximations that could be used, including quasi-likelihood inference, Gauss-Hermite approximations, and the Laplace approximation (Raudenbush, Yang, & Yousef, 2000). In this dissertation, the Laplace approximation was used, as implemented in the HLM software, because research has shown that it produces accurate parameter estimates (Raudenbush, Yang, & Yousef, 2000). All analyses were run at the domain level. The results are reported in four sections: one section for each of the four domains.

The Ideas Domain

In the ideas domain, the examinee's essay is rated on its level of focus and development with respect to the assigned prompt. There is a specific rubric raters used to evaluate each essay in the ideas domain, and the rating scale for the rubric ranges from 1-5. Higher scores on the rating scale denote greater focus and deeper development. Appendixes C and D show descriptive statistics and the distribution of scores assigned by experts and raters, respectively, in the ideas domain to the 50 common set essays. The mean raw score in the ideas domain for both experts and raters was 3, and the standard deviation was approximately 1. The experts and raters assigned approximately the same proportion of 1s, 2s, 3s, 4s, and 5s in the ideas domain. These distributions suggest that the essays in the common sets reflected a range of proficiency in the ideas domain. Ultimately, raters' raw scores in the ideas domain were compared to the resolved experts' scores and then converted to distance accuracy scores using Equation 3. From these distance accuracy scores, each observation was then coded 1 (accurate) if the distance accuracy score was zero, meaning there was no difference between the resolved expert score and the rater's score. If the distance accuracy score was 1, the observation was coded 0 (inaccurate). The logistic mixed models in Equations 5.3, 6.2, and 7.3 were then fitted to these binary data. Results from fitting these models are presented in turn.

The results of fitting the model in Equation 5.3, the unconditional model with the log-odds of accuracy in the ideas domain as the dependent variable, are presented in Table 3.

Table 3

Results of Fitting the Unconditional Model (Ideas Domain)

Parameter	Estimate	SE	*p*
γ_{00}	0.90	0.05	<.001*
τ_{00}	0.04	0.02	.03*

Note. -2 log *L* = 9,425.5
Note. * $p < .05$

The sequence of HLM software commands used to obtain these results can be found in Appendix E. The unconditional model contains no predictors. The purpose of fitting this model was to explore whether raters varied in overall log-odds of accuracy in the ideas domain, thereby addressing research question 1 for this domain. The fixed effect γ_{00} demonstrates that, overall, a rater's log-odds of scoring essays accurately in the ideas domain was 0.9 logits. This estimate differs significantly from zero logits, which, expressed as a probability, denotes a fifty percent probability of accuracy. This can be shown by expressing a logit value (η_{ij}) of 0 as a probability, using the following equation,

$$\phi_{ij} = \frac{1}{1+\exp(-\eta_{ij})}. \quad (8)$$

By contrast, inserting the predicted value of γ_{00} (0.9) for η_{ij} and solving Equation 8 results in an overall probability of .71 that rater *j* will score essays accurately in the ideas domain. The variance around the overall log-odds of accuracy was 0.04 logits, as indicated by the estimate for τ_{00}. This estimate was significant, meaning there was a rater effect on the log-odds of accuracy in the ideas domain. Said differently, when there are no predictors in the model, raters were not equivalent in their estimated log-odds of scoring essays accurately in the ideas domain.

To address the second research question about differences in rater accuracy as a function of training method, the model in Equation 6.2, with training method as a categorical covariate, was fitted to the data. Results are found in Table 4.

Table 4

Results of Fitting the Conditional Model from Equation 6.2 (Ideas Domain)

Parameter	Estimate	SE	*p*
γ_{00} (Self-paced)	0.909	0.07	<.001*
γ_{01}(Collab.)	-0.009	0.10	.93
τ_{00}	0.036	0.03	.02*

Note. -2 log *L* = 9,425.49
Note. * $p < .05$

The sequence of HLM software commands used to obtain these results can be found in Appendix E. Beginning with the fixed effects, the estimate for γ_{00} takes on a different meaning in this model. It now signifies the overall log-odds of scoring essays accurately in the ideas domain for a rater who received self-paced frame-of-reference training. The estimate for γ_{00} was 0.909 logits, which differs significantly from 0 logits. By comparison, the fixed effect γ_{01} reflects the overall log-odds of scoring essays accurately in the ideas domain for a rater who received collaborative frame-of-reference training. This estimate, 0.9 logits, is obtained by summing γ_{00} and $\gamma_{01.}$ Relative to a rater who received self-paced training, a rater who received collaborative training had slightly smaller log-odds of accuracy in the ideas domain— 0.009 logits, which was not significant. This difference does not provide sufficient evidence to reject the null hypothesis

that raters from both training conditions have equivalent log-odds of accuracy in the ideas domain.

As the estimate of the random effect τ_{00} in Table 4 illustrates, there was a significant rater effect on accuracy in the ideas domain, even after controlling for training condition.

To compare the fit of these two models, a deviance test was used. A deviance test, also known as a likelihood ratio test, can be used to compare the fit of models that are nested, meaning that a simpler, or reduced, model can be derived from a more complex, or full, model by removing parameters (Hox, 2010). The simpler model in Equation 5.3 can be derived from the model in Equation 6.2 by removing the fixed intercept parameter $\gamma_{01.}$ To conduct the deviance test, the deviance for the model in Equation 6.2 is subtracted from the deviance for the model in Equation 5.3. The deviance is expressed as "-2 × ln (*Likelihood*), where *Likelihood* is the value of the likelihood function at convergence" (Hox, 2010, p. 47). The difference in deviances follows, approximately, a chi-square distribution, with degrees of freedom equal to the difference in parameters estimated in the full model relative to the reduced model (Hox, 2010). As reported in Tables 3 and 4, the deviance for the model in Equation 5.3 and 6.2 was 9,425.5 and 9,425.49, respectively. This difference is 0.01. It is distributed, approximately, as a chi-square with one degree of freedom because the model in Equation 6.2 estimated one more parameter. The resulting chi-square is not significant at p =.05. The conditional model was thus rejected in favor of the unconditional model. However, attempting to fit the conditional model allowed for exploration of research question 2. There was no evidence of a training method effect on the overall log-odds of accuracy in the ideas domain.

To address the third research question about change in accuracy as a function of essay type, the model in Equation 7.3 was fitted to the binary data. The model in Equation 7.3

introduces the slope variable Z_{ij}, which allows for the modeling of change in the log-odds of accuracy in the ideas domain as a function of the type of essay the rater scored. Again, essay type, denoted by Z_{ij}, reflects the maximum absolute difference between the scores that each expert *independently* assigned in the ideas domain to the common set essay in question. That is, higher values of Z_{ij} highlight essays for which it was more difficult for experts to determine an accurate score in the ideas domain. The theoretical range of Z_{ij} in the ideas domain is 0-4, but the data reveal that the observed range was 0-1. The results of fitting the model in Equation 7.3 are reported in Table 5.

Table 5

Results of Fitting the Conditional Model from Equation 7.3 (Ideas Domain)

Parameter	Estimate	SE	*p*
γ_{00} (Self-paced)	1.31	0.12	<.001*
γ_{01} (Collab.)	-0.05	0.16	.77
γ_{10} (Self-paced)	-0.77	0.16	<.001*
γ_{11} (Collab.)	0.07	0.2	.74
τ_{00}	0.04	0.03	.01*

Note. -2 log L = 9,339.8
Note. * $p < .05$

The sequence of HLM software commands used to obtain these results can be found in Appendix E. Beginning with the fixed effect $\gamma_{00,}$ the log-odds of accuracy in the ideas domain for a rater who received self-paced training was 1.31 logits when Z_{ij} took a value of zero (i.e., the experts independently assigned the same score in the ideas domain). For each unit increase in Z_{ij}, the

log-odds went down by 0.77 logits. This is significantly different than zero logits, as reflected by the estimate for the slope parameter γ_{10}. This finding suggests that for a rater who received self-paced training, the log-odds of accuracy in the ideas domain was contingent on the type of essay being scored. A similar trajectory was found for raters who received collaborative training. The estimate for γ_{01} indicates that the log-odds of accuracy in the ideas domain for a rater who received collaborative training was 1.26 logits when Z_{ij} took a value of zero. This estimate is obtained by summing γ_{00} and $\gamma_{01.}$ For each unit increase in Z_{ij}, the log-odds went down by 0.7 logits. This estimate is obtained by summing γ_{10} and $\gamma_{11.}$ Compared to a rater who received self-paced training, a rater who received collaborative training was slightly less accurate when Z_{ij} was zero but slightly more accurate as Z_{ij} took higher values. It should be noted, however, that these differences were not significant, as reflected in Table 5. There is not sufficient evidence to reject the null hypothesis that raters from both training conditions have equivalent log-odds of accuracy as the essay type changes.

As the estimate of the random effect τ_{00} in Table 5 illustrates, there was a significant rater effect on accuracy in the ideas domain, even after controlling for both training condition and essay type.

To compare the fit of the models in Equations 5.3 and 7.3, a deviance test was used. The deviance for the models in Equations 5.3 and 7.3 was 9,425.5 and 9,339.8, respectively. This difference is 85.7. It is distributed, approximately, as a chi-square with three degrees of freedom because the model in Equation 7.3 estimated three more parameters. The resulting chi-square is significant at p <.05. The model in Equation 5.3 was thus rejected in favor of the model in Equation 7.3. That is, the model with the slope variable Z_{ij} predicted the log-odds of accuracy in the ideas domain better than the model with no predictors or the model with only the categorical

covariate for training condition. This finding, which provides evidence that some essays are more difficult to score accurately in the ideas domain, is consistent with the general findings of Engelhard (1996).

It should be noted that a model with a random effect for the slope was fitted to the data but rejected because it did not improve model fit. The deviance for this model was 9339.4, and it estimated two more parameters than the model in Equation 7.3. These parameters were a random effect for the slope and a covariance parameter between the intercepts and slopes. The difference in deviances between this model and the model in Equation 7.3 was 0.4. It is distributed, approximately, as a chi-square with two degrees of freedom because the model with a random effect for the slope estimated two additional parameters. The resulting chi-square is not significant at $p = .05$. Therefore, the model in Equation 7.3 was retained. While there was a rater effect on the intercept (i.e., the overall log-odds of accuracy), there was no rater effect on the slope. Raters did not vary significantly in the change in log-odds of accuracy for each unit increase in Z_{ij}.

As a final note, Appendix F features a normal quintile-quintile (Q-Q) plot of the level-2 random effect for the intercept from the model in Equation 7.3. In this plot, the distribution of the estimated level-2 random effect for the intercept (x-axis) is plotted against an expected normal distribution (y-axis). Because the points in this plot form approximately a straight line, there is evidence that the distribution of the level-2 random effect for the intercept (τ_{00}) was normal, as it is assumed to be.

The Organization Domain

In the organization domain, the examinee's essay is rated on the cohesiveness of the overall plan (i.e., its overall organizational structure). Again, there is a specific rubric raters used to evaluate each essay in the organization domain, and the rating scale for the rubric ranges from 1-5. Higher scores on the rating scale denote a more cohesive overall plan. Appendixes C and D show descriptive statistics and the distribution of scores assigned by experts and raters, respectively, in the organization domain to the 50 common set essays. The mean raw score in the organization domain for both experts and raters was 3, and the standard deviation was approximately 1. The experts and raters assigned approximately the same proportion of 1s, 2s, 3s, 4s, and 5s in the organization domain. These distributions suggest that the essays in the common sets reflected a range of proficiency in organization. Ultimately, raters' raw scores in the organization domain were compared to the resolved experts' scores and then converted to distance accuracy scores using Equation 3. From these distance accuracy scores, each observation was then coded 1 (accurate) if the distance accuracy score was zero, meaning there was no difference between the resolved expert score and the rater's score. If the distance accuracy score was 1 or 2, the observation was coded 0 (inaccurate). The logistic mixed models in Equations 5.3, 6.2, and 7.3 were then fitted to these binary data. Results from fitting these models are presented below.

The results of fitting the model in Equation 5.3, the unconditional model with log-odds of accuracy in the organization domain as the dependent variable, are presented in Table 6.

Table 6

Results of Fitting the Unconditional Model (Organization Domain)

Parameter	Estimate	SE	*p*
γ_{00}	1.16	0.05	<.001*
τ_{00}	0.003	0.02	.38

Note. -2 log L = 9,104.8
Note. * $p <$ .05

The sequence of HLM software commands used to obtain these results can be found in Appendix E. The unconditional model contains no predictors. The purpose of fitting this model was to address research question one: do raters vary in overall log-odds of accuracy in the organization domain? Beginning with overall accuracy in organization, the fixed effect γ_{00} demonstrates that, overall, a rater's log-odds of scoring essays accurately in the organization domain was 1.16 logits. This estimate differs significantly from zero logits, which, expressed as a probability, denotes a fifty percent probability of accuracy. This result is evident by inserting a value of zero for η_{ij} in Equation 8 and solving. By contrast, inserting the predicted value of γ_{00} (1.16) for η_{ij} and solving Equation 8 results in an overall probability of .76 that rater *j* will score essays accurately in the organization domain. The variance around the overall log-odds of accuracy was 0.003 logits, as indicated by the estimate for τ_{00}. This estimate was not significant, meaning there was not a rater effect on the log-odds of accuracy in the organization domain. Said differently, when there are no predictors in the model, raters were equivalent in their estimated log-odds of scoring essays accurately in the organization domain.

To address the second research question about differences in rater accuracy as a function of training method, the model in Equation 6.2, with training method as a categorical covariate, was fitted to the accuracy scores in the organization domain. Results are given in Table 7.

Table 7

Results of Fitting the Conditional Model from Equation 6.2 (Organization Domain)

Parameter	Estimate	SE	*p*
γ_{00} (Self-paced)	1.14	0.07	<.001*
γ_{01}(Collab.)	0.03	0.09	.75
τ_{00}	0.003	0.02	.35

Note. -2 log L = 9,104.7
Note. * $p <$ .05

The sequence of HLM software commands used to obtain these results can be found in Appendix E. Beginning with the fixed effects, the estimate for γ_{00} takes on a different meaning in this model. It now signifies the overall log-odds of scoring essays accurately in the organization domain for a rater who received self-paced frame-of-reference training. The estimate for γ_{00} was 1.14 logits, which differs significantly from 0 logits. By comparison, the fixed effect γ_{01} reflects the overall log-odds of scoring essays accurately in the organization domain for a rater who received collaborative frame-of-reference training. This estimate of 1.17 logits is obtained by summing γ_{00} and $\gamma_{01.}$ Relative to a rater who received self-paced training, a rater who received collaborative training had slightly greater log-odds of scoring essays accurately in the organization domain— 0.03 logits, which was not significant. This difference does not provide sufficient evidence to reject the null hypothesis that both groups of raters are equivalently accurate in the organization domain.

