SOMESHWAR, SHONAI. Ph.D. Quality Control and the Impact of Variation and Prediction Errors on Item Family Design (2024).
Directed by Dr. Richard Luecht. 121 pp.

This two-part study examined the impact of variation within item families and errors associated with predicted item difficulty parameters on examinee test scores. Part A served as an extension of Shu et al.'s (2010) study to address how much variation matters within item families before they begin to negatively impact scores. Part A also evaluated the impact of two calibration strategies on examinee scores – CS_1 or calibrating task model families and CS_2 or calibrating individual items. Part B attempted to verify Bejar's (1983) proposition, which stated that an explained variance of 80 percent needs to be met before predicted item difficulties could be used as a substitute for empirical estimates obtained from pre-testing. Both parts relied on a simulation approach to generate differential quality of item families and predicted item difficulties across different degrees of explained variance. Some quality control (QC) statistics were used to assess any variation in IRT statistics and their impact on examinee scores.

The results from Part A suggested that CS_1 and CS_2 were appropriate for low variation ($< 0.2\ \sigma$) and high variation conditions ($0.2\ \sigma$ to $0.5\ \sigma$), respectively. While a within task model family variation of 0.2σ and 0.5σ showed increased trends in bias and RMSE for moderate and high conditions under CS_1, this variation ultimately did not result in significant score differences between the two calibration strategies, especially for longer tests.

The findings from Part B showed how IRT models are robust enough to withstand error introduced by poorly predicted difficulty parameters used to score examinees. While the estimated scores remained relatively unaffected, the residual-based fit statistics (for the probability of an examinee endorsing an item based on the estimated scores and predicted item parameters) revealed larger errors as the correlations between the true and predicted item difficulties decreased. Results

from the person fit analysis revealed that misfit is more likely to occur for the lower R^2 conditions. Overall, the results from both Part A and Part B showed that developing a QC system for modern item and test development approaches is feasible and even necessary.

QUALITY CONTROL AND THE IMPACT OF VARIATION AND PREDICTION ERRORS ON ITEM FAMILY DESIGN

by

Shonai Someshwar

A Dissertation
Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro

2024

Approved by

Dr. Richard M. Luecht
Committee Chair

DEDICATION

To my grandfather, Mohan Someshwar

APPROVAL PAGE

This dissertation written by Shonai Someshwar has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

Dr. Richard Luecht

Committee Members

Dr. Robert Henson

Dr. Micheline Chalhoub-Deville

Dr. Kyung Yong Kim

May 14, 2024

Date of Acceptance by Committee

May 14, 2024

Date of Final Oral Examination

ACKNOWLEDGEMENTS

This dissertation may bear my name, but it would not have been possible without the support system I have been fortunate to have over the last five years. Central to that support system is my advisor, Dr. Richard Luecht. I am indebted to him for his unwavering support, reassurance, and guidance during my doctoral journey. My committee members – Dr. Micheline Chalhoub-Deville, Dr. Robert Henson, and Dr. Kyung Yong Kim – have provided me with invaluable feedback and insight throughout the dissertation process. Additionally, I am deeply grateful to Dr. Tiffany Tovey and Dr. Devdass Sunnassee for their guidance through OAERS, coursework, and research. I would like to thank my mentor in the certification world, Anjali Weber, who helped shape my career in the United States. Lastly, to my dear friends - both within and outside the ERM department – whom I cannot thank enough for the constant check-ins and words of motivation.

I believe that my current position in life and career would be unfathomable without the many sacrifices that my parents made to send me halfway across the world to pursue my goals. I owe them everything. My sister, Mishri, generously opened her home to me before, during, and post COVID-19. She has been my constant source of reassurance and support. I would be remiss for not thanking her lovely children, Rahm and Michael, for always reminding me to never take life too seriously. My brother (in-law) Simon has been instrumental in keeping me calm throughout this process. He has always been in my corner, even if at times it meant going up against some very opinionated Someshwars. I would not have survived these past few years without the love and support of my extended family – the Tonses, the Slobodniks, the Rayzs, and the Kellys.

Finally, to the one who taught me the value of discipline and hard work – my grandfather, my biggest cheerleader. PA, while you are no longer with me, I will continue to make you proud for the rest of my days.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Some modern item generation and test development strategies are shifting from a focus on developing individual items to instead focus on one of three strategies: (1) automatic item generation using parent items, smart items or item shells to mass produce items (e.g., clones or nominal variants; Haladyna & Shindoll, 1989; Bejar, 2002); (2) cognitive design of task models and item families adhering to complexity feature designs and specifications (Sheehan & Mislevy, 1990; Embretson, 1999; Embretson and Daniel, 2010; Luecht, 2012; Luecht & Burke, 2020) ; or (3) using on-the-fly item predictions using artificial intelligence (AI) models (e.g., Settles, et al, 2020). All three strategies can reduce or eliminate the costs, resource demands, item-exposure security risks, and statistical stability of traditional item-level pretesting. However, they can also introduce errors in scoring or decision making when using a common set of item parameter estimates for an entire family of items or when making model-based predictions of the item parameters from item features contain nontrivial amounts of variation.

The purpose of this study is to show how different magnitudes of variation within item families from an Assessment Engineering (AE; Luecht, 2006, 2007, 2009, 2012) perspective or different degrees of item-parameter predictions affect score and decision accuracy. This study will consider both within item family variation and variation due to item parameter prediction errors, and is thus divided into two parts – Part A and Part B. Part A of this study considers within item family variation by expanding the work of Shu et al's (2010) study, which evaluated differential variation in item characteristics at three levels - task models, templates, and individual items, for a multidimensional formative assessment. The present study sought to systematically vary the quality of items within task models and examine their impact on examinee scores through a series of quality-control indicators for a unidimensional summative assessment. The second part of this

study, Part B, focused on evaluating the impact of item parameters - obtained through varying degrees of prediction – on examinee scores. More specifically, Part B sought to verify Bejar's (1983) claim that there needs to be an 80-percent variance explained in predicted item difficulty before the predictions can be used as a substitute for empirical difficulty estimates for scoring purposes. Thus, this study addresses the central issue of quality-control (QC) regarding the performance of items within families and the impact of using predicted item parameters to detect *how much variation matters* in terms of statistical isomorphicity before it affects score precision.

The subsequent sections of this chapter provide the context and rationale for this study, specifically comparing traditional versus modern item and test development methods, highlighting the problem areas in the former, and the implicit requirements for item-level pre-testing as a key part of item QC.

Different Perspectives of Item and Test Development

Traditional methods of test development such as those explained in *The Handbook of Test Development* (2016), Schmeiser and Welch (2006), and the *Standards of Educational and Psychological Testing* (the Standards; AERA, APA, NCME, 2014), rely on a series of well-defined guidelines that support the effective creation of test specifications, item development, test assembly, and scoring rubrics. Their primary focus is on developing individual items, largely depending on the judgment and expertise of subject matter experts (SMEs) (Luecht & Burke, 2020). These guidelines state that perhaps the most important step is recruiting qualified content experts and training them to develop high-quality items. They go on to discuss the many stages of item development: from writing items to reviewing and editing the items according to an organization's editorial standards, and finally, conducting appropriateness and fairness reviews. Once items have passed the many stages of review, they may be embedded in operational test

forms to be pretested within the target examinee population. Statistical item analytics based on the pretest data may call for further qualitative evaluation of the items (e.g., answer key validation or confirmation). Items that "pass" the operational pretesting phase are typically added to an item bank and designated as eligible for inclusion on future test forms. Items with exceptionally poor item pretest statistics and/or other unrepairable qualitative issues such as confusing or inappropriate wording may be removed altogether from an item bank. These sets of guidelines serve as testaments to the QC mechanisms in place at every stage of item (and test) development. In short, QC is not a novel problem in testing. Furthermore, these QC procedures can be statistical or qualitative in nature.

Statistical QC of newly developed items using traditional methods typically means performing an item analysis, a fundamental psychometric process that affects both test construction and scoring (Luecht, 2014a). Item analysis as crucial as it is, cannot begin without an elaborate process of extracting, cleaning, and restructuring testing data, thereby ensuring its quality, integrity, and accuracy (Luecht, 2014b). Once assured of the quality and integrity of the data, an item analysis can be performed to supplement the qualitative evaluation process during item review. Item analysis provides quantitative indicators in the form of statistical properties of examinee responses to each item on the test (Crocker & Algina, 1986; Luecht, 2014a). It further examines whether these statistical properties are appropriate for the ability level of examinees who take the test (Allen & Yen, 1979).

Common statistical indices in a typical item analysis can be grouped into three categories (Crocker & Algina, 1986): 1) indices that are based on the distribution of responses to each item, such as item difficulty (Allen & Yen, 1979; Schemiser & Welch, 2006); 2) indices that express the relationship between the examinee responses to items and the criterion of interest such as item

discrimination (Allen & Yen 1979); and 3) indices that are a function of both the means and variances as well as the criterion of interest, such as reliability and validity coefficients. These item performance statistics are usually obtained using classical test theory methods that use the number correct raw scores (Allen & Yen, 1979; Crocker & Algina, 1986; Raykov & Marcoulides, 2010) and item response theory methods that places item difficulty, item discrimination, and examinee proficiency on a common scale of latent ability (Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Lord, 1980; Wright & Stone, 1979). Both sets of methods require adequate data to yield stable item parameter estimates. Moreover, other variables such as innovative/technology-enhanced item types, computer-adaptive testing and multistage testing, restriction of range in test scores and homogenous samples further contribute to the complexity of item analysis (Luecht, 2014b). While IRT models are sophisticated enough to handle complex test design factors, the calibration process still requires large sample sizes to render useful item parameter estimates (Luecht & Burke, 2020). In addition, the continuous pre-testing of items for to support large-scale assessments is an expensive process that increases the risk of item exposure, enhancing the threat to test security. Given the high costs associated with developing and pre-testing new items, only to find that a little over half the items can be used in operational test forms seems to be a poor return on investment (Case et al., 2001).

Modern Approaches to Item and Test Development

Alternatives to traditional item and test development seek to address the issues concerning their traditional counterparts by conceptualizing families of items to promote mass production and generalizability of item parameter estimates. One of these approaches, Automatic Item Generation (AIG), uses parent items as item shells, item templates, or item models (Hively et al., 1968; Osburn, 1968; Bormouth, 1970) to produce multiple children or sibling items using a generative

mechanism such as a generator (LaDuca, et al, 1986; Bejar & Yocum,1991; Bejar, 2002, 2010; Gorin & Embretson, 2013; Gierl & Haladyna, 2013a). Considering that the cost of developing a single new item using traditional methods is between $1500 to $2500, the rationale for using AIG is to reduce item production costs (Rudner, 2009; Gierl & Lai, 2012; Gierl & Lai, 2013a; Gierl et al., 2021). This generative form of item development also suggests that items nested within an item model are expected to behave the same way, having inherited the psychometric characteristics of the previously validated parent item, and thus granting them statistical or psychometric isomorphicity (Bejar, 2002). Thus, when the resulting item instances have similar psychometric characteristics, they are called isomorphs (Bejar, 2002). If their psychometric properties vary within an item family, they are then referred to as variants (Gierl & Lai, 2012). The psychometric isomorphicity of item families can therefore, go a long way in addressing the constraints produced by traditional pre-testing methods.

Cognitive design of task models and item families that abide by specific task complexity features is the second approach to modern item and test development. This approach includes principled assessment design (PAD) frameworks such as cognitive design systems (Embretson, 1998); evidence-centered design (Mislevy et al., 2003; Mislevy et al. 2002); and assessment engineering (Luecht, 2006, 2007, 2010, 2012). These are top-down approaches that use cognitive or task models that specify construct-relevant knowledge and skills as well as the relationships among them in terms of their complexity. These cognitively oriented specifications are intended to determine the difficulty of items in the family (Embretson & Daniel, 2008; Luecht et al., 2009; Luecht, 2013; Luecht & Burke, 2020). The purpose of using cognitive task models is to replace traditional test blueprints and specifications that show no connection between the blueprint categories and empirical estimates of item difficulty, to design item families that are intended to

behave the same way in terms of their content and psychometric characteristics (Luecht, 2008, 2009; Luecht et al., 2010). Thus, one of the core principles of AE and other PAD frameworks is that they design classes of items to be substantively isomorphic in terms of content and complexity and statistically isomorphic in terms of item difficulty. In other words, by basing families of items on explicitly stated cognitive specifications and the evidence that is needed to demonstrate proficiency, there is deliberate control over their content and *complexity by design* (Mislevy, 2007; Mislevy & Haertel, 2007; Burke et al., 2020).

Calibration of Item Families

PAD approaches can be used to implement AIG since they are based on underlying cognitive specifications or cognitive theories that support the development of task models and templates/item models. Even though the assumption is that the resulting items from each item model are expected to be psychometrically isomorphic i.e., exhibiting no meaningful differences in their psychometric properties, this assumption needs to be verified through calibration of the item families. There are some methods for examining item families for their psychometric isomorphicity, three of which were noted in Sinharay et al's. 2003 study. The first is the unrelated siblings model. This model ignores family membership of items, assuming independent item response functions for all tasks. Sinharay et al. (2003) applied the unrelated siblings model in the context of the one-parameter logistic or Rasch item response theory model. The researchers noted that the limitation of this model is ignoring family membership, therefore requiring larger sample sizes for calibration compared to other models. The second model, the identical siblings model, assumes that all items within the same family have the same item response function. However, this model assumes total psychometric isomorphism among all items, and ignores the possible variability of items within the item family, thereby providing incorrect item parameter estimates.

Finally, the third model is the related siblings model that does not assume that all the items in the same family have the same item response function. It is expressed as a hierarchical model, where the first component is expressed as a Rasch or one-parameter item response function, where the difficulties for the item family are assumed to be equal. The hierarchical piece specifies a distribution of parameters for relating each item in the family, indicating that each item-specific parameter is modeled as varying around a family-specific mean (Faye, et al., 2018). All three models have been applied in both dichotomous IRT models (Glas & van der Linden, 2003; Lathrop & Cheng, 2017) and polytomous IRT models (Cho et al., 2014; Geerlings et al., 2011).

The third and final approach to modern item and test development is using AI models to generate items and predict their difficulties on-the-fly. Recent studies in the field of computational linguistics have used a combination of machine learning and natural language processing models to develop tests of language proficiency (Alsubait et al., 2013; Kurdi et al., 2017; Ha & Yaneva, 2018; Settles et al., 2020). Of these studies, Settles et al. (2020) proposed that these methods could solve the "‘‘cold start’’ problem in language test development, by relaxing manual item creation requirements and alleviating the need for human pilot testing altogether." (p. 247). In other words, their study sought to address the issue of obtaining difficulty estimates for generated items without the need for any response data. The researchers sought to automate the estimation of item difficulty from linguistic features of the Common European Framework Reference (CEFR) difficulty scale. They utilized linguistic models to estimate item difficulty of vocabulary and passage-based items that were spread across five item formats. The combination of machine learning, NLP, and linguistic models generated plausible items across the five item formats and subsequently computed the difficulty for each item. The researchers observed high correlations between the scores of their automatically generated English language test with those of the TOEFL iBT and

IELTS. They also found a "strong relationship" between the machine-learning and NLP estimates of item difficulty and IRT difficulty estimates from operational data.