As the estimate of the random effect τ_{00} in Table 7 illustrates, there was not a significant rater effect on accuracy in the organization domain when controlling for training condition.

To compare the fit of the models in Equations 5.3 and 6.2, a deviance test was used. As reported in Tables 6 and 7, the deviance for both the unconditional and conditional models was

9,104.8 and 9,104.7, respectively. This difference is 0.1. It is distributed, approximately, as a chi-square with one degree of freedom because the model in Equation 6.2 estimated one more parameter. The resulting chi-square is not significant at p =.05. The conditional model was thus rejected in favor of the unconditional model. However, attempting to fit the conditional model allowed for exploration of research question 2. There was no evidence of a training method effect on the overall log-odds of accuracy in the organization domain.

To address the third research question about change in rater accuracy as a function of essay type, the model in Equation 7.3 was fitted to the accuracy scores in the organization domain. The model in Equation 7.3 introduces the slope variable Z_{ij}, which allows for the modeling of change in the log-odds of accuracy in the organization domain as a function of the type of essay the rater scores. As before, essay type, denoted by Z_{ij}, reflects the maximum absolute difference between the scores that each expert *independently* assigned in the organization domain to the common set essay in question. That is, higher values of Z_{ij} indicate essays for which it was more difficult for experts to determine an accurate score in the organization domain. The theoretical range of Z_{ij} in the organization domain is 0-4, but the data reveal that the observed range was 0-2. The results of fitting the model in Equation 7.3 are reported in Table 8.

Table 8

Results of Fitting the Conditional Model from Equation 7.3 (Organization Domain)

Parameter	Estimate	SE	*p*
γ_{00} (Self-paced)	1.41	0.1	<.001*
γ_{01} (Collab.)	0.18	0.13	.15
γ_{10} (Self-paced)	-0.6	0.12	<.001*
γ_{11} (Collab.)	-0.31	0.2	.05*
τ_{00}	0.007	0.03	.28

Note. -2 log L = 9,006.9
Note. * $p < .05$

The sequence of HLM software commands used to obtain these results can be found in Appendix E. Beginning with the fixed effect $\gamma_{00,}$ the log-odds of accuracy in the organization domain for a rater who received self-paced training was 1.41 logits when Z_{ij} took a value of zero (i.e., the experts independently assigned the same score in the organization domain). For each unit increase in Z_{ij}, the log-odds decreased by 0.6 logits, which is significantly different than zero logits, as reflected by the estimate for the slope parameter γ_{10}. This finding suggests that for a rater who received self-paced training, accuracy in the organization domain was contingent on the type of essay being scored. A similar but more pronounced trajectory was found for raters who received collaborative training. The estimate for γ_{01} demonstrates that the log-odds of accuracy in the organization domain for a rater who received collaborative training was 1.59 logits when Z_{ij} took a value of zero. This estimate is obtained by summing γ_{00} and $\gamma_{01.}$ For each unit increase in Z_{ij}, the log-odds went down by 0.91 logits. This estimate is obtained by summing γ_{10} and $\gamma_{11.}$ Compared to a rater who received self-paced training, a rater who received

collaborative training was slightly more accurate when Z_{ij} was zero but became *significantly* less accurate as Z_{ij} took higher values. The difference in intercepts was not significant, but the difference in slopes was, at $p = .05$. Therefore, there is sufficient evidence to reject the null hypothesis that raters from both training conditions are equivalently accurate as the essay type changes. That is, there was a training method effect on the strength of association between accuracy in the organization domain and the type of essay being scored. A rater who received self-paced training had greater log-odds of accuracy as Z_{ij} took higher values.

As the estimate of the random effect τ_{00} in Table 8 illustrates, there was not a significant rater effect on accuracy in the organization domain when controlling for both training condition and essay type.

To compare the fit of the models in Equations 5.3 and 7.3, a deviance test was used. The deviance for the models in Equations 5.3 and 7.3 was 9,104.8 and 9,006.9, respectively. This difference is 97.9. It is distributed, approximately, as a chi-square with three degrees of freedom because the model in Equation 7.3 estimated three more parameters. The resulting chi-square is significant at $p < .05$. The model in Equation 5.3 was thus rejected in favor of the model in Equation 7.3. This model, with the slope variable Z_{ij}, predicted the log-odds of accuracy in the organization domain better than the model with no predictors or the model with only the categorical covariate for training condition. This finding provides evidence that some essays are more difficult to score accurately in the organization domain. It is consistent with the general findings of Engelhard (1996).

A model with a random effect for the slope was fitted to the data but rejected because it did not improve model fit. The deviance for this model was 9006.9, and it estimated two more parameters than the model in Equation 7.3. These parameters were a random effect for the slope

and a covariance parameter between the intercepts and slopes. The difference in deviances between this model and the model in Equation 7.3 was 0. It is distributed, approximately, as a chi-square with two degrees of freedom because the model with a random effect for the slope estimated two additional parameters. The resulting chi-square is not significant at $p = .05$. Therefore, the model in Equation 7.3 was retained.

As a final note, Appendix F features a normal quintile-quintile (Q-Q) plot of the level-2 random effect for the intercept from the model in Equation 7.3. In this plot, the distribution of the estimated level-2 random effect for the intercept (x-axis) is plotted against an expected normal distribution (y-axis). Because the points in this plot form approximately a straight line, there is evidence that the distribution of the level-2 random effect for the intercept (τ_{00}) was normal, as it is assumed to be.

The Style Domain

In the style domain, the examinee's essay is rated on how effectively he/she controls language to engage the reader. Emphasis is placed on how carefully the examinee chooses words and crafts phrases and sentences to enhance the persuasiveness of the essay. Again, there is a specific rubric raters used to evaluate each essay in the style domain, and the rating scale for the rubric ranges from 1-5. Higher scores on the rating scale denote more effective control of language. Appendixes C and D show descriptive statistics and the distribution of scores assigned by experts and raters, respectively, in the style domain to the 50 common set essays. The mean raw score in the style domain for both experts and raters was 3, and the standard deviation was approximately 1. The experts and raters assigned approximately the same proportion of 1s, 2s, 3s, 4s, and 5s in the style domain. These distributions suggest that the essays in the common sets

reflected a range of proficiency in style. Ultimately, raters' raw scores in the style domain were compared to the resolved experts' scores and then converted to distance accuracy scores using Equation 3. From these distance accuracy scores, each observation was then coded 1 (accurate) if the distance accuracy score was zero, meaning there was no difference between the resolved expert score and the rater's score. If the distance accuracy score was 1 or 2, the observation was coded 0 (inaccurate). The logistic mixed models in Equations 5.3, 6.2, and 7.3 were then fitted to these binary data. Results from fitting these models are presented below.

The results of fitting the model in Equation 5.3, the unconditional model with log-odds of accuracy in the style domain as the dependent variable, are presented in Table 9.

Table 9

Results of Fitting the Unconditional Model (Style Domain)

Parameter	Estimate	SE	*p*
γ_{00}	1.02	.05	<.001*
τ_{00}	.066	.03	.001*

Note. -2 log L = 9,292.9
Note. * $p <$ .05

The sequence of HLM software commands used to obtain these results can be found in Appendix E. The unconditional model contains no predictors. The purpose of fitting this model was to explore whether raters varied in overall accuracy in the style domain, thereby addressing research question 1 for this domain. Beginning with overall accuracy in style, the fixed effect γ_{00} indicates that, overall, a rater's log-odds of scoring essays accurately in the style domain was 1.02 logits. This estimate differs significantly from zero logits, which, expressed as a probability, denotes a fifty percent probability of accuracy, evident by inserting a value of zero

for η_{ij} in Equation 8 and solving. By contrast, inserting the predicted value of γ_{00} (1.02) for η_{ij} and solving Equation 8 results in an overall probability of .73 that rater *j* will score essays accurately in the style domain. The variance around the overall log-odds of accuracy was .066 logits, as indicated by the estimate for τ_{00}. This estimate was significant, meaning there was a rater effect on the log-odds of accuracy in the style domain. Said differently, when there are no predictors in the model, raters were not equivalent in their estimated log-odds of scoring essays accurately in the style domain.

To address the second research question about differences in rater accuracy as a function of training method, the model in Equation 6.2, with training method as a categorical covariate, was fitted to the accuracy scores in the style domain. Results are found in Table 10.

Table 10

Results of Fitting the Conditional Model from Equation 6.2 (Style Domain)

Parameter	Estimate	SE	*p*
γ_{00} (Self-paced)	0.985	0.09	<.001*
γ_{01}(Collab.)	0.07	0.11	.53
τ_{00}	0.065	0.03	<.001*

Note. -2 log *L* = 9,292.4
Note. * $p <$ .05

The sequence of HLM software commands used to obtain these results can be found in Appendix E. Beginning with the fixed effects, the estimate for γ_{00} takes on a different meaning in this model. It now signifies the overall log-odds of scoring essays accurately in the style domain for a rater who received self-paced frame-of-reference training. The estimate for γ_{00} was 0.985 logits, which differs significantly from 0 logits. By comparison, the fixed effect γ_{01} reflects the

overall log-odds of scoring essays accurately in the style domain for a rater who received collaborative frame-of-reference training. The estimate, 0.992 logits, is obtained by summing γ_{00} and $\gamma_{01.}$ Relative to a rater who received self-paced training, a rater who received collaborative training had slightly greater log-odds of scoring essays accurately in the style domain— 0.07 logits, which was not significant. This difference does not provide sufficient evidence to reject the null hypothesis that both groups of raters are equivalently accurate in the style domain.

As the estimate of the random effect τ_{00} in Table 10 illustrates, there was a significant rater effect on the log-odds of accuracy in the style domain, even after controlling for training condition.

To compare the fit of the models in Equations 5.3 and 6.2, a deviance test was used. As reported in Tables 9 and 10, the deviance for both the unconditional and conditional models was 9,292.9 and 9,292.4, respectively. This difference is 0.5. It is distributed, approximately, as a chi-square with one degree of freedom because the model in Equation 6.2 estimated one more parameter. The resulting chi-square is not significant at p =.05. The conditional model was thus rejected in favor of the unconditional model. However, attempting to fit the conditional model allowed for exploration of research question 2. There was no evidence of a training method effect on the overall log-odds of accuracy in the style domain.

To address the third research question about change in rater accuracy as a function of essay type, the model in Equation 7.3 was fitted to the accuracy scores in the style domain. The model in Equation 7.3 introduces the slope variable Z_{ij}, which allows for the modeling of change in the log-odds of accuracy in the style domain as a function of the type of essay the rater scores. Again, essay type, denoted by Z_{ij}, reflects the maximum absolute difference between the scores that each expert *independently* assigned in the style domain to the common set essay in question.

That is, higher values of Z_{ij} highlight essays for which it was more difficult for experts to determine an accurate score in the style domain. The theoretical range of Z_{ij} in the style domain is 0-4, but the data reveal that the observed range was 0-2. The results of fitting the model in Equation 7.3 are reported in Table 11.

Table 11

Results of Fitting the Conditional Model from Equation 7.3 (Style Domain)

Parameter	Estimate	SE	*p*
γ_{00} (Self-paced)	1.23	0.11	<.001*
γ_{01} (Collab.)	0.09	0.13	.5
γ_{10} (Self-paced)	-0.63	0.1	<.001*
γ_{11} (Collab.)	-0.05	0.18	.79
τ_{00}	0.07	0.03	<.001*

Note. -2 log L = 9,219.4
Note. * $p <$ .05

The sequence of HLM software commands used to obtain these results can be found in Appendix E. Beginning with the fixed effect $\gamma_{00,}$ the log-odds of accuracy in the style domain for a rater who received self-paced training was 1.23 logits when Z_{ij} took a value of zero (i.e., the experts independently assigned the same score in the style domain). For each unit increase in Z_{ij}, the log-odds went down by 0.63 logits, which is significantly different than zero logits, as reflected by the estimate for the slope parameter γ_{10}. This finding suggests that for a rater who received self-paced training, accuracy in the style domain was contingent on the type of essay being scored. A similar trajectory was found for raters who received collaborative training. The log-odds of accuracy in the style domain for a rater who received collaborative training was 1.32

logits when Z_{ij} took a value of zero. This estimate is obtained by summing γ_{00} and $\gamma_{01.}$ For each unit increase in Z_{ij}, the log-odds went down by 0.68 logits. This estimate is obtained by summing γ_{10} and $\gamma_{11.}$ Compared to a rater who received self-paced training, a rater who received collaborative training was slightly more accurate when Z_{ij} was zero but slightly less accurate as Z_{ij} took higher values. As can be seen in Table 11, however, these differences in intercepts and slopes were not significant. There was not sufficient evidence to reject the null hypothesis that raters from both training conditions are equivalently accurate as the essay type changes. That is, there was no training method effect on the strength of association between accuracy in the style domain and the type of essay being scored.

As the estimate of the random effect τ_{00} in Table 11 illustrates, there was a significant rater effect on accuracy in the style domain, even after controlling for both training condition and essay type.

To compare the fit of the models in Equations 5.3 and 7.3, a deviance test was used. The deviance for the models in Equations 5.3 and 7.3 was 9,292.9 and 9,219.4, respectively. This difference is 73.5. It is distributed, approximately, as a chi-square with three degrees of freedom because the model in Equation 7.3 estimated three more parameters. The resulting chi-square is significant at p <.05. The model in Equation 5.3 was thus rejected in favor of the model in Equation 7.3. The model with the slope variable Z_{ij} predicted the log-odds of accuracy in the style domain better than the model with no predictors or the model with only the categorical covariate for training condition. This finding, which provides evidence that some essays are more difficult to score accurately, is consistent with the general findings in Engelhard (1996).

A model with a random effect for the slope was fitted to the data but rejected because it did not improve model fit. The deviance for this model was 9218.2, and it estimated two more

parameters than the model in Equation 7.3. These parameters were a random effect for the slope and a covariance parameter between the intercepts and slopes. The difference in deviances between this model and the model in Equation 7.3 was 1.2. It is distributed, approximately, as a chi-square with two degrees of freedom because the model with a random effect for the slope estimated two additional parameters. The resulting chi-square is not significant at $p = .05$. Therefore, the model in Equation 7.3 was retained. While there was a rater effect on the intercept (i.e., the overall log-odds of accuracy), evident by the significant estimate for τ_{00}, there was no rater effect on the slope. Raters did not vary significantly in the change in log-odds of accuracy for each unit increase in Z_{ij}.

As a final note, Appendix F features a normal quintile-quintile (Q-Q) plot of the level-2 random effect for the intercept from the model in Equation 7.3. In this plot, the distribution of the estimated level-2 random effect for the intercept (x-axis) is plotted against an expected normal distribution (y-axis). Because the points in this plot form approximately a straight line, there is evidence that the distribution of the level-2 random effect for the intercept (τ_{00}) was normal, as it is assumed to be.

The Conventions Domain

In the conventions domain, the examinee's essay is rated on how effectively he/she demonstrates control of standard English. Emphasis is placed on the correctness and complexity of sentence formation, usage, and mechanics. Again, there is a specific rubric raters used to evaluate each essay in the conventions domain, and the rating scale for the rubric ranges from 1-5. Higher scores on the rating scale denote greater control of standard English. Appendixes C and D show descriptive statistics and the distribution of scores assigned by experts and raters,

respectively, in the conventions domain to the 50 common set essays. The mean raw score in the conventions domain for both experts and raters was 3, and the standard deviation was approximately 1. The experts and raters assigned approximately the same proportion of 1s, 2s, 3s, 4s, and 5s in the conventions domain. These distributions suggest that the essays in the common sets reflected a range of proficiency in conventions. Ultimately, raters' raw scores in the conventions domain were compared to the resolved experts' scores and then converted to distance accuracy scores using Equation 3. From these distance accuracy scores, each observation was then coded 1 (accurate) if the distance accuracy score was zero, meaning there was no difference between the resolved expert score and the rater's score. If the distance accuracy score was 1 or 2, the observation was coded 0 (inaccurate). The logistic mixed models in Equations 5.3, 6.2, and 7.3 were then fitted to these binary data. Results from fitting these models are presented below.

The results of fitting the model in Equation 5.3, the unconditional model with log-odds of accuracy in the conventions domain as the dependent variable, are presented in Table 12.

Table 12

Results of Fitting the Unconditional Model (Conventions Domain)

Parameter	Estimate	SE	*p*
γ_{00}	0.99	0.05	<.001*
τ_{00}	0.03	0.02	.04*

Note. -2 log *L* = 9,322.8
Note. * $p < .05$

The sequence of HLM software commands used to obtain these results can be found in Appendix E. The unconditional model contains no predictors. The purpose of fitting this model was to

explore whether raters varied in overall accuracy in the conventions domain, thereby addressing research question 1 for this domain. The fixed effect γ_{00} indicates that, overall, a rater's log-odds of scoring an essay accurately in the conventions domain was 0.99 logits. This estimate differs significantly from zero logits, which, expressed as a probability, denotes a fifty percent probability of accuracy, evident by inserting a value of 0 for η_{ij} in Equation 8 and solving. By contrast, inserting the predicted value of γ_{00} (0.99) for η_{ij} and solving Equation 8 results in an overall probability of .73 that rater *j* will score essays accurately in the conventions domain. The variance around the overall log-odds of accuracy was 0.03 logits, as indicated by the estimate for τ_{00}. This estimate was significant, meaning there was a rater effect on the log-odds of accuracy in the conventions domain. Said differently, when there are no predictors in the model, raters were not equivalent in their estimated log-odds of scoring essays accurately in the conventions domain.

To address the second research question about differences in rater accuracy as a function of training method, the model in Equation 6.2, with training method as a categorical covariate, was fitted to the accuracy scores in the conventions domain. Results are found in Table 13.

Table 13

Results of Fitting the Conditional Model from Equation 6.2 (Conventions Domain)

Parameter	Estimate	SE	*p*
γ_{00} (Self-paced)	0.9	0.06	<.001*
γ_{01}(Collab.)	0.18	0.09	.05*
τ_{00}	0.02	0.03	.08

Note. -2 log *L* = 9,318.8
Note. * $p <$ .05

The sequence of HLM software commands used to obtain these results can be found in Appendix E. Beginning with the fixed effects, the estimate for γ_{00} takes on a different meaning in this model. It now signifies the overall log-odds of scoring essays accurately in the conventions domain for a rater who received self-paced frame-of-reference training. The estimate for γ_{00} was 0.9 logits, which differs significantly from 0 logits. By comparison, the fixed effect γ_{01} reflects the overall log-odds of scoring essays accurately in the conventions domain for a rater who received collaborative frame-of-reference training. This estimate, 1.08 logits, is obtained by summing γ_{00} and γ_{01}. Relative to a rater who received self-paced training, a rater who received collaborative training had significantly greater log-odds of scoring essays accurately in the conventions domain — 0.18 logits. This difference provides sufficient evidence to reject the null hypothesis that raters from both training conditions are equivalently accurate in the conventions domain.

As the estimate of the random effect τ_{00} in Table 13 illustrates, the rater effect on the log-odds of accuracy in the conventions domain was not significant when controlling for training condition.

To compare the fit of the models in Equations 5.3 and 6.2, a deviance test was used. As reported in Tables 12 and 13, the deviance for both the unconditional and conditional models was 9,322.8 and 9,318.8, respectively. This difference is 4. It is distributed, approximately, as a chi-square with one degree of freedom because the model in Equation 6.2 estimated one more parameter. The resulting chi-square is significant at p =.05. This result means the difference in deviances is significant. The unconditional model was thus rejected in favor of the conditional model. There was evidence of a training method effect on the overall log-odds of accuracy in the

conventions domain. Raters who received collaborative training had higher log-odds of accuracy when scoring essays in conventions.

To address the third research question about change in rater accuracy as a function of essay type, the model in Equation 7.3 was fitted to the accuracy scores in the conventions domain. The model in Equation 7.3 introduces the slope variable Z_{ij}, which allows for the modeling of change in the log-odds of accuracy in the conventions domain as a function of the type of essay the rater scores. As before, essay type, denoted by Z_{ij}, reflects the maximum absolute difference between the scores that each expert *independently* assigned in the conventions domain to the common set essay in question. That is, higher values of Z_{ij} highlight essays for which it was more difficult for experts to determine an accurate score in the conventions domain. The theoretical range of Z_{ij} in the conventions domain is 0-4, but the data reveal that the observed range was 0-2. The results of fitting the model in Equation 7.3 are reported in Table 14.

Table 14

Results of Fitting the Conditional Model from Equation 7.3 (Conventions Domain)

Parameter	Estimate	SE	*p*
γ_{00} (Self-paced)	0.99	0.09	<.001*
γ_{01} (Collab.)	0.17	0.13	.2
γ_{10} (Self-paced)	-0.16	0.09	.07
γ_{11} (Collab.)	0.02	0.18	.92
τ_{00}	0.02	0.02	.08

Note. -2 log L = 9,312.8
Note. *p < .05

The sequence of HLM software commands used to obtain these results can be found in Appendix E. Beginning with the fixed effect $\gamma_{00,}$ the log-odds of accuracy in the conventions domain for a rater who received self-paced training was 0.99 logits when Z_{ij} took a value of zero (i.e., the experts independently assigned the same score in the conventions domain). For each unit increase in Z_{ij}, the log-odds went down by 0.16 logits, which is not significantly different than zero logits, as reflected by the estimate for the slope parameter γ_{10}. This finding suggests that for a rater who received self-paced training, accuracy in the conventions domain was not contingent on the type of essay being scored. A similar trajectory was found for raters who received collaborative training. The estimate for γ_{01} indicates that the log-odds of accuracy in the conventions domain for a rater who received collaborative training was1.16 logits when Z_{ij} took a value of zero. This estimate is obtained by summing γ_{00} and $\gamma_{01.}$ For each unit increase in Z_{ij}, the log-odds went down by 0.14 logits. This estimate is obtained by summing γ_{10} and $\gamma_{11.}$ Compared to a rater who received self-paced training, a rater who received collaborative training was slightly more accurate when Z_{ij} was zero and slightly more accurate as Z_{ij} took higher values. These differences in intercepts and slopes were not significant, as reflected in Table 14. Therefore, there was not sufficient evidence to reject the null hypothesis that raters from both training conditions are equivalently accurate as the essay type changes. That is, there was no training method effect on the strength of association between the log-odds of accuracy in the conventions domain and the type of essay being scored.

As the estimate of the random effect τ_{00} in Table 14 illustrates, there was not a significant rater effect on accuracy in the conventions domain when controlling for both training condition and essay type.

To compare the fit of the models in Equations 6.2 and 7.3, a deviance test was used. The deviance for the models in Equations 6.2 and 7.3 was 9,318.8 and 9,312.8, respectively. This difference is 6. It is distributed, approximately, as a chi-square with two degrees of freedom because the model in Equation 7.3 estimated two more parameters. The resulting chi-square is significant at $p < .05$. The model in Equation 6.2 was thus rejected in favor of the model in Equation 7.3. The model with the slope variable Z_{ij} predicted the log-odds of accuracy in the conventions domain better than the model with no predictors or the model with only the categorical covariate for training condition. Because the model in Equation 7.3 fit the data better, the parameter estimates from this model are preferred to those from the model in Equation 6.2. This point bears emphasis because the difference in intercepts was significant in the model in Equation 6.2 but not in Equation 7.3. Therefore, it is concluded that there was not a fixed training method effect on the intercept. The same is true with respect to the slope.

A model with a random effect for the slope was also fitted to the data but rejected because it did not improve model fit. The deviance for this model was 9312.8, and it estimated two more parameters than the model in Equation 7.3. These parameters were a random effect for the slope and a covariance parameter between the intercepts and slopes. The difference in deviances between this model and the model in Equation 7.3 was 0. It is distributed, approximately, as a chi-square with two degrees of freedom because the model with a random effect for the slope estimated two additional parameters. The resulting chi-square is not significant at $p = .05$. Therefore, the model in Equation 7.3 was retained.

As a final note, Appendix F features a normal quintile-quintile (Q-Q) plot of the level-2 random effect for the intercept from the model in Equation 7.3. In this plot, the distribution of the estimated level-2 random effect for the intercept (x-axis) is plotted against an expected

normal distribution (y-axis). Because the points in this plot form approximately a straight line, there is evidence that the distribution of the level-2 random effect for the intercept (τ_{00}) was normal, as it is assumed to be.

CHAPTER 5

DISCUSSION

A large body of research has shown that raters differ in both accuracy and degree of errors (Engelhard, 1996, 2002; Gyagenda & Engelhard, 2009; Leckie & Baird, 2011; Weigle, 1998). Studies have investigated whether these differences are a function of factors such as professional experience (Shohamy, Gordon, & Kraemer, 1992), amount of rating experience (Leckie & Baird, 2011), differences in cognitive processing (Wolfe, 2005), and training (Weigle, 1998). With respect to the training factor, this dissertation built on prior research by using an experimental design in which raters were randomly assigned to one of the two most commonly used training methods in large-scale writing assessment: self-paced or collaborative frame-of-reference training (Johnson, Penny, & Gordon, 2009). The relative effect of these training conditions on rater accuracy was of particular interest. Accuracy at the domain level was used as the dependent variable because it provided direct comparison of a rater's and expert's scores; therefore, it provided direct evidence of a rater's proficiency (Engelhard, 2013). To control for the possibility that some essays might be more difficult to score accurately than others, the essay type was collected and added to the conditional logistic mixed model in Equation 7.3 to predict the slope. Therefore, in addition to exploring the relative effects of two training methods on rater accuracy, this dissertation also explored the influence of essay type on rater accuracy.

In general, no training method effect was found on either the overall log-odds of accuracy or on the strength of association between the log-odds of accuracy and essay type. That is, raters

who received self-paced training or collaborative training did not differ, in general, in their predicted intercept and slope. Raters who received collaborative training tended to be slightly more accurate when the essay type Z_{ij} took a value of zero. By contrast, raters who received self-paced training tended to be slightly more accurate as essays became somewhat more difficult to score, that is, as Z_{ij} took higher values. The vast majority of these differences were not significant. There was an exception, however. In the organization domain, there was a significant interaction effect between training method and essay type. As Z_{ij}took higher values, raters who received collaborative training were significantly less accurate than raters who received self-paced training.

The mostly non-significant differences in accuracy between the two training methods provide evidence that the self-paced method helps raters score as accurately as the collaborative method. This finding has specific implications for large-scale writing assessment programs, given the relative time-intensiveness, cost, and delivery capabilities of the self-paced method. Raters in this study that were assigned to the self-paced method were allotted as much time to complete training as raters assigned to the collaborative method, but few took all of the allotted time. In fact, the majority of the raters assigned to the self-paced method completed training in six hours, whereas raters assigned to the collaborative method completed training in 9 hours. Because the self-paced method generally requires less time, it is more economical. Further, it is more versatile. While it is possible to conduct collaborative training online, the logistics of online collaborative training are more cumbersome than online self-paced training. For example, it would be burdensome to arrange a videoconference with a group of raters to hold a dialogue about a set of training or practice benchmarks. It would be easier for a rater and a trainer to

discuss one-on-one, using online conferencing technologies, which is possible when a self-paced training model is delivered online.

The raters who participated in this study were not absolute novices. The reason for this is that Georgia Supplemental Writing Assessment protocol requires all raters to have had experience on at least one scoring project. Subsequent research on the effects of rater training could usefully focus on differences in response to training of novice and experienced raters.

Although most of the contrasts between the two training methods were not significant, most of the main effects were. In all analyses, the intercepts (i.e., the overall log-odds of accuracy) were significant. That is, in all domains, the overall log-odds of accuracy were significantly different than zero, which would denote a .5 probability of accuracy. Raters from both training conditions had overall log-odds of accuracy ranging from 0.9 in the ideas domain to 1.17 in the organization domain. Expressed as probabilities, using Equation 8, these log-odds translate to approximately 0.7 in each domain. Overall, raters appeared to have scored accurately in each domain.