Rationale for the Present Study

The research on modern approaches to item and test development described above demonstrates three important improvements over traditional methods:1) leveraging advancements in technology to support the item generation increases efficiency and reduces the cost per item over a period of time; 2) calibrations are either taking place at the family level or being produced on-the-fly, which at the very least reduces the sample size of items requiring traditional pre-testing; and 3) the items and tests generated from these methods support a stronger construct validity argument compared to traditional methods. Yet, relying on family-level sets of calibrations or predicted item parameters from item features can potentially introduce error in scoring, thereby affecting decisions based on these scores. It therefore becomes crucial to decide on an appropriate calibration strategy for item families that maintains low within-family variance or acceptable variation within it. Moreover, it may be essential to use predictions of item parameters that came from regression models with a high proportion of explained variance or R^2. These two points are addressed by Shu et al. (2010) and Bejar (1983), respectively. Shu et al. (2010) compared three calibration strategies at the level of task models, templates, and individual items for a multidimensional formative assessment. Test forms were assembled from twelve separate item pools in various mixtures across seven test assembly conditions (from best to worst in terms of item pool quality), where each item pool varied in terms of its quality of item parameter estimates. One of the objectives of the study was to demonstrate that task model and template-level calibrations could be used only if the item templates were actually able to control difficulty. The second objective was to evaluate the impact of using these calibrated statistics on scoring

examinees, especially when using item of poor quality to assemble the test forms. This study provided two important findings: 1) an evaluation of root mean square deviation (RMSD) plots revealed that while individual item difficulty estimates varied within each test assembly condition or calibration strategy, the locations were more or less maintained; 2) the template based calibrations performed as well as the individual item calibrations even for the worst test assembly conditions. The findings also suggest that if a particular minimum level of item discrimination were achieved and variance of item difficulties within templates were within "acceptable tolerances" then template-level calibrations could be used for task models while meeting QC requirements. However, how much tolerance or variation is acceptable remains unclear.

Bejar (1983) studied the degree to which subject matter experts could predict the difficulty and discrimination of items on the Test of Standard Written English. The objective of this study was to determine if, after substantial training of subject matter experts in rating items for their difficulty and discrimination, these ratings could be used to replace empirical estimates of item difficulty for scoring purposes. The results showed that despite extensive training of subject matter experts, the accuracy of the predicted item statistics was too low to be used for any further analysis. Moreover, it was concluded that a requirement of an explained variance of 80 percent. ($R^2 = 0.8$) needs to be met before predicted item difficulties could be used as a substitute for empirical estimates obtained from pre-testing. Forty years later, with the use of AI models to predict item parameters, would the same explained variance need to be met for predictions to substitute empirical estimates? Examining the different degrees of predictions of item parameters in terms of their impact on examinee scores may provide some insight in answering that question.

Research Questions

The present study sought to build on the work of Shu et al. (2010) and Bejar (1983) to specifically address:

1) How much variation matters in task model families without impacting examinee scores?

   1a. Under which conditions would calibrating the task models and/or calibrating the individual items be most appropriate?

   1b. Which calibration strategy would have the most (or least) impact on examinee scores?

2) How do different degrees of explained variance in predicted item parameters impact scores and person fit?

This study aimed to answer these questions by applying a simulation approach in both parts A and B. Part A addressed the first research question by using item parameter estimates that were generated from three conditions of variation – low, moderate, and high – at three different test lengths, each assembled using the seven test assembly conditions from Shu et al's (2010) study. It also answered the first research question as it pertained to the first two modern item development approaches mentioned earlier – AIG and PAD approaches. Part B addressed the second research question by generating predicted item parameters across different degrees of explained variance or R^2. This study intentionally varied the quality of predicted item parameters based on the strength of the correlation between the predicted item difficulties and empirical item difficulty estimates, thereby testing Bejar's (1983) claim. Ultimately, the goal of parts A and B was to extend, test, and verify findings from previous research and assess how they impact examinee scores. Parts A and B also intended to serve as guidelines for choosing calibration strategies that are appropriate for the quality of item families used in practice and provide insight regarding the use of predicted item parameter estimates for scoring examinees.

CHAPTER II: LITERATURE REVIEW

This chapter contains two main sections. The first section examined the literature supporting traditional item development and quality control procedures. The literature in this section briefly focused on the process of developing traditional items and tests while also describing mainstream statistical quality control indexes used in practice. This section also discussed some issues with traditional statistical QC that warranted a shift in conceptualizing items. The second section discussed modern approaches to item and test development and contained two parts. The first addressed relevant literature on the development and use of item families, thus focusing on Automatic Item Generation (AIG) and Principled Assessment Design (PAD). The second part examined literature concerning on-the-fly generation of items and their psychometric characteristics while also providing a brief overview of literature that has investigated item difficulty prediction. Finally, the literature review concluded with gaps found in previous studies that the present study aimed to address.

Traditional Item and Test Development

Drasgow et al. (2006) envisioned 21st century testing programs as those that leverage the power of technological advancements and innovative assessment frameworks to become an integrated system of systems. With large scale testing programs requiring an increasing number of high-quality items to support continuous testing, it becomes imperative to reconceptualize how items and tests are developed to meet these demands and cater to the evolving needs of testing programs. Traditional item and test development practices are limited in their ability to meet these demands for a few reasons: 1) they are labor-intensive and time consuming; 2) they require continuous item pretesting; 3) they lack the robustness to handle complex assessment design.

Traditional item development is a rigorous process that requires suitably trained item writers to develop and refine items that can eventually perform well on an exam (Welch, 2006). These methods place the onus of producing high-quality items on the item writer (Mosier et al., 1945). The items are usually multiple-choice questions, which are still widely used today even as innovative item types gain popularity (Parshall et al., 2002). Mosier et al. (1945) who provided one of the earliest guidelines for writing multiple choice items, said that items must be phrased in such a manner that all examinees can understand the task set, and those who have the requisite knowledge will provide the correct answers. Since then, there have been extensive guidelines developed across the years for item and test development. However, violations still occur despite following these guidelines, making it challenging to develop test items that meet psychometric standards without adequate pretesting. Ultimately, current item development practices, even when executed by the most skilled item writers, represent creative expressions of their ideas of the construct. Each item, is therefore, designed to be unique and less generalizable, thereby favoring more of an artisanal result over a scientific process.

Second, once items are developed, they are subjected to both qualitative and statistical review. It is at this point that items are pretested and then analyzed to obtain their performance statistics. Since there is no mechanism for establishing reliable a priori estimates of item performance statistics, pretesting is essential. Moreover, previous studies have shown that despite receiving extensive training on item performance, primarily focusing on understanding what makes an item difficult, item writers' predictions of item difficulty are inaccurate, reinforcing the need for pretest statistics obtained through item analysis to determine next steps (Camerer & Johnson, 1991; Lorge & Diamond, 1954; Nathan & Koedinger, 2000; Nathan et al., 2001; Tinkelman & Sherman, 1947).

Item analyses are typically performed using classical test theory and/or item response theory methods to produce indexes such as item difficulty, item discrimination, and reliability to gauge whether an item is suitable for the test. In classical test theory, item difficulty, more commonly referred to as a p-value, refers to the proportion of examinees who answered the item correctly and reflects whether the level of difficulty is well-suited for the intended purpose of the test (Crocker & Algina, 1986; Luecht, 2014a). While testing organizations may vary in their choice of appropriate item difficulties for their tests based on operational policy, item difficulties in the range of 0.3 – 0.7 tend to provide the most information about differences in examinees taking the test (Allen & Yen, 1979). Items that have difficulties close to 0 or 1 are typically flagged for further review or completely discarded from use.

Item discrimination is an index that denotes how well items differentiate between examinees who are relatively high on the criterion of interest versus those who are comparatively low (Crocker & Algina, 1986; Schmeiser & Welch, 2006). While there are five parameters of item discrimination – index of discrimination, point biserial correlation, biserial correlation, phi coefficient, and the tetrachoric correlation coefficient – the point biserial and biserial correlations are commonly used in both credentialing exams and achievement tests (Luecht, 2014a; Schemiser & Welch, 2006). The point biserial correlation represents a Pearson product-moment correlation between a dichotomously scored item and the total test score, where the total score is computed with that item removed. High point biserial correlations are desirable while low or negative point biserial correlations signify that the item is either not discriminating well between examines or indicating that low scoring students on the total test score are performing better relative to high scoring students. When the underlying distribution of the dichotomously scored item is normal, the correlation is then called a biserial correlation (Allen & Yen, 1979). In practice, items that have

a point biserial correlation lower than 0.10 or biserial correlation lower than 0.21 are flagged for further review (Luecht, 2014a).

Other statistical indices include key validation analysis (Luecht, 2014a) and the reliability of test items generally measured by coefficient alpha (Crocker & Algina, 1986; Cronbach, 1951). A key validation analysis or distractor analysis helps in identifying incorrect response options that show high positive correlations with the total scores or high proportion of examinees choosing them over the answer key. Such results warrant a review of item distractors for multiple-choice and other forms of selected-response items. A reliability index representing the internal consistency of the items on the test – coefficient alpha - is a function of the number of items on the test and the average inter-correlation among items.

Item response theory (IRT), considered to be an improvement over classical test theory methods, estimates item and person characteristics such that they are on the same underlying scale (De Ayala, 2009; Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Luecht, 2014). Most IRT models assume that the latent variable or trait is unidimensional, although there are models that can be used to tackle multidimensional representations of the latent variable (Reckase, 2009). The basic premise of IRT is that the items can differentiate among people located at different points along the unidimensional continuum (De Ayala, 2009). These models attempt to explain the relationship among the characteristics of items, the responses to the items, and the latent variable of interest. They also make three important assumptions of monotonicity, unidimensionality, and invariance that when held, offer several analytical advantages over classical test theory (Hambleton et al., 1991). There are a variety of models that can be used to evaluate the quality of dichotomous and polytomous items that have either ordered or nominal score categories.

Items with negative slopes or item discrimination parameters, extreme values of item difficulty, and high values of the pseudo-guessing parameter are typically flagged for additional review or are ultimately discarded from use. In addition, item and test characteristic curves (TCCs), item and test information functions (TIFs) are plotted to examine their performance. Since IRT models are regression models with unobserved explanatory variables, regression-based diagnostic statistics can be used to evaluate fit between the observed item response scores and model-based predictions (Luecht, 2014b). Ames and Penfield (2015) provide a comprehensive overview of item fit statistics that they group as chi-squared approaches and likelihood ratio approaches.

Despite having a host of techniques to establish statistical quality control over traditionally developed items, challenges remain. Luecht (2014b) identified several issues in item analysis that are encountered in credentialing programs: 1) small sample size; 2) homogeneity of the population and restriction of range; 3) establishing a consistent reference group; 4) complexities arising from using technology-enhanced items and computer-adaptive or multistage tests. Item analysis requires a large enough sample size to produce stable item parameter estimates. The sample size may vary depending on the measurement model used and while classical test theory offers some flexibility, IRT models are generally quite data hungry. In other words, these measurement models require adequate high quality response data to provide useful item parameter estimates. Thus, item pretesting can be an expensive process, especially when considering small testing programs.

Homogenous populations, meaning those professionals who share the same qualifications and training to meet eligibility criteria for certification exams, impact item analysis by imposing a restriction on the range of total scores. These restrictions on the variance of total scores can diminish their correlations with item scores or sub scores. While corrections for these variance restrictions are available, they may not provide accurate results when using small sample sizes.

Moreover, with credentialing tests being designed to provide maximum precision at the cut score, items may be designed to be very easy or difficult, systematically reducing their variance. Furthermore, the item-total score correlation may be lower as well. When statistically evaluating these items, it may deceptively appear that they are of low quality, when in fact they are meeting the specific requirements of test design.

Luecht (2014b) defines a reference group as, "a designated subset of the larger population of test takers and can be useful to help maintain reasonable consistency of the sampling used in IAs and other data analyses over time". (p. 138). Even though all examinees obtain scores, it is only the reference group that is included in the item analysis, and on which quality assurance is performed. Although driven by policy, the reference group is useful in evaluating results of the item analysis and therefore, an important consideration.

As the needs of testing programs evolve to include innovative item types or technological enhanced items, and computer adaptive tests, item analysis becomes increasingly more complex to support their continuous use. Technology-enhanced items require the identification of a valid score scale to compute their item-total correlations, which can be challenging to achieve in practice. Computer-adaptive tests and multistage tests aim to customize the difficulty of test forms to examinee ability to provide them with an enhanced testing experience. In addition to placing a burden on item banks, such tests further add to the complexity of item analysis. For these reasons, traditional item analysis may be inappropriate for evaluating innovative items and items for computer adaptive or multistage tests.

Despite the rigor involved in developing and analyzing traditionally developed items, the fact remains that they are individual items which are unique and not exchangeable in terms of their psychometric properties (Luecht & Burke, 2020). If these items operate in unpredictable ways and

fail to survive pretesting protocols, then the concern is whether traditional item development and quality control mechanisms can be sustainable in the long run. Alternative approaches may be needed to address these issues, particularly, redefining how items are conceptualized and used for testing purposes. Some alternatives such as model-based approaches have proposed a paradigm shift in thinking from developing individual items to developing families of items. These item families have been argued to be exchangeable in terms of their content and psychometric characteristics such as difficulty, thereby making the item development more efficient and sustainable than traditional methods. The next section describes modern alternatives to traditional item and test development, where two of the approaches involve developing item families instead of individual items, while the third demonstrates automatic generation of items along with their estimates of difficulty.

Alternatives to Traditional Pre-Testing

Some alternatives to traditional pre-testing take the form of model-based approaches. Model-based approaches indicate a template-based type of item generation, involving an item model or template to develop a family of items that operate in psychometrically and structurally predictable ways. The literature on these approaches proposes the use of an item model or template to develop large numbers of items that can satisfy the need for high demand in large scale assessments. The other type of model-based of approach forgoes the use of templates in favor of large language models to automatically produce items and estimate their difficulty on-the-fly. This approach has primarily been used in the area of language assessment. Both template-based and non-template-based methods will be discussed in the subsequent sections.

Automatic Item Generation (AIG)

AIG suggests an iterative process of item generation that is guided by the systematic manipulation of elements present in an item model – a template that houses fixed and variables elements for the generation of multiple items (LaDuca et al., 1986; Haladyna, 2013; Osburn, 1968). More recently, Gierl and Lai (2013) defined AIG as, "the process of using models to generate items with the aid of computer technology." (p. 1).

While AIG may be a modern form of item development, its conception can be dated back to the 1960s and 1970s; specifically, Hively et al. (1968), who used item forms to develop items for a test of basic arithmetic skills. Their approach towards developing items for the arithmetic test was based on templates (item forms), which had fixed and variable elements. The items were intended to be "hypothetically equivalent", assessing the same skill and having the same difficulty for each diagnostic category. Bormuth (1972) proposed a theory of item generation from prose and believed that traditional methods of item writing were subjective and inefficient. His theory stated that prose text could be transformed into "wh-questions". While Bormuth's theory may have propelled the conversation surrounding item generation from prose, Finn (1975) highlighted certain limitations of syntactic analysis and proposed a solution to these limitations using Fillmore's (1968) grammar. These studies asserted a rule-based method of item generation that could potentially suppress the subjectivity involved in item writing and promote similar structural and psychometric characteristics amongst the items produced. More recent literature has utilized item models and rule-based generation to argue a three-step approach to generating items (Gierl & Lai, 2013a).