With respect to the main effect for the slope (i.e., the change in the log-odds of accuracy for each unit increase in Z_{ij}), the analysis in each domain revealed significant estimates, except in the conventions domain. As Z_{ij}took on higher values, the log-odds of accuracy decreased. Said differently, as it became more difficult for experts to determine an accurate score on an essay, it became more difficult for raters to be accurate. This dissertation is one of only a few studies to have explored the effect of essay type on accuracy. The research in this dissertation produced results similar to, and also built on, the findings in Engelhard (1996) by offering an explanation for why some essays were more difficult to score accurately: as the experts had a more difficult time agreeing exactly on essays, raters became less accurate. Further research,

perhaps including qualitative analyses of rater responses, would be useful for determining whether essays for which Z_{ij} took higher values had similar characteristics and to probe why some experts disagreed. It would be useful to include examples of such essays for rater training to determine whether this might improve training and, therefore, scoring accuracy. As reflected in Appendix A, only one-quarter to one-third of the training benchmarks included in this dissertation were essays where Z_{ij} took a value of 1. All others took a value of zero.

This issue has specific implications for advances in automated scoring, which has received considerable attention as the number of constructed response items on large-scale assessments has increased (Williamson, Xi, & Breyer, 2012). Nearly all automated engines are trained in much the same way as human raters –through training benchmarks, typically selected and resolved by experts (Williamson, Xi, & Breyer, 2012). Like humans, automated systems are likely to be less accurate on essays where experts have greater difficulty determining an accurate score, unless categories of such essays are identified and appropriately sampled in training. In a review of automated scoring engines, Dikli (2006) reported that automated engines require anywhere from 100 to 1,000 training benchmarks, with most engines requiring 200-300. Williamson, Xi, and Breyer's (2012) more current review reports similar numbers. That said, another question for further research involves determining what proportion of these benchmarks should be "clear," defined by unanimous agreement among independent expert scores, versus more "ambiguous," defined by less agreement among independent expert scores, prior to resolution. Of course, the same question applies to training human raters. Further research along these lines could involve training raters on differing proportions of clear and more ambiguous training benchmark papers, as determined by experts' independent scores prior to resolution. After treatment, these groups of raters could score a common set of essays made up of essays

that present differing degrees of ambiguity with respect to the scoring decision. It may be true that more rigorous training on ambiguous benchmarks leaves raters better prepared to score such responses accurately.

To conclude, if the number of constructed-response items on large-scale assessments continues to rise, rater accuracy will be increasingly important to demonstrate. As defined in this dissertation, accuracy denotes the degree to which a rater's scores match an expert's (Engelhard, 1996; Sulsky & Balzer, 1988). That is, accuracy involves expertise. Sulsky and Balzer (1988) note that expertise must be established with evidence, meaning the criteria for what constitutes expertise should not be taken for granted. To this end, the experts in this dissertation had considerable experience applying the Georgia Supplemental Writing Assessment rubrics to essays. They also helped design the assessment, its rubrics, and its training protocols. Therefore, the experts possess understanding of how the examinee's essays were intended to be scored. Consequently, comparing a rater's scores with an expert's provides a direct link between assessment development, how the responses were intended to be scored, and how they were in fact scored. Such evidence can be used to demonstrate scoring fidelity within the broader validity argument of the writing assessment (Kane, Crooks, & Cohen, 1999). In fact, this evidence is arguably stronger than indices of inter-rater agreement and inter-rater reliability. In other words, it is one thing for raters to agree or rank order essays similarly; it is another for raters and experts to do the same. This is particularly true given the abundance of research showing that raters differ in accuracy, severity, and other measures of rater proficiency (Engelhard, 1996, 2002; Gyagenda & Engelhard, 2009; Leckie & Baird, 2011; Weigle, 1998). Experts can provide a more stable frame of reference, both during training and scoring. Certainly, there are opportunities for further research on rater training and monitoring techniques

involving experts. Such avenues, some described in this dissertation, hold promise for addressing the perennial challenge of crafting plausible validity arguments for direct writing assessments and other tests involving raters.

REFERENCES

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, *63*, 301-308.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, *6*, 205-212.

Bernardin, H. J. & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology*, *65*, 60-66.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, *64*, 410-421.

Bruning, R., Schraw, G., & Norby, M. (2011). *Cognitive psychology and instruction*. Boston, MA: Pearson.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cronbach, L.J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin*, *52*, 177-193.

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

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*. Retrieved from http://www.jtla.org.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*, 171-191.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*, 56-70.

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Erlbaum.

Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.

Fitzmaurice, G.M., Laird, N.M., & Ware, J.H. (2011). *Applied longitudinal analysis* (2nd ed.). Hoboken, NJ: Wiley.

Gyagenda, I. & Engelhard, G., Jr. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement, 10*, 1-22.

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: Guilford.

Kane, M.T., Crooks, T., & Cohen, A.S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5-17.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26-43.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology*, *60*, 550-555.

Lane, S., & Stone, C. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 388-431). Portsmouth, CT: American Council on Education and Praeger.

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*, 399–418.

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

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, *69*, 147-156.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Murphy, K. R., & Cleveland, J. N. (1995*). Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69,* 581-588.

Raczynski, K. (2012, April). *Essay characteristics associated with examinee misfit: Implications for rater training and monitoring in a large-scale writing assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.

Raczynski, K., Cohen, A. S., & Lu, Z. (2013, August). *Modeling the effects of training and trainers on rater accuracy*. Poster presented at the annual meeting of the American Psychological Association, Honolulu, HA.

Raudenbush, S. W., Yang, M. L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via higher-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9,* 141-157.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA*:* SAGE.

Raudenbush S. W., Bryk A. S., Cheong Y. F., Congdon R. T., & du Toit M. (2011). HLM 7: Hierarchical linear and nonlinear modeling [Software]. Chicago, IL*:* Scientific Software International.

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., & Hill, C. (2011). Optimal Design Plus Empirical Evidence (Version 3.0) [Software]. Available from http://www.wtgrantfoundation.org/resources/consultation-service-and-optimal-design.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88,* 413-428.

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

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science, Vol. 3* (pp. 1570–1573). Chicester, UK: Wiley.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73,* 497-506

Templin, J. (2012). Multilevel models for clustered data workshop: Summer 2012 (UGA) [PDF document]. Retrieved from http://jonathantemplin.com/multilevel-models-clustered-data-workshop-summer-2012-uga/

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15,* 263-287.

Williamson, D. M., Xi, X., & Breyer, J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31,* 2-13.

Woehr, D., & Huffcutt, A. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*, 189-205.

Wolfe, E. W. (2005). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment*, *2*, 37-56.

Wolfe, E. W., & McVay, A. (2010). "Rater effects as a function of rater training context" (White paper). Pearson Assessments. Retrieved 7/12/13 from https://psychcorp.pearsonassessments.com/NR/rdonlyres/6435A0AF-0C12-46F7-812E-908CBB7ADDFF/0/RaterEffects_101510.p

APPENDIX A:
RESOLVED SCORES AND INDEPENDENT SCORES FOR ALL TRAINING BENCHMARKS AND COMMON SET ESSAYS (BY DOMAIN)

Training Benchmarks (Ideas Domain)

Benchmark	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	1	1	1	1	1	0
2	1	1	1	1	1	0
3	1	1	2	1	1	1
4	2-	1	2	2	2	1
5	2	2	2	2	2	0
6	2	2	2	2	2	0
7	2+	2	2	3	2	1
8	3-	3	3	2	3	1
9	3	3	3	3	3	0
10	3	3	3	3	3	0
11	3+	3	3	3	3	0
12	4-	3	4	4	4	1
13	4	4	4	4	4	0
14	4	4	4	4	4	0
15	4+	4	4	4	5	1
16	5-	5	5	5	5	0
17	5	5	5	5	5	0
18	5	5	5	5	5	0

Training Benchmarks (Organization Domain)

Benchmark	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	1	1	1	1	1	0
2	1	1	1	1	1	0
3	1	1	1	1	1	0
4	2-	2	2	2	2	0
5	2	2	2	2	2	0
6	2	2	2	2	2	0
7	2+	2	3	2	2	1
8	3-	2	3	3	3	1
9	3	3	3	3	3	0
10	3	3	3	3	3	0
11	3+	3	3	3	3	0
12	4-	4	4	3	4	1
13	4	4	4	4	4	0
14	4	4	5	4	4	1
15	4+	4	5	4	4	1
16	5-	5	5	4	5	1
17	5	5	5	5	5	0
18	5	5	5	5	5	0

Training Benchmarks (Style Domain)

Benchmark	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	1	1	1	1	1	0
2	2-	2	1	2	2	1
3	2	2	2	2	2	0
4	2	2	2	2	2	0
5	2	2	2	2	2	0
6	3-	3	3	3	3	0
7	3	3	3	3	3	0
8	3	3	3	3	3	0
9	3	3	3	3	3	0
10	3+	3	3	3	3	0
11	4-	4	3	4	4	1
12	4	4	4	4	4	0
13	4	4	4	4	4	0
14	4	4	4	4	4	0
15	4+	4	5	4	4	1
16	5-	5	4	5	5	1
17	5	5	5	5	5	0
18	5	5	5	5	5	0

Training Benchmarks (Conventions Domain)

Benchmark	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	1	1	1	1	1	0
2	1	1	1	1	1	0
3	1	1	2	1	1	1
4	2-	1	2	2	2	1
5	2	2	2	2	2	0
6	2	2	2	2	2	0
7	2+	2	2	3	2	1
8	3-	3-	3	2	3	1
9	3	3	3	3	3	0
10	3	3	3	3	3	0
11	3+	3	3	3	3	0
12	4-	3	4	4	4	1
13	4	4	4	4	4	0
14	4	4	4	4	4	0
15	4+	4	4	4	5	1
16	5-	5-	5	5	5	0
17	5	5	5	5	5	0
18	5	5	5	5	5	0

Practice Benchmarks (Ideas Domain)

Benchmark	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	3	3	3	3	3	0
2	2	2	2+	2	2	0
3	4	4	4	4	4	0
4	3-	3	3-	2	3-	1
5	2	2	2	2	2	0
6	5	5	5	5	5	0
7	3	3+	3+	3	3	0
8	3	3	3	3	3	0
9	4+	4+	4+	3+	4	1
10	2+	3-	2+	2	2	1

Practice Benchmarks (Organization Domain)

Benchmark	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	3	3	3	3	3	0
2	3-	2	3-	3	3-	1
3	4	4+	4	4	4	0
4	3-	3-	3	2	3	1
5	2	2	2-	2	2	0
6	5	5	5	5	5	0
7	3	3	3+	3	3	0
8	3	3	3	3	3	0
9	4+	4+	4+	4	4	0
10	2+	3-	2+	2	2	1

Practice Benchmarks (Style Domain)

Benchmark	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	4-	4	4	3	4-	1
2	2	2	2	2+	2	0
3	4+	4	4	4	5	1
4	3-	3-	2+	3	3	1
5	2	2-	2-	2	2	0
6	5	5	5	5	5	0
7	2	2	3	2	2	1
8	3	3-	3	3	3	0
9	4	4+	4+	4	4	0
10	3	3	3	3	3	0

Practice Benchmarks (Conventions Domain)

Benchmark	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	4	4	4	4	4-	0
2	2+	2	2	3	2	1
3	5	5	4	5	5-	1
4	3	3	3	3	3	0
5	1	1	1	2	1	1
6	5	5	5	5	5	0
7	2	2	2	2	2-	0
8	3	3	3	3	3	0
9	4	4	4	4	4	0
10	3	3	3	3	3	0

Common Set 1 (Ideas Domain)

Essay	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	3	3	3	3	4	1
2	3	3	3	3	3	0
3	4	4	4	4	4	0
4	3	3	3	3	3	0
5	3	3	3	2	3	1
6	3	3	3	3	3	0
7	1	1	1	1	1	0
8	3	3	2	3	3	1
9	2	2	2	2	2	0
10	3	3	3	3	3	0
11	4	4	4	5	4	1
12	2	2	2	3	2	1
13	2	2	2	2	2	0
14	5	5	5	5	4	1
15	3	3	4	3	3	1
16	3	3	4	3	3	1
17	2	2	2	2	1	1
18	3	3	3	3	3	0
19	3	3	4	3	3	1
20	3	3	3	3	2	1
21	4	4	4	4	4	0
22	3	2	3	3	3	1
23	2	2	3	2	2	1
24	2	2	2	2	2	0
25	3	3	3	3	3	0

Common Set 1 (Organization Domain)

Essay	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	4	3	4	4	4	1
2	3	3	3	3	3	0
3	4	4	3	4	4	1
4	3	3	3	3	3	0
5	3	3	3	3	3	0
6	3	3	3	3	3	0
7	1	1	1	1	1	0
8	3	3	3	3	3	0
9	2	2	2	2	2	0
10	3	3	3	3	3	0
11	4	4	4	4	4	0
12	2	2	2	2	3	1
13	2	2	2	2	2	0
14	5	5	5	5	5	0
15	3	2	3	3	3	1
16	3	3	4	3	3	1
17	2	2	2	2	1	1
18	3	3	3	3	3	0
19	4	4	4	4	4	0
20	3	3	3	3	3	0
21	4	4	4	4	4	0
22	3	2	3	3	3	1
23	2	2	3	2	2	1
24	2	2	2	2	2	0
25	3	4	3	3	3	1

Common Set 1 (Style Domain)

Essay	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	4	3	4	5	4	2
2	3	3	3	3	3	0
3	3	3	3	3	3	0
4	3	3	3	3	3	0
5	3	3	3	3	3	0
6	3	3	3	3	3	0
7	1	1	1	1	1	0
8	3	3	3	3	3	0
9	2	2	2	2	2	0
10	3	3	3	3	3	0
11	4	4	4	4	4	0
12	2	2	2	2	2	0
13	2	2	2	2	2	0
14	5	5	5	5	5	0
15	3	3	3	3	3	0
16	3	3	3	3	3	0
17	1	1	1	1	1	0
18	3	3	3	3	3	0
19	4	4	4	4	4	0
20	3	3	3	3	2	1
21	4	4	4	3	4	1
22	3	2	3	3	3	1
23	3	2	3	3	3	1
24	3	3	3	2	3	1
25	3	3	3	3	3	0

Common Set 1 (Conventions Domain)