Prior to item generation, test developers must first identify the content and cognitive domain of interest by collaborating with content experts. The purpose of specifying both a content

and cognitive domain is to establish the construct to be measured as well as the knowledge, skills, and abilities (KSAs) that are required to solve assessment tasks associated with that construct. Next, cognitive models are developed either using a cognitive theory (Bejar, 1993; Embretson, 1998, 1999) or parent items (a sample of what should be modelled) (Bejar, 2003; Drasgow et al., 2006; Gierl & Lai, 2016; Gierl et al., 2021). Thus, the cognitive model specifies how the content is organized in the knowledge domain.

Once the cognitive model is developed, the development of an item model involves organizing the content in a test item format complete with the stem, response options, and any auxiliary information deemed necessary by content experts (Gierl & Lai, 2013a). Item models can be developed in two ways – to develop isomorphs or variants. To develop isomorphs, the incidentals or features that are unlikely to affect item difficulty are identified and manipulated (Irvine and Kyllonen, 2002; Gierl & Lai, 2012; Gierl & Lai, 2013a). However, when the goal of item generation is to develop variants, then radicals or elements of an item model that are likely to affect item difficulty can be altered (Irvine and Kyllonen, 2002; Gierl & Lai, 2012; Gierl & Lai, 2013a).

Implementing AIG in practice can be achieved in two ways – strong theory and weak theory (Drasgow et al., 2006). Some early examples of using strong theory to generate items in the domains of mental rotation (Bejar, 1993) and matrix completion items (Embretson, 1998). Bejar (1993) explored the feasibility of using response generative modeling (RGM) to the domain of mental rotation. Both studies explored the use of cognitive theory and processes underlying solving tasks, indicating a top-down approach towards item generation. They also found that the generated items were well supported by cognitive theory, which could also be used to control their difficulty. Other studies have further explored the top-down approach towards item generation – using

schema theory to generate mathematics items (Singley & Bennett, 2002), generating abstract reasoning items (Embretson, 2002), figural inductive reasoning items (Arendasy & Sommer, 2005), mental rotation items (Arendasy & Sommer, 2010), and developing item models for the broad content domain of medical surgery (Gierl & Lai, 2013b). These studies suggest that item models developed using strong theory yield instances with acceptable psychometric characteristics.

Weak theory on the other hand, relies on a parent item with known psychometric properties, wherein content experts identify and vary incidental features of items that can be manipulated to generate isomorphs (Drasgow et al., 2006). Several studies have used weak theory to generate items. LaDuca et al. (1986) described a complex method to generate test items for medical tests. They used existing items with good psychometric properties and identified changes to the stem and associated changes to the options. Bejar et al. (2003) used weak theory to develop item models for the quantitative section of the GRE. Since there was no underlying cognitive model to support item calibration, the item models were calibrated using the Expected Response Function (ERF) (Mislevy et al, 1994). A sample of variants were pretested, and the resulting information was used to calibrate the item model. Since the variants were not empirically calibrated, their item parameters remained unknown. The ERF was the approach used by the researchers to account for the uncertainty. They first conducted a simulation study to examine the impact of different levels of isomorphicity on score precision. The results of the simulation study suggested that an adaptive test based on item models would be psychometrically feasible provided there was only a moderate lack of isomorphicity. Thus, the researchers proceeded with an experimental on-the-fly adaptive test where items on the GRE quantitative section were exclusively developed by item models used in the simulation study. This was followed by an

experimental linear test also administered to the same sample who took the adaptive test. The previous GRE quantitative scores of the test taking population were compared to their scores on the experimental adaptive and linear tests. The correlation between the two sets of scores (experimental and operational) was found to be 0.87. Thus, despite using the ERF approach (attenuated item parameters) to calibrate the item models, the overall test appeared to perform similarly to the operational GRE quantitative test.

Other studies that used weak theory to implement AIG include Bejar and Yocom's (1991) work with hidden figure items and Lai et al.'s (2016) work with oral radiology. Both studies used parent items to generate item clones, rendering these items psychometrically isomorphic. For both domains (spatial ability and oral radiology), cognitive models did not exist at the time of their publication. Moreover, these studies suggest that weak theory can be beneficial for generating item clones for the purpose of repeatedly evaluating examinees on a specific topic.

Benefits of AIG. AIG has several benefits over traditional item generation methods and has demonstrated that it is a promising method of developing item families. Some of its benefits are discussed below.

Cost and efficiency. Perhaps the most well-known benefit of AIG is its ability to generate hundreds or thousands of items from various item models, thereby meeting the high demand for items and offering an economical alternative to traditional item development (Bejar et al., 2002; Graf et al., 2005). Studies such as Lai et al. (2009) and Choi et al. (2016) found that after training content experts in AIG, 64, 280 items across various subjects (math, science, literature, and social studies) and more than two million unique math items for grades 6 and 7 from 350 item templates could be generated, respectively. These studies implied that while item modeling efforts may

require more time than developing individual items, the time and costs may be offset by the large number of instances developed from them.

Until recently, there were few studies that discussed the costs associated with AIG. Kosh et al. (2019) found that when implementing AIG for the first time in a K-12 education setting, AIG is "more effective than traditional item writing if a testing program requires 247 or more items within a content area." (p. 51). This statement was made based on the assumption that one manually developed item costs $1 to produce. They also noted that while the initial cost of implementing AIG was high because the technological infrastructure was non-existent, subsequent implementations became more cost effective provided the need for items is substantial within one content area.

Intelligent calibration and efficient item analysis. Item models developed based on strong theory's use of cognitive models and/or theoretical underpinnings of test performance provides an underlying structure that adequately explains examinee test performance and facilitates prediction of psychometric properties of instances without having to field-test each one (Bejar, 1990; 1993; 1996; Embretson, 1998; 1999; 2002; Gierl & Lai, 2012; Gierl et al., 2012). Enright et al. (2002a) used item models to evaluate the impact of item design features on difficulty and discrimination. They found that feature variation accounted for 90% of the variance in difficulty and 50% of the variance in discrimination in rate word problems. Feature variation accounted for 61% of the variance in difficulty but did not explain the variance in discrimination of probability problems. Well-designed item models, that is, those based on a sound understanding of the item features that influence item discrimination and difficulty, can yield instances that are psychometrically similar. However, understanding which features to extract from items and contribute to their difficulty may require extensive research.

Maintaining test security. Item exposure, inevitable while using computer-based and computer-adaptive tests (CAT), leaves tests highly susceptible to content theft that threatens overall test security (Drasgow et al., 2006; Lai et al., 2009; Wainer, 2002). Moreover, test security breaches are known to take place when large-scale assessments reuse test items (Wainer, 2000 as cited in Lai, et al., 2009). Using AIG to develop item models that generate both isomorphs and variants can enlarge and replenish item banks on which CAT is dependent. For example, Gitomer and Bennett (2003) found that using AIG to produce variants of the same item meant that similar items could be used from the item bank, decreasing item exposure altogether.

Diagnostic and formative assessments. Based on previous research, ideally, a cognitive model of task performance is needed to support item model development. Cognition is also a main component of Cognitive Diagnostic Assessment (CDA), identifying the attributes to be measured (de la Torre & Minchen, 2014). Research and theory surrounding AIG is largely based on the Item Response Theory (IRT) framework, focusing on unidimensional assessments with limited diagnostic and formative information (Bai, 2019). Futhermore, Bai (2019) argues that since CDA and AIG overlap in the first stage (developing a cognitive model), perhaps this overlap can substantially reduce the difficulty of implementing AIG in formative assessments. For example, Choi et al. (2018) conducted an empirical study, evaluating the usefulness of administering AIG math items via a formative assessment system to Korean high school students. Students could instantly view feedback for incorrectly scored items, which provided explanations for solving the item correctly. Moreover, once they understood the feedback, they could take several more practice problems that were similar yet different to the question they got wrong to further improve their skills. The results found that 45.6% of students were satisfied with the iterative process using item clones for practice.

The literature on AIG highlights its ability to mass produce items and over time, reduce the high costs and inefficiencies associated with traditional item development. However, it also introduces the problem of calibrating these items, possibly amplifying the issues involved in item analysis described in the previous section. This is likely to be the case if traditional item analyses are used to calibrate the AIG items. However, procedures for calibrating families of items have been proposed and rely on the premise that items within families are statistically and structurally exchangeable or isomorphic in nature. For AIG to be truly successful at overcoming the limitations of traditional item development, it needs to be successful in mass production and calibration of items to avoid pre-testing bottlenecks. Isomorphic instances can help in reducing the size of the item pool that needs to be pretested, thereby saving resources that would have otherwise been engaged in pretesting each item. However, the extent to which item families are isomorphic rests completely on how the item model is designed. When item models are based on either cognitive theory or task models, their design reflects both content *and* complexity features, thereby making isomorphicity a greater possibility.

Cognitive Design with Task Model Specification Features

Creating isomorphic item families requires a granular understanding of their cognitive and/or content complexity. The literature on Principled Assessment Design (PAD) offers frameworks that can be used to develop item families that are driven exclusively by cognitive specifications and theory. Some of the PAD approaches are cognitive design systems (Embretson, 1998), evidence-centered design (ECD; Mislevy, et al., 2003), and assessment engineering (AE; Luecht, 2006, 2007, 2013).

Item families designed by cognitive theory can follow Embretson's (1998) cognitive design system. This approach involves developing cognitive models for each type of item, specifically

identifying processes for solving it as well as stimulus features present in the item that could impact its difficulty. Thus, difficulty of an item is decomposed into the cognitive processes responsible for solving an item, the stimulus features of the item, and the relationship between the two. The item stimulus features can also be manipulated to vary their relationship with the response processes, thereby impacting the difficulty of the item. As a result, difficulty and sources of cognitive complexity are identified at the item level. Moreover, Embretson (1998, 1999) argued that such manipulations affect construct representation (Embretson, 1983), one of the key elements for formulating a construct validity argument. Therefore, the cognitive design approach has several advantages over traditional item development efforts. Using cognitive models, the psychometric properties of items, namely their difficulty, can be predicted reasonably well – comparable to a multiple correlation of 0.70 (Embretson, 1999). Incorporating cognitive operations into psychometric models ensures that there is adequate knowledge concerning the underlying process of how items are solved, which when controlled by an item generation algorithm can produce items with predicted psychometric characteristics – a generative approach to psychometric modeling (Bejar, 1990; Bejar, 1993). Adopting the cognitive system of item generation therefore implies that difficulty can be controlled and used to produce isomorphic item families based on specific combinations of item features.

Other applications of Embretson's (1998) cognitive design system include Gorin and Embretson (2006)'s study, which examined the relationship between the cognitive features of reading comprehension items and item difficulty. They found that the two issues that could affect the psychometric properties are the relative proportion of variable elements in the item structures and the differences between the substituted elements (Embretson, 2002). They also state that one

of the benefits of the cognitive design system approach is that construct validity is assessed at the item level.

Embretson and Daniel (2008) modified earlier cognitive models on mathematical problem solving, proposing five stages of cognitive processing – encoding, integration, solution planning, solution execution, and decision. They identified 12 variables across the five stages that affected cognitive complexity. They used both the Linear Logistic Test Model (LLTM; Fischer ,1973) and regression models to investigate the impact of quantifying cognitive complexity in mathematical solving problems on item difficulty. The results suggested that two out of the 12 variables had a statistically significant impact on item difficulty using regression analysis, while the LLTM showed eight out of the 12 variables had a statistically significant impact on difficulty. Moreover, standard errors for the model coefficients were much smaller for the LLTM than the regression model coefficients.

Daniel and Embretson (2010) further investigated the impact of the cognitive variables identified in Embretson and Daniel (2008), specifically, source of equation and number of subgoals on item difficulty by intentionally controlling for these features in item design. The source of the equation and number of subgoals correspond to the encoding and solution planning stages of their earlier study. Thus, items without the presentation of an equation are significantly more cognitively complex than those with the equations. Together, the work of Gorin and Embretson (2006), Embretson and Daniel (2008), and Daniel and Embretson (2010) show that the cognitive design system can be used to automatically generate items and predict their difficulty to some extent.

Although not a direct application of the cognitive design system, Bejar and Yocom‘s (1991) study preceded Embretson’s (1998) work and suggested a generative approach to psychometric modeling. The generative approach proposed in their study encoded the cognitive

structures involved in solving hidden figure items into an item generation algorithm to produce isomorphs. A Hugh transform pattern recognition algorithm was used as templates to generate item clones. Since the objective of the study was to use a modeling approach that would represent both response consistency and response difficulty, the templates were able to explain both reasonably well. Moreover, the researchers advocated this approach of generating isomorphs as a method of continuous validation by testing the knowledge of response processes during each administration of the test. Embretson's and Bejar's work, along with supporting studies by Gorin (2005) and Embretson and Kingston (2018), demonstrate that item generation can rapidly mass produce item families while simultaneously anticipating their psychometric properties.

Item family design based on *task models* typically stem from evidence-centered design (ECD; Mislevy et al., 2003) or assessment engineering (AE; 2006a, 2006b, 2012a). Luecht and Burke (2020) state that, "the term "task model" is generally associated with evidence-centered design (ECD) to conceptualize the tasks that elicit evidence in support of proficiency claims (Mislevy et al., 2003; Mislevy, 2006; Mislevy & Riconscente, 2006; Mislevy & Haertel, 2007)." (p. 11). ECD proposes an evidence-based framework that supports a structured and systematic method of developing assessment tasks. Mislevy et al. (2003) defined ECD as:

A set of activities and artifacts that facilitate explicit thinking about (a) given the purpose of the assessment, what content and skills are both useful and interesting to claim about examinees; (b) what is the reasonable and observable evidence in student work or performance required to support the claims; and (c) how tasks (items) can be developed within the constraints of the assessment to provide students with an optimal opportunity to provide the observable evidence that is consistent with their achieving the intended claim. (pp. 314–315).

Huff et al.'s (2012) study developed task models to automatically generate items using the ECD framework. It serves as a prominent example of how task models can be used to generate isomorphic item families. The researchers designed task models for an Advanced Placement (AP) science course. In their study, they used the ECD-based task model to address two salient requirements for developing item templates: 1) ensuring that the templates had sufficient information for an item generation algorithm to generate items; and 2) explaining the factors that affect difficulty well enough such that resulting item instances operate in psychometrically similar ways. They state that the task models used in the study provided specific requirements extracted from broader proficiency claims and evidence, thereby providing the foundation for item templates to operationalize them. Thus, task models and item templates operate along a continuum, moving from general claims to more specific items. Moreover, when task models target a particular proficiency claim, they consider how students' knowledge progress along a proficiency continuum and how task features affect its cognitive complexity. Such an evidence-based development of item templates involves translating task complexity into features of the item template that remained fixed or variable, along with any necessary constraints on the range of each feature. Although not explicitly based on cognitive processing or theory, ECD provides evidentiary support for explaining and controlling difficulty of resulting items through task models and item templates. Finally, Huff et al. (2012) developed less challenging, moderately challenging, and more challenging item families from four templates, resulting in a total of 1,787 items. While these items were not administered or evaluated for their psychometric characteristics, the researchers argue that their approach enhanced the generation of high-quality items. They also encourage quality control procedures and strict evaluation methods for task models and item templates to throughout the design process to further improve their quality.