Essay	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
1	4	3	4	5	4	2
2	2	2	2	2	2	0
3	3	2	3	4	3	2
4	2	2	2	2	2	0
5	3	3	3	3	3	0
6	4	4	4	4	3	1
7	1	1	1	1	1	0
8	3	3	3	3	3	0
9	2	1	2	2	2	1
10	3	4	3	3	3	1
11	4	4	4	4	4	0
12	2	2	2	2	2	0
13	2	2	2	2	1	1
14	5	5	5	5	5	0
15	3	2	3	3	3	1
16	3	2	3	4	3	2
17	1	1	1	1	1	0
18	3	3	3	3	3	0
19	4	4	4	4	4	0
20	2	2	2	3	2	1
21	4	4	4	4	4	0
22	3	2	3	3	3	1
23	3	2	3	3	3	1
24	2	2	2	3	2	1
25	3	3	3	3	3	0

Common Set 2 (Ideas Domain)

Essay	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
26	3	3	3	3	3	0
27	4	5	4	4	4	1
28	4	4	4	4	5	1
29	2	2	2	2	2	0
30	1	1	1	2	1	1
31	4	4	4	4	4	0
32	3	3	3	3	3	0
33	3	3	3	3	3	0
34	2	2	2	1	2	1
35	3	3	3	3	3	0
36	1	1	2	1	1	1
37	4	4	4	4	4	0
38	3	4	3	3	3	1
39	4	4	3	4	4	1
40	3	3	3	3	3	0
41	3	3	4	3	3	1
42	2	2	2	2	2	0
43	4	5	4	4	4	1
44	2	2	2	2	3	1
45	3	3	3	3	3	0
46	5	5	5	5	5	0
47	4	4	5	4	4	1
48	3	3	3	3	3	0
49	3	3	3	3	3	0
50	3	3	3	3	3	0

Common Set 2 (Organization Domain)

Essay	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
26	3	3	3	3	3	0
27	4	5	4	3	4	2
28	5	5	5	4	5	1
29	2	2	2	2	2	0
30	1	1	1	1	1	0
31	4	4	4	4	4	0
32	3	3	3	3	3	0
33	3	3	3	3	3	0
34	2	2	2	2	2	0
35	3	3	3	3	3	0
36	1	1	2	1	1	1
37	4	4	5	4	4	1
38	3	4	3	3	3	1
39	4	4	3	4	4	1
40	3	3	3	3	3	0
41	4	4	4	4	4	0
42	3	3	3	3	2	1
43	4	5	4	4	4	1
44	2	2	2	2	2	0
45	3	3	4	3	3	1
46	5	5	5	5	5	0
47	4	4	5	4	4	1
48	3	3	3	3	3	0
49	3	3	3	3	3	0
50	3	3	3	3	3	0

Common Set 2 (Style Domain)

Essay	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
26	3	3	4	3	3	1
27	4	5	4	4	4	1
28	5	4	5	5	5	1
29	2	2	2	2	2	0
30	1	1	1	2	1	1
31	4	4	4	4	5	1
32	3	3	3	3	3	0
33	3	3	3	3	3	0
34	2	2	2	2	2	0
35	3	3	3	3	3	0
36	1	1	1	1	1	0
37	4	4	4	4	4	0
38	3	4	3	3	3	1
39	4	4	4	4	4	0
40	3	3	3	3	3	0
41	4	4	4	4	4	0
42	3	3	2	3	3	1
43	4	5	4	4	4	1
44	2	2	2	2	2	0
45	3	3	3	4	3	1
46	5	5	5	5	5	0
47	5	4	5	5	5	1
48	3	3	3	3	3	0
49	3	3	3	3	3	0
50	3	3	2	3	3	1

Common Set 2 (Conventions Domain)

Essay	Resolved Score	Expert 1	Expert 2	Expert 3	Expert 4	Type
26	4	4	4	4	4	0
27	4	5	4	4	4	1
28	5	5	5	5	5	0
29	2	2	2	3	1	2
30	1	1	1	1	1	0
31	4	4	4	4	5	1
32	3	3	3	3	3	0
33	3	3	3	4	3	1
34	2	2	2	2	2	0
35	2	2	2	3	2	1
36	1	1	1	1	1	0
37	4	4	4	4	4	0
38	3	4	3	3	3	1
39	4	4	3	4	4	1
40	3	3	3	3	3	0
41	4	4	4	4	4	0
42	3	3	2	3	3	1
43	5	5	5	5	4	1
44	2	2	2	2	3	1
45	3	3	3	4	3	1
46	5	5	5	5	5	0
47	5	5	5	5	5	0
48	3	3	3	3	3	0
49	3	3	3	3	3	0
50	3	3	3	3	3	0

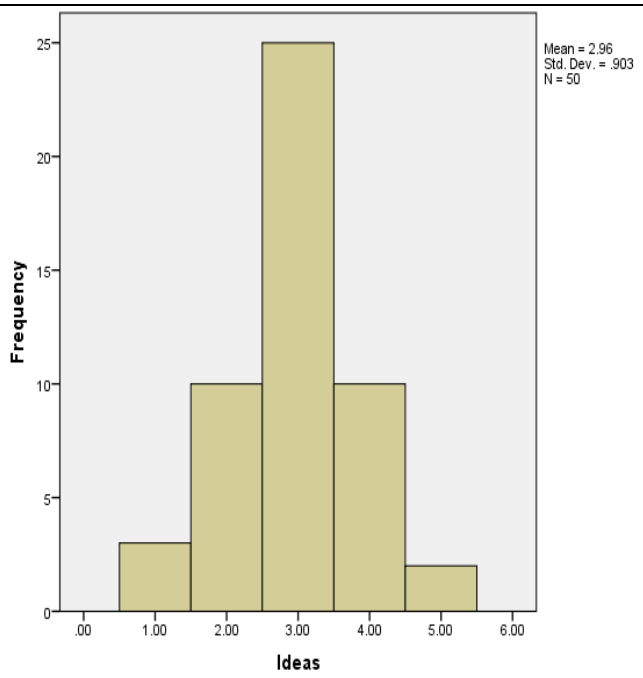
APPENDIX B:
NOTES TABLE THAT RATERS USED DURING TRAINING

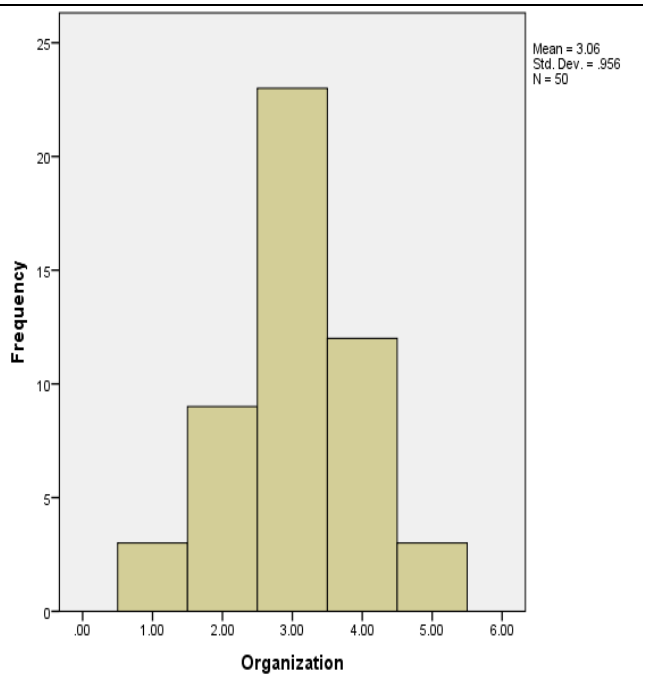
Domain: _____

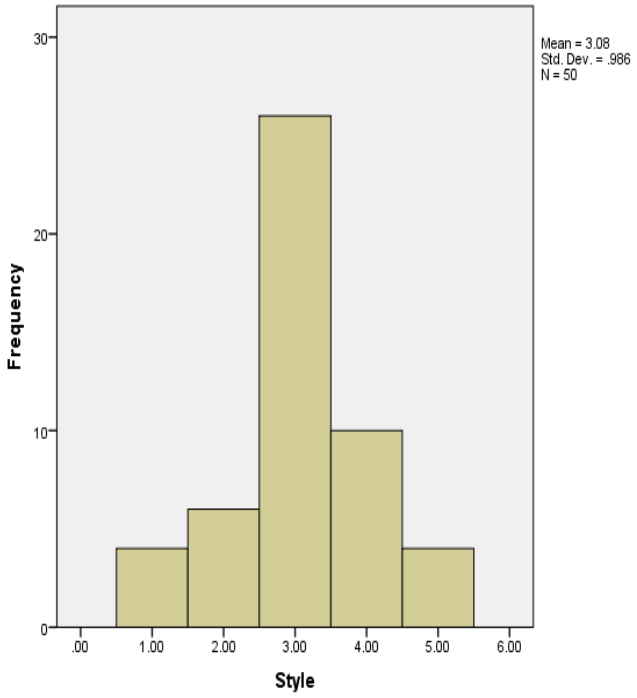
As you work your way through the benchmarks, please jot down some brief notes in the table below. Your notes will help keep the group discussions/conferences focused.

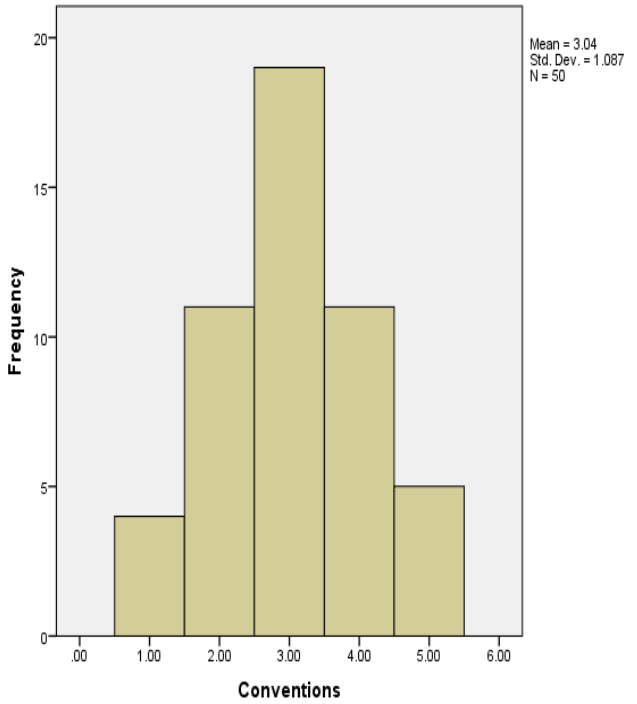
Score line	Key difference	Any benchmarks where this distinction was difficult to make? How so?
1 / 2		
2 / 3		
3 / 4		
4 / 5		

APPENDIX C:
DESCRIPTIVE STATISTICS AND HISTOGRAMS FOR RESOLVED EXPERTS' SCORES ON COMMON SET ESSAYS (BY DOMAIN)

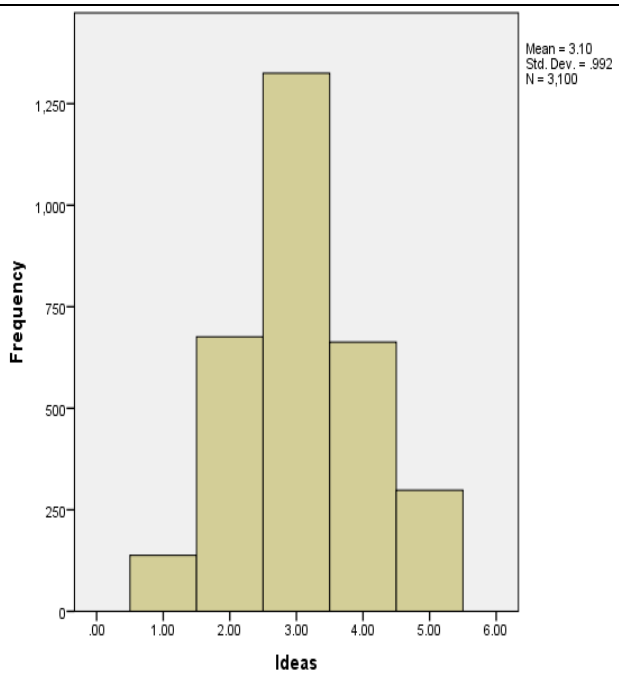


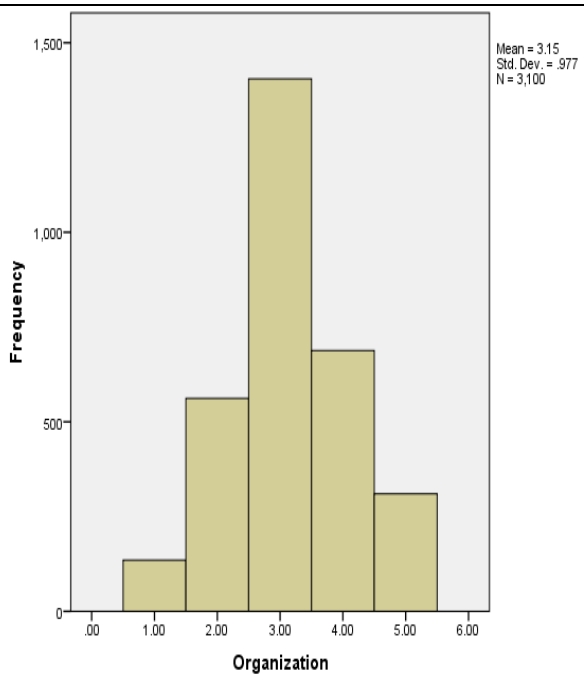


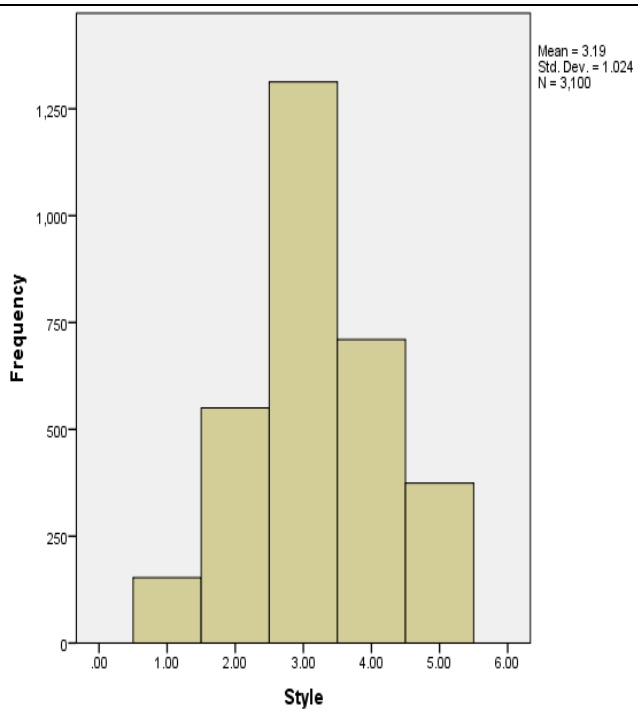


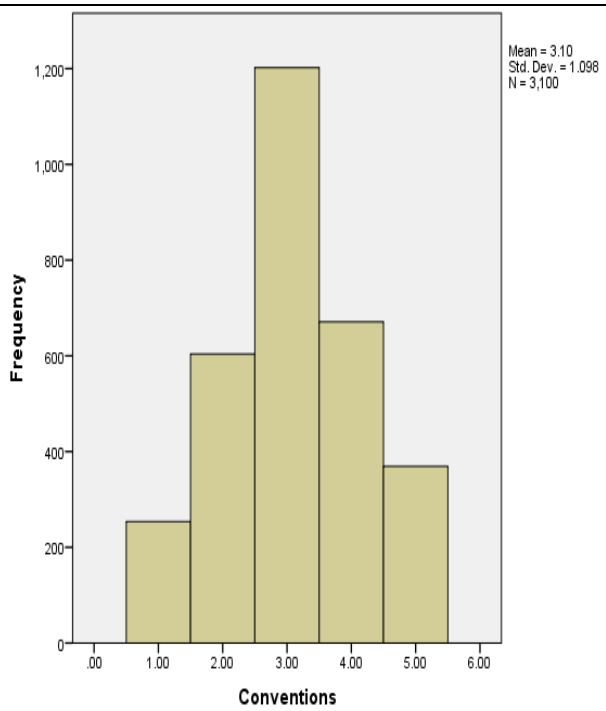


APPENDIX D:
DESCRPTIVE STATISTICS AND HISTOGRAMS FOR RATERS' SCORES ON COMMON SET ESSAYS (BY DOMAIN)