The second method of developing item families using task models is Assessment Engineering (Luecht, 2008, 2009, 2012a). The concept of a task model under the AE framework focuses on "a structured specification of the cognitive complexity undergirding an entire family of items." (Luecht & Burke, 2020, p. 33). AE and ECD operate in similar ways. As Luecht (2012a) states, "AE shares many features and some terminology with other approaches to principled assessment design, including evidence-centered design (ECD; Mislevy, 2006; Mislevy et al., 2003; Mislevy & Riconscente, 2006)." (p. 59). Central to the AE system are task models and item templates. Task models are based on construct maps comprise of ordered, measurable tasks, along a scale of proficiency, from low to high. The ordering of such proficiency claims assumes that those on the high-end of the scale have mastered claims on the lower end of the scale (Luecht, 2007, 2008; Luecht et al., 2009). In other words, cognitive complexity is at the forefront of construct mapping and task model development. Task models are developed to explicitly specify the combination of declarative knowledge and procedural skills required to support the proficiency claims along the ordinal scale of the construct. Luecht (2007) described the rules for building task models as:

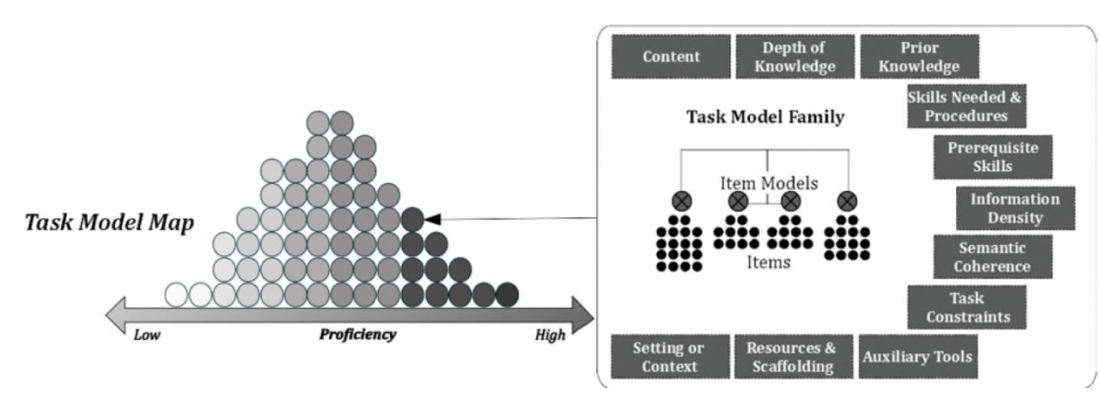
1. Task models should be incremental–that is, ordered by complexity.
2. Task models at the same level must reflect conjunctive performance.
3. Higher performance assumes that lower-level knowledge and skills have been successfully mastered. (p. 32)

When ordered by complexity, task models or task model families (containing between a few to several item templates) essentially differ in their location on the proficiency scale (Luecht, 2013a). This is because they target proficiency claims that are relatively more or less difficult to

other task model families on the scale. Figure 1 presents a task model map that illustrates how task model families are positioned along the continuum of proficiency.

The development of the task models, templates, and items can be targeted at different points of the difficulty scale proportional to the needed measurement precision (Masters, 2010). In terms of quality control, Masters and Luecht (2010) note that items produced using AE are engineered to maintain the template and/or task model location (difficulty), such that their psychometric characteristics reside within an acceptable degree of variation. They go on to say that if the templates are validated through pilot testing, and function along the intended scale of difficulty, or that the variation is within tolerable limits, then pre-testing can be relaxed. In other words, item families developed using AE-based task models are completely exchangeable or isomorphic (Luecht, 2007; Luecht et al., 2010).

Figure 1. A Task Model Map with Task Model Families, Item Models, and Items



Note. A Task Model Map with Task Model Families Comprising of Item Models and Items. From *Reconceptualizing Items: From Clones and Automatic Item Generation to Task Model Families,* by R.M. Luecht and M. J. Burke, 2020. Copyright 2020 by Richard Luecht. Reprinted with permission.

There are several examples of how task models have been used to develop item families in various contexts. Luecht et al. (2009) provided an empirical proof-of-concept of task modeling in a large-scale assessment context, using computerized performance exercises measuring accounting and auditing simulation problems. The computerized performance exercises were modified versions of accounting simulations on a national licensing exam and were reverse engineered to create task models. The exercises were reverse engineered with the help of subject matter experts, in such a manner as to express their difficulty using cognitive specifications. Complexity scores for each computerized performance exercise was determined based on the descriptions provided by the content experts and a task model coding schema. Content experts were also asked to provide ratings of difficulty for each exercise. Finally, the computerized performance exercises were also field-tested in professional accounting schools, where a partial credit model was used to analyze the response data. Correlations between these estimates and the complexity scores for the task models was 0.92, suggesting an ordered relationship between the AE task models and empirical difficulties of the computerized performance exercises. The results also demonstrate the benefits of using task models for operational use to develop a large number of items for the licensure exam. Task models can be used in place of traditional test blueprints to support operational item development with the controlled psychometric characteristics of items.

Luecht et al. (2010) developed task models for a multidimensional high school algebra test and a multidimensional high school reading comprehension test. Task models were developed for four diagnostic algebra constructs and four constructs for reading comprehension using the task model grammar defined earlier. Their study shows that using task models to replace traditional test blueprints can advocate for an "integrated test specification" that can allow items to be written to a specific target difficulty.

Masters (2010) described a method of using AE and item templates to develop sets of extended matching, true-false, and multiple-choice items with statistical characteristics that are suited to the needs of a licensure testing program. His study showed that item variants developed using AE item templates fit the Rasch calibration/scoring model as well, if not better than items developed in traditional ways; the item variants from the same template yield similar classical and IRT statistics. One key result of the study was a method to use AE to evaluate the performance of item writers over time.

Luecht (2012b) used AE task model and templates to demonstrate how item family hierarchical structures could be used for computer-adaptive testing (CAT). Luecht (2012b) states:

> From a CAT perspective, we need to adapt on the radical features by either selecting families having particular combinations of radical attributes or features (Geerlings et al., 2011) or actually manipulating on the fly those features that would provide the most precision at the current, provisional estimate of an examinee's proficiency score (Luecht, 2009; also see Bejar et al., 2003). What is apparent is that either conception of an adaptive test involves family-based hierarchy of units and fundamentally changes both the nature of item selection as well as the nature of the scoring outlined earlier. (pp. 206-207).

There were also two main findings from this CAT simulation study: 1) random variation within item families did not affect IRT scores of examinees as long as the items were randomly selected from item families; and 2) conditional covariances between the within family item difficulty and item discrimination parameters should be carefully examined to reduce any potential residual covariance between them.

Lai and Gierl (2012) developed AE task models using two cognitive models in mathematics and reading comprehension. Only one item template was created for each task model. They used

a three-step approach to create item templates: 1) applying relevant features of the task model within a given context; 2) applying each feature of the task model to the item template based on the context such that all constraints are met; and 3) ensuring that the item template is constrained to control non-cognitive effects on difficulty. Fifteen item templates were developed for mathematics and reading comprehension. Using their IGOR software, they found that the median number of items generated per template was 18, although IGOR generated a total of 10,301 items from the 15 templates. They found that mathematics item templates generated more items than reading comprehension item templates. Furthermore, they explained some of the benefits of using AE to generate items: 1) task models allow the generated items to be explicitly linked to the test construct; 2) the items generated from templates based on task models also allow different items to test the same set of cognitive features; and 3) AE-based item generation can be used to generate large item banks in a systematic manner.