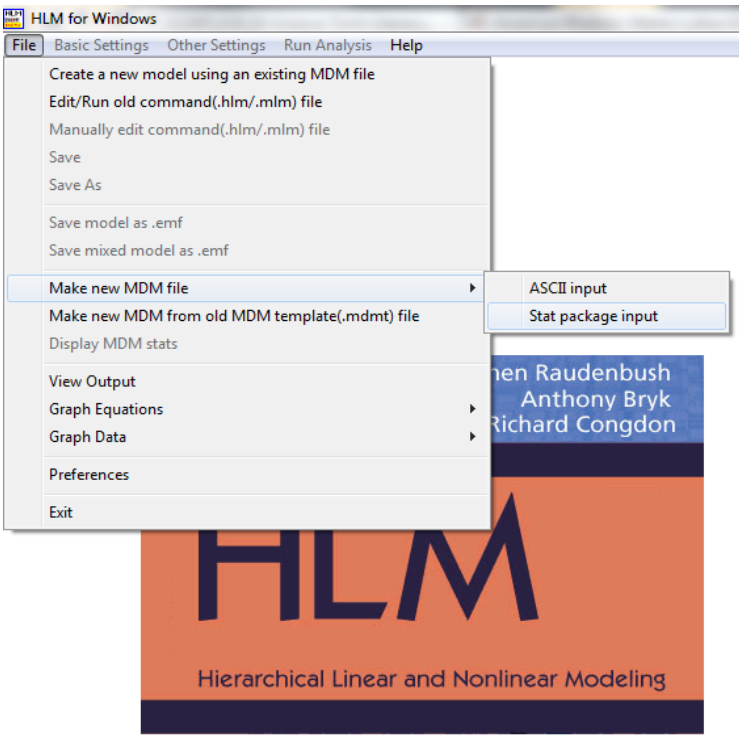


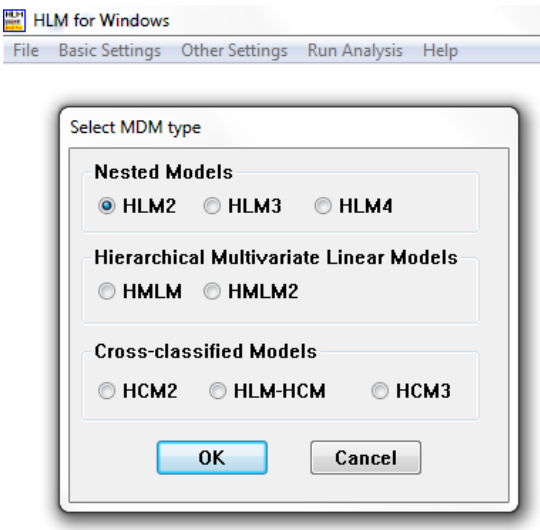




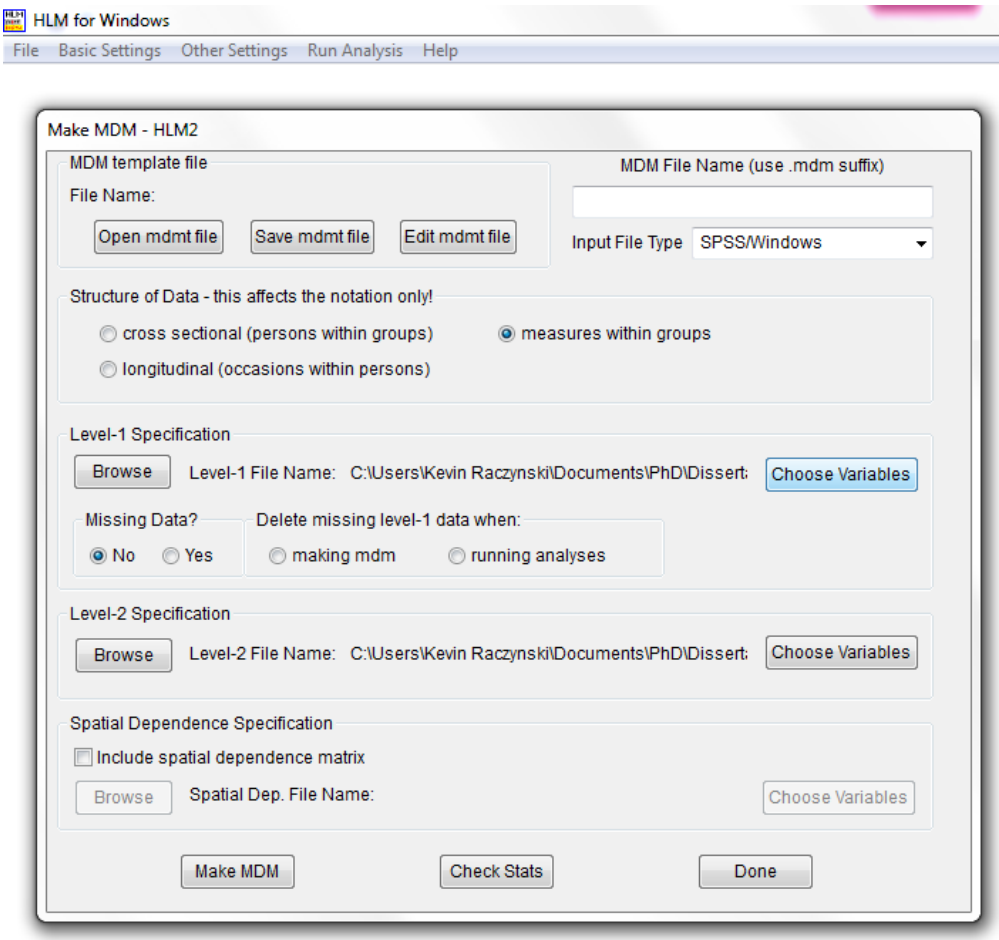


APPENDIX E:
HLM SOFTWARE COMMANDS (ALL VARIABLES SELECTED, AS IN EQUATION 7.3; FOR THE REDUCED MODELS IN 5.3 AND 6.2, NOT ALL VARIABLES WOULD BE SELECTED)

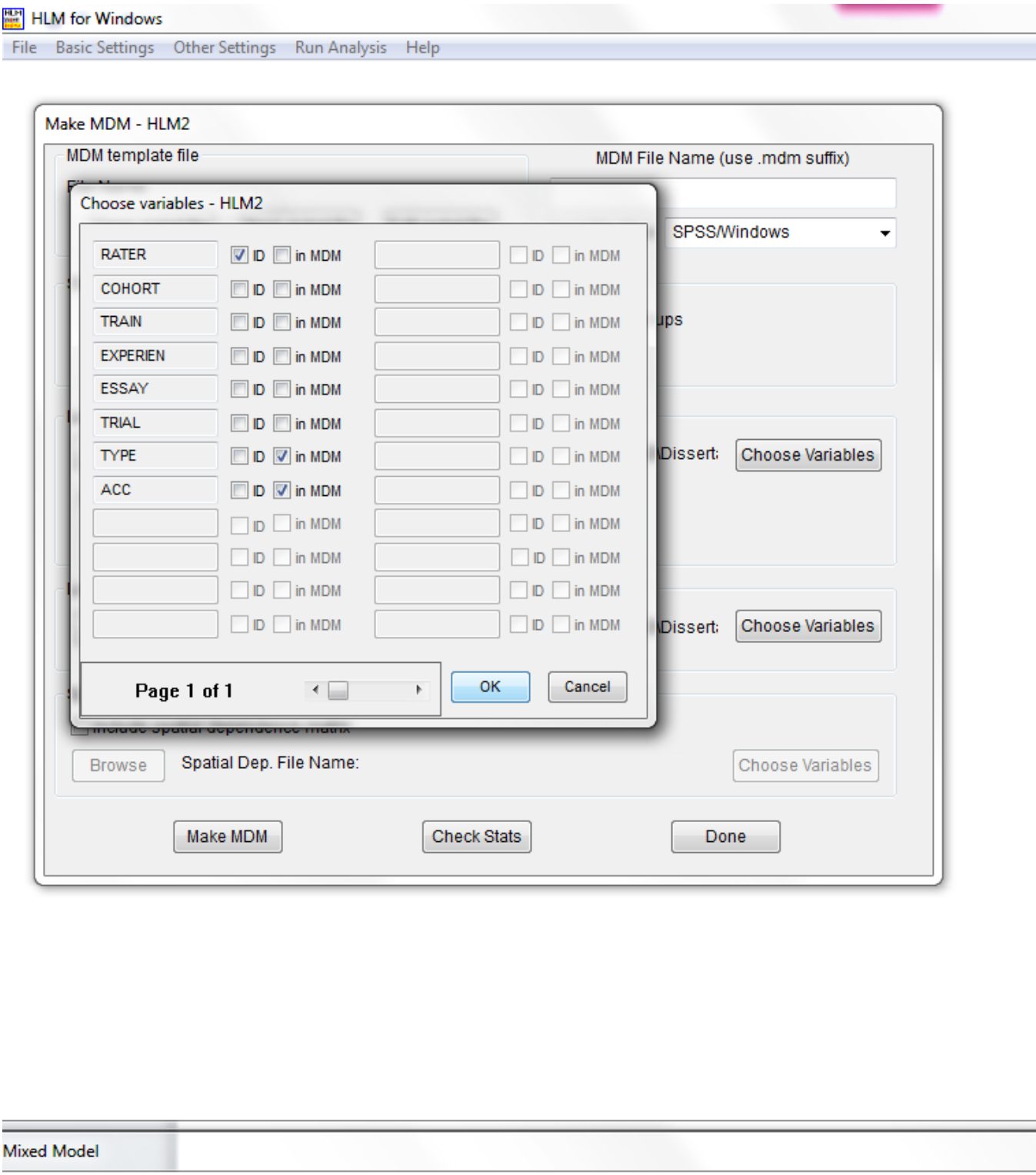


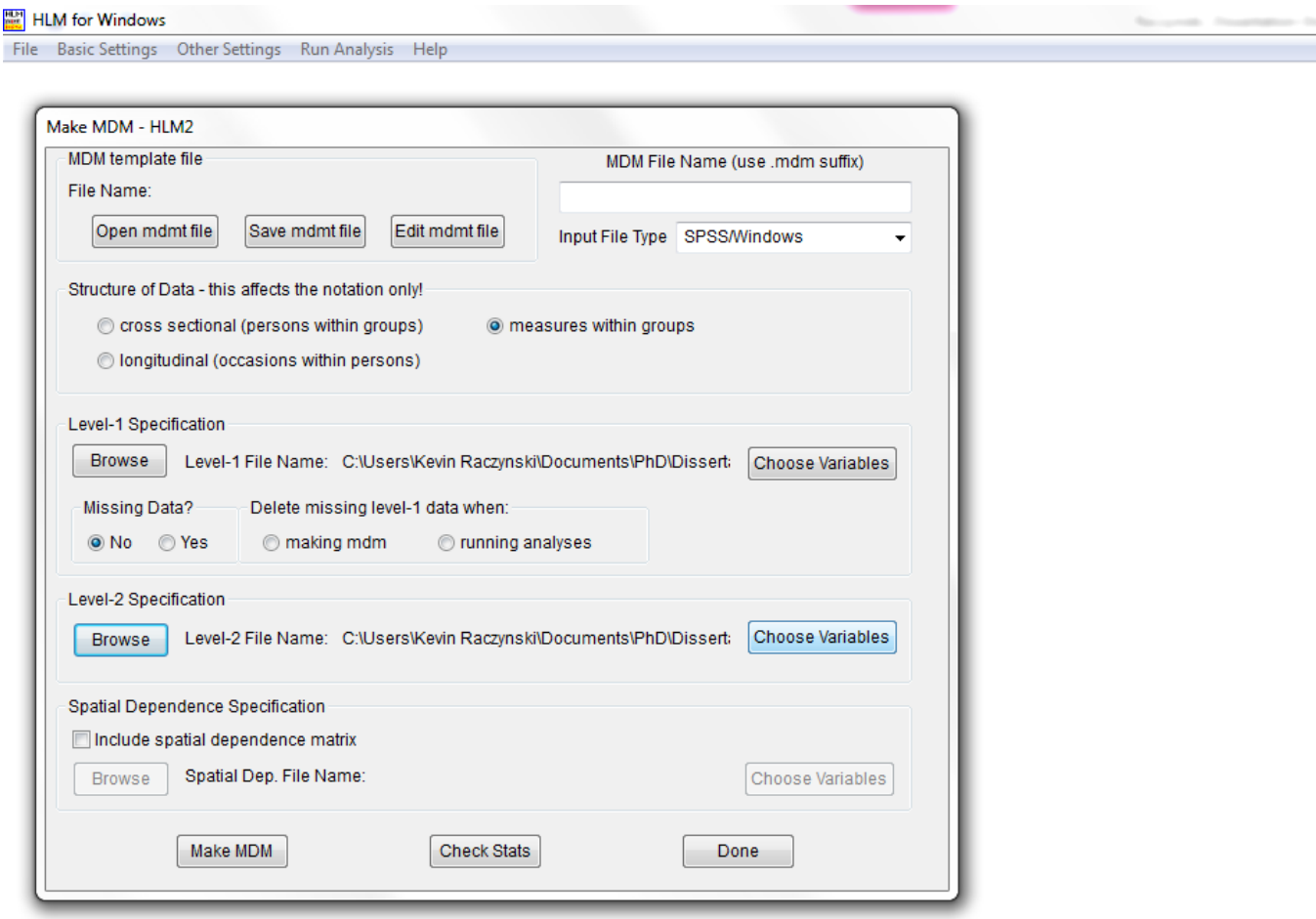

HLM for Windows
File
Basic Settings
Other Settings
Run Analysis
Help
Select MDM type
Nested Models
HLM2
HLM3
HLM4
Hierarchical Multivariate Linear Models
HMLM
HMLM2
Cross-classified Models
HCM2
HLM-HCM
HCM3
OK
Cancel

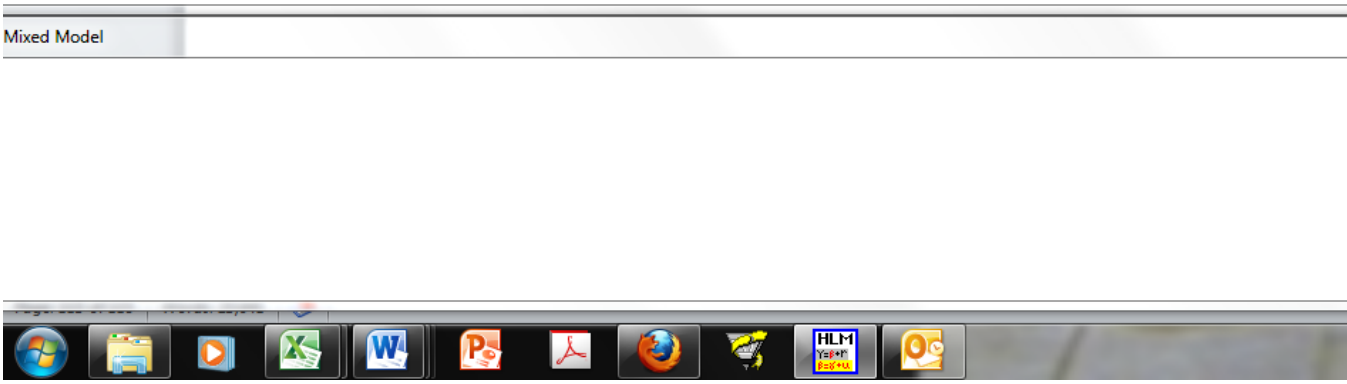

Mixed Model

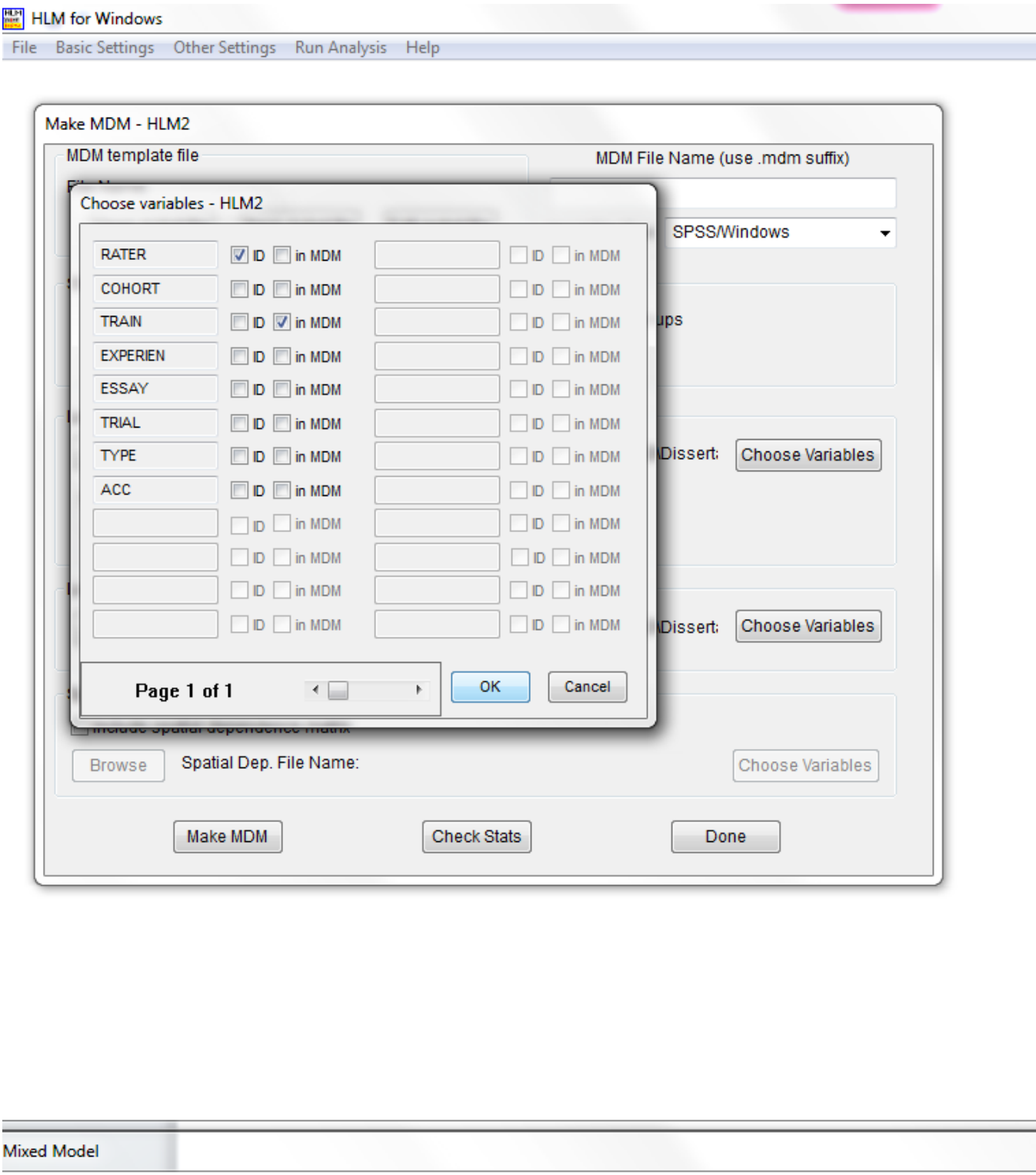

HLM for Windows
File Basic Settings Other Settings Run Analysis Help
Make MDM - HLM2
MDM template file
File Name:
Open mdmt file
Save mdmt file
Edit mdmt file
MDM File Name (use .mdm suffix)
Input File Type SPSS/Windows
Structure of Data - this affects the notation only!
cross sectional (persons within groups)
measures within groups
longitudinal (occasions within persons)
Level-1 Specification
Browse
Level-1 File Name: C:\Users\Kevin Raczynski\Documents\PhD\Dissert
Choose Variables
Missing Data?
No Yes
Delete missing level-1 data when:
making mdm
running analyses
Level-2 Specification
Browse
Level-2 File Name: C:\Users\Kevin Raczynski\Documents\PhD\Dissert
Choose Variables
Spatial Dependence Specification
Include spatial dependence matrix
Browse
Spatial Dep. File Name:
Choose Variables
Make MDM
Check Stats
Done

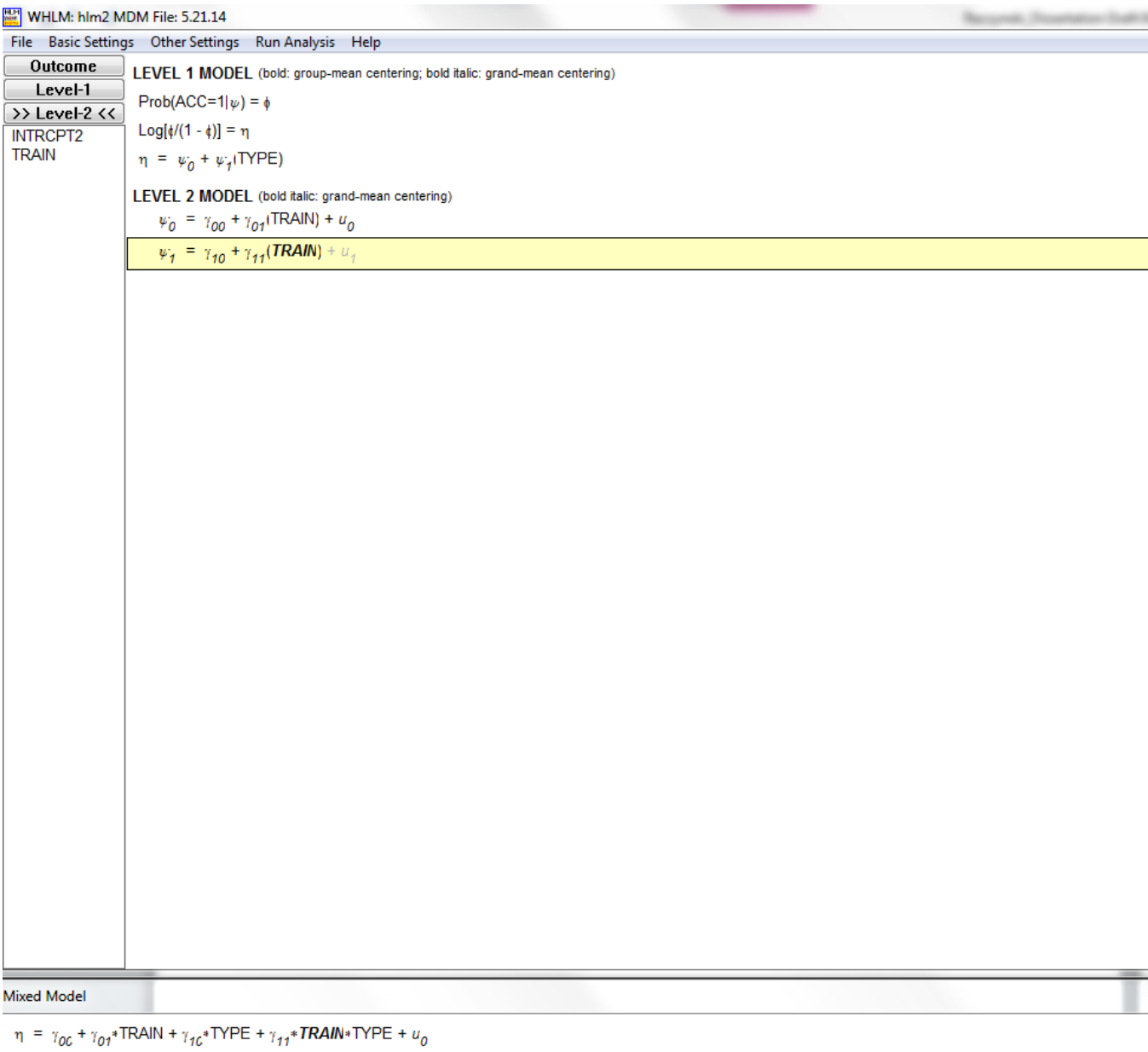

Mixed Model

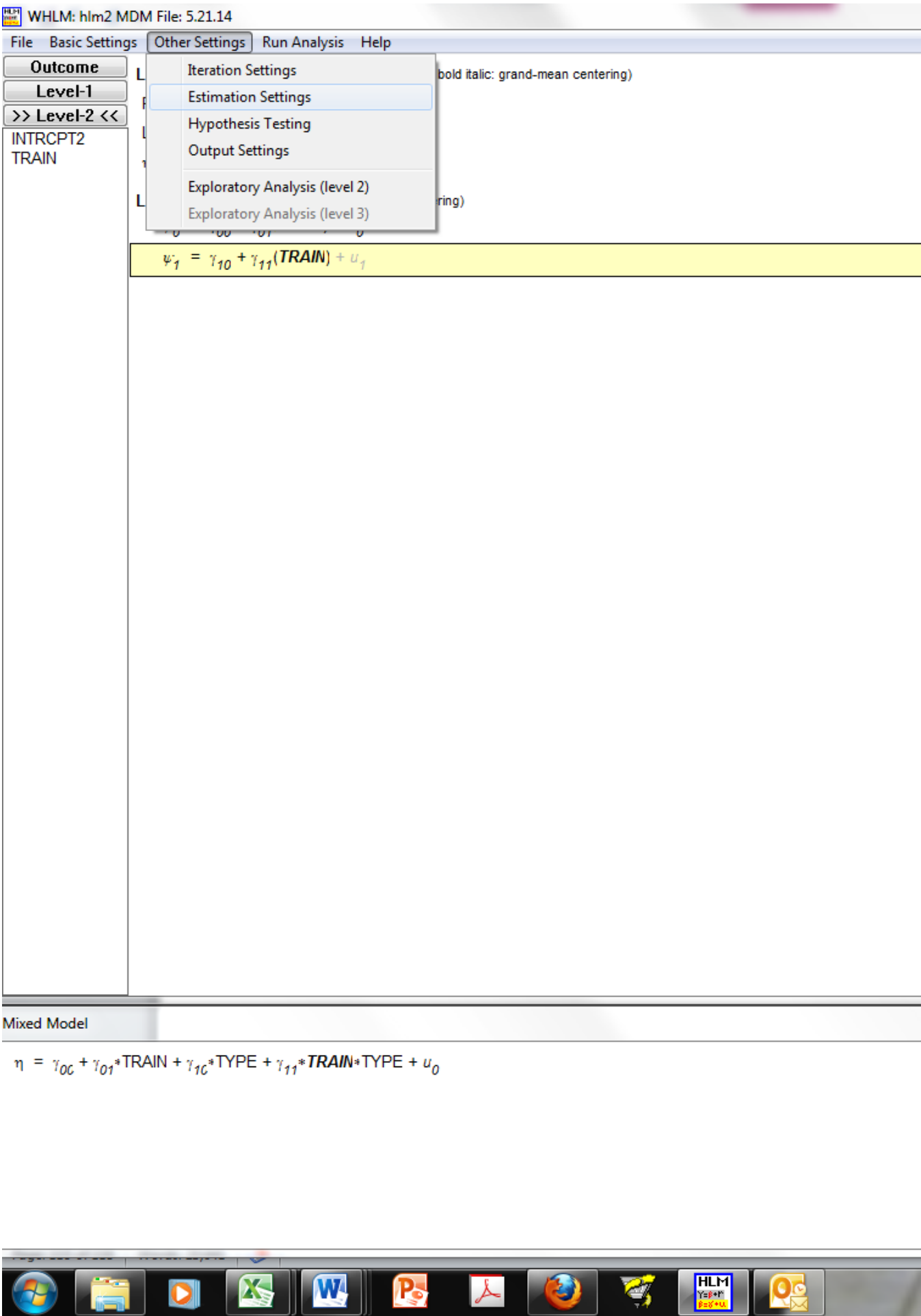

HLM for Windows
File Basic Settings Other Settings Run Analysis Help
Make MDM - HLM2
MDM template file
MDM File Name (use .mdm suffix)
SPSS/Windows
Choose variables - HLM2
RATER ID in MDM
COHORT ID in MDM
TRAIN ID in MDM
EXPERIEN ID in MDM
ESSAY ID in MDM
TRIAL ID in MDM
TYPE ID in MDM
ACC ID in MDM
Page 1 of 1
OK
Cancel
Dissert:
Choose Variables
Browse
Spatial Dep. File Name:
Make MDM
Check Stats
Done
Mixed Model

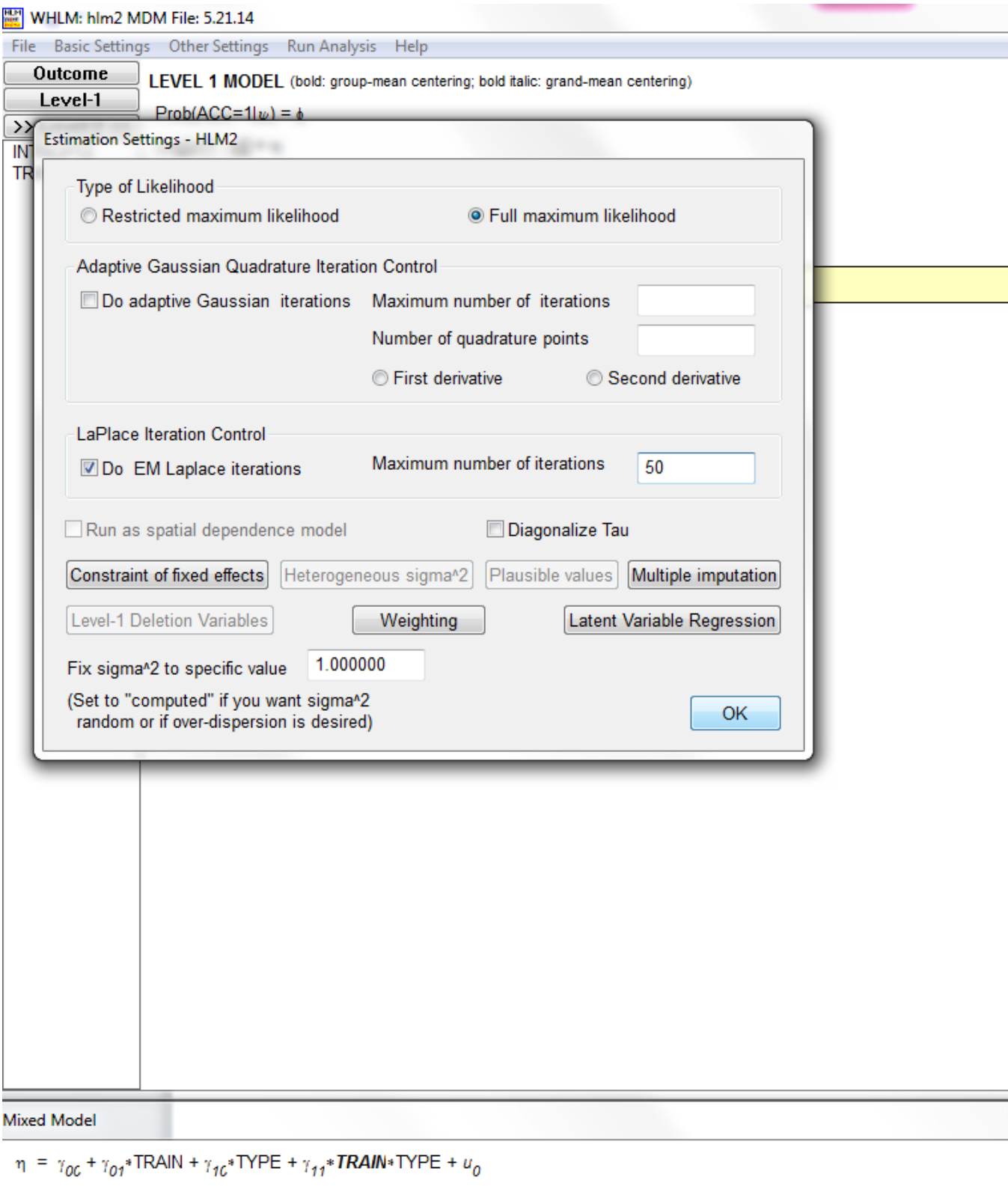

HLM for Windows
File Basic Settings Other Settings Run Analysis Help
Make MDM - HLM2
MDM template file
File Name:
Open mdmt file
Save mdmt file
Edit mdmt file
MDM File Name (use .mdm suffix)
Input File Type SPSS/Windows
Structure of Data - this affects the notation only!
cross sectional (persons within groups)
measures within groups
longitudinal (occasions within persons)
Level-1 Specification
Browse
Level-1 File Name: C:\Users\Kevin Raczynski\Documents\PhD\Dissert
Choose Variables
Missing Data?
No Yes
Delete missing level-1 data when:
making mdm
running analyses
Level-2 Specification
Browse
Level-2 File Name: C:\Users\Kevin Raczynski\Documents\PhD\Dissert
Choose Variables
Spatial Dependence Specification
Include spatial dependence matrix
Browse
Spatial Dep. File Name:
Choose Variables
Make MDM
Check Stats
Done

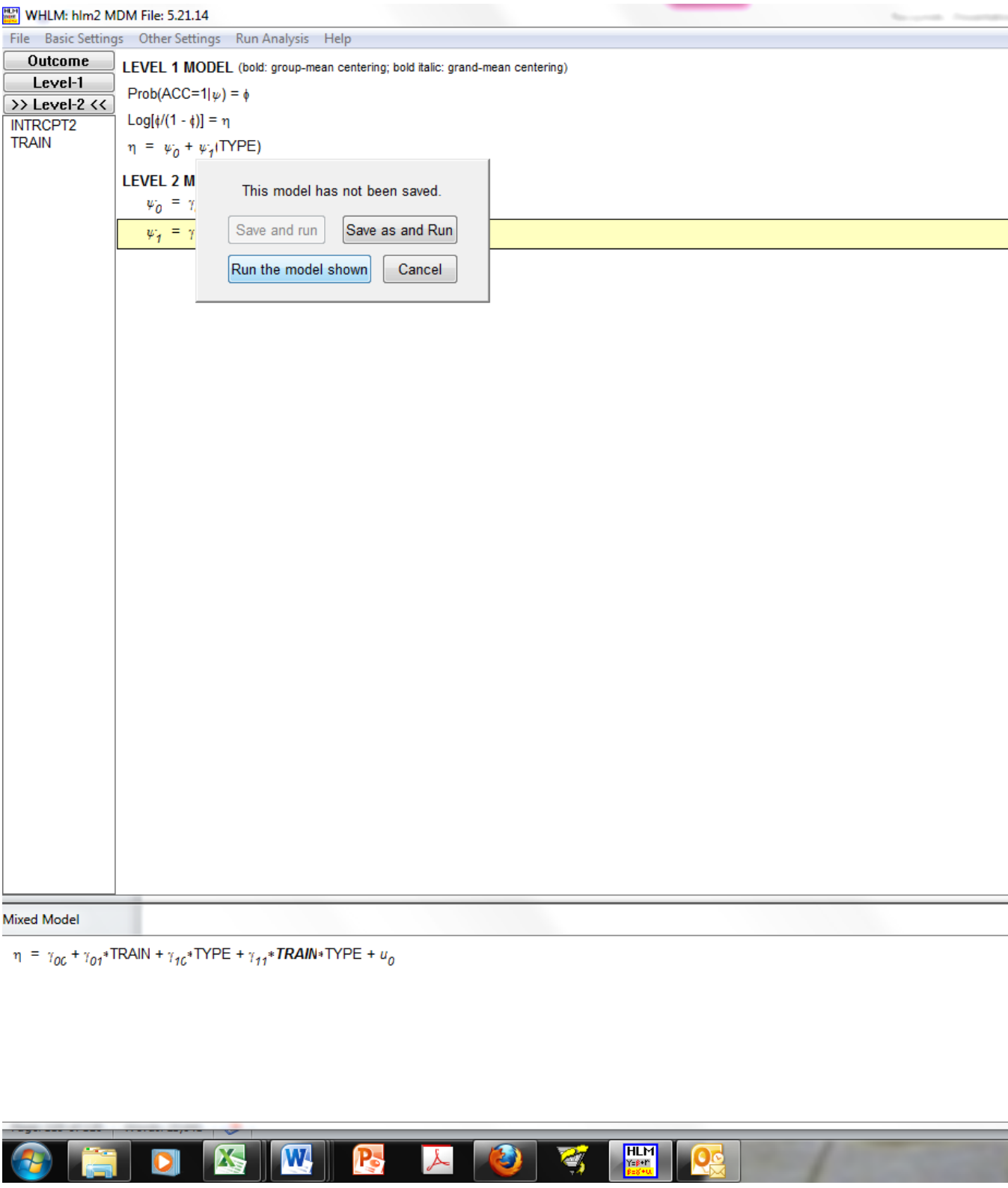

Mixed Model


HLM for Windows
File Basic Settings Other Settings Run Analysis Help
Make MDM - HLM2
MDM template file
MDM File Name (use .mdm suffix)
SPSS/Windows
Choose variables - HLM2
RATER ID in MDM
COHORT ID in MDM
TRAIN ID in MDM
EXPERIEN ID in MDM
ESSAY ID in MDM
TRIAL ID in MDM
TYPE ID in MDM
ACC ID in MDM
Page 1 of 1
OK
Cancel
Dissert:
Choose Variables
Browse
Spatial Dep. File Name:
Make MDM
Check Stats
Done
Mixed Model


WHLM: hlm2 MDM File: 5.21.14
File Basic Settings Other Settings Run Analysis Help
Outcome
Level-1
>> Level-2 <<
INTRCPT2
TRAIN
LEVEL 1 MODEL (bold: group-mean centering; bold italic: grand-mean centering)
Prob(ACC=1|φ) = φ
Log[φ/(1 - φ)] = η
η = β0 + β1*TYPE)
LEVEL 2 MODEL (bold italic: grand-mean centering)
β0 = γ00 + γ01*TRAIN) + u0
β1 = γ10 + γ11*(TRAIN) + u1
Mixed Model
η = γ0C + γ01*TRAIN + γ1C*TYPE + γ11*TRAIN*TYPE + u0


WHLM: hlm2 MDM File: 5.21.14
File
Basic Settings
Other Settings
Run Analysis
Help
Outcome
Level-1
>> Level-2 <<
INTRCPT2
TRAIN
Iteration Settings
Estimation Settings
Hypothesis Testing
Output Settings
Exploratory Analysis (level 2)
Exploratory Analysis (level 3)
bold italic: grand-mean centering)
ring)
$\psi_1 = \gamma_{10} + \gamma_{11}(\boldsymbol{TRAIN}) + u_1$
Mixed Model
$\eta = \gamma_{0C} + \gamma_{01}$*TRAIN + γ_{1C}*TYPE + γ_{11}*$\boldsymbol{TRAIN}$*TYPE + u_0


WHLM: hlm2 MDM File: 5.21.14
File Basic Settings Other Settings Run Analysis Help
Outcome
Level-1
LEVEL 1 MODEL (bold: group-mean centering; bold italic: grand-mean centering)
Prob(ACC=1|φ) = φ
Estimation Settings - HLM2
Type of Likelihood
Restricted maximum likelihood
Full maximum likelihood
Adaptive Gaussian Quadrature Iteration Control
Do adaptive Gaussian iterations
Maximum number of iterations
Number of quadrature points
First derivative
Second derivative
LaPlace Iteration Control
Do EM Laplace iterations
Maximum number of iterations
50
Run as spatial dependence model
Diagonalize Tau
Constraint of fixed effects
Heterogeneous sigma^2
Plausible values
Multiple imputation
Level-1 Deletion Variables
Weighting
Latent Variable Regression
Fix sigma^2 to specific value
1.000000
(Set to "computed" if you want sigma^2 random or if over-dispersion is desired)
OK
Mixed Model
η = γ00 + γ01*TRAIN + γ10*TYPE + γ11*TRAIN*TYPE + u0


WHLM: hlm2 MDM File: 5.21.14
File Basic Settings Other Settings Run Analysis Help
Outcome
Level-1
>> Level-2 <<
INTRCPT2
TRAIN
LEVEL 1 MODEL (bold: group-mean centering; bold italic: grand-mean centering)
Prob(ACC=1|φ) = ϕ
Log[ϕ/(1 - ϕ)] = η
η = ψ0 + ψ1(TYPE)
LEVEL 2 M
ψ0 = γ
ψ1 = γ
This model has not been saved.
Save and run
Save as and Run
Run the model shown
Cancel
Mixed Model
η = γ00 + γ01*TRAIN + γ10*TYPE + γ11*TRAIN*TYPE + u0

APPENDIX F:
Q-Q PLOTS OF LEVEL-2 RANDOM EFFECT FOR THE INTERCEPT (BY DOMAIN)