Statistical Isomorphism within PAD. According to Luecht and Burke (2020), statistical isomorphism in the context of task model families means establishing a set of family-level parameters that can be used for scoring examinees. They further argue that statistical isomorphism only means that items within a family have similar but not identical psychometric characteristics. Thus, even though using family-level parameters for scoring purposes does ignore the within family variation in item parameters, the variation itself is inconsequential in terms of its impact on scores.

Procedures for calibrating item families are varied in their approach and complexity. For example, Bejar et al. (2003) used a three-parameter logistic model to compute the expected response function for an item family, averaging the item characteristic curve over all the item

instances for the item family. They found that under certain circumstances there was no bias, but measurement precision was reduced.

Sinharay et al. (2003) proposed three approaches for modeling item families in the context of the one-parameter logistic or Rasch IRT model. The first approach was the unrelated siblings model that assumed independent item response functions for all items within an item family. This model ignores item family membership. The model can be expressed as follows:

$$P\left(Y_{ij} = 1 \middle| \theta_i, b_j\right) = \frac{exp\left(\theta_i - b_j\right)}{1 + exp\left(\theta_i - b_j\right)} \quad (1)$$

where $P\left(Y_{ij} \middle| \theta_i, b_j\right)$ is the probability of success on item *j* for examinee *i,* θ_i is the latent trait for the examinee *i,* and b_j is the difficulty of item *j.* For this and subsequent models, the scale indeterminacy in the latent trait is addressed by setting the mean of the θs to 0.

The identical siblings model assumes that all items with the family have identical item response functions, and is expressed as:

$$P\left(Y_{ij} = 1 \middle| \theta_i, b_{jg}\right) = \frac{exp\left(\theta_i - b_{jg}\right)}{1 + exp\left(\theta_i - b_{jg}\right)} \quad (2)$$

where b_{jg} is the difficulty of item *j* that is a member of item family *g*, thus all b_{jg}s for item family *g* are assumed to be equal. This model therefore represents complete psychometric isomorphism among all the items within an item family. Sinharay et al. (2003) stated that the limitation of this model was that it failed to account for the variability among the items within a family, providing incorrect item parameter estimates. Fay et al. (2018) added that the identical siblings model could

not investigate psychometric isomorphism on its own but could be used together with the related siblings model to assess it in limited ways. They describe the related siblings model as one that does not assume that all items within a family have the same item response function, thereby allowing for some departure from true psychometric isomorphism. Instead, the related siblings model is described as having its first component from equation 2 while the second is a hierarchical model, that specified a distribution to relate the parameters for tasks from the same family:

$$b_{jg} \sim N\left(\mu_{b_g}, \sigma^2_{b_g}\right) \quad (3)$$

which specifies that each item-specific parameter b_{jg} is modeled as varying around a family-specific mean, μ_{b_g}.

Although these models can be used to calibrate automatically generated items, evaluating their psychometric isomorphicity requires additional procedures. Sinharay and Johnson (2005) advocated the use of Bayes factors (BFs) to compare the unrelated and related siblings models, which would possibly provide evidence in favor of the related siblings model over the identical siblings model, since it considers the family structure in the model. However, Fay et al. (2018) note that Bayes factors only indicated the presence of psychometric isomorphism versus a complete absence of it. Their study attempted to estimate the amount of possible psychometric isomorphism in items within a family using a set of statistical and graphical procedures. They go on to state several benefits of their approach: 1) requires only one model – the related siblings model; 2) allows for separate characterization of psychometric isomorphism for each of the item families; 3) can be instantiated differently depending on the number of items within an item family; and 4) uses graphical approaches using the IRT model as well as the classical test theory approaches to detect isomorphism in multiple choice items. They used BFs to provide evidence in

favor of invariant items within a family, where BFs of three or larger confirmed psychometric isomorphism while BFs of 0.33 or lower indicated non-invariance of items within a family. Their study found that BFs were useful for flagging items within a family that were not performing as expected. They also found that using multiple sources of evidence - BFs, item characteristic curves, and percent of examinees selecting each response option for an isomorphic item that was said to be invariant over test forms – was helpful in determining whether items within a family were truly psychometrically isomorphic.

Embretson's (1999) generalization of the linear logistic test model (LLTM; Fischer, 1973) serves as an alternative to the previously discussed item family calibration methods by using linear combinations of item features or PAD specifications to predict family-level item parameters (Luecht & Burke, 2020). Thus, the item family discrimination and difficulty parameters can be expressed as weighted combinations of the item features, where the weights are specified by designers of the item models or through empirical research (Luecht & Burke, 2020). Moreover, statistical isomorphism via item modeling efforts resurrected the idea of being able to predict item characteristics, especially item difficulty, in a systematic manner, since previous studies had directed difficulty prediction of traditionally developed items (Bejar, 1990, 1993; Drum et al., 1981; Embretson & Wetzel, 1987; Enright et al., 2002a; Mitchell, 1983; Freedle & Kostin, 1993; Kostin, 2004).

More complex approaches towards family-level calibrations involve hierarchical IRT models. Glas and van der Linden (2003) developed a hierarchical item response model to deal with differences between the distributions of item parameters of families of item clones. They demonstrated that task models and/or templates can be calibrated instead of individual items, using a hierarchical Bayes framework. In other words, there was one set of parameters estimated for an

entire family of items from a task model or template. The advantages of such an approach are: 1) less pretesting; 2) robust parameter estimation; and 3) misfit is minimized if the families are well formed (Luecht, 2007).

Geerlings et al.'s (2011) study also applied a hierarchical IRT model for item families generated through different combinations of design rules. Items within families were said to differ based on their surface features (incidentals; Irvine, 2002). The researchers combined the item cloning model (ICM; see Glas & van der Linden, 2001, 2003) and the LLTM for "the expected value of the item difficulty parameters for each family." Using a data-augmented Gibbs sampler, the researchers estimated the parameters of the model based on a Bayesian framework. The researchers fit three linear item cloning models (LICM) to test hypotheses based on previously validated theory concerning item difficulty for non-verbal intelligence tests. The first model was a baseline model, containing a different dummy rule for each of the 11 item families created. The second and third models were constructed based on two of the design rules supported by theory. The three models were compared using the deviance information criterion (DIC). Results showed that easier items had larger item discrimination parameters than difficult items. The study indicates that the LICM can be applied to automatically generated item families, increasing the cost-effectiveness of item generation.

Quality Control Procedures for Task Model and Item Families. Evaluating the statistical isomorphism of task models under the AE framework is imperative as it ensures that they are functioning as intended. Focusing on task models also means that any random or systematic variation occurring within them can be controlled by adjusting the item models (Luecht, 2012a). Luecht and Burke (2020) argue that statistical isomorphism can be evaluated in three ways: 1) deviations of task model family characteristics; 2) comparing the expected response functions

of family-level characteristics with item level characteristics; and 3) person fit across item families. They explain that deviations of task model family characteristics involve examining two types of residuals based on the expected response function of the task model family, given by:

$$ERF(\theta;\xi) = \pi_f = \sum_{c=1}^{C} = U_c P_{fc}(\theta) \quad (4)$$

where C denotes score categories, f denotes the task model family, U represents the item response score, and P is the probability.

The first method is to examine the deviations between the ERF and the scored item responses (Luecht & Burke, 2020):

$$\varepsilon_u = u_{i:f} - \pi_f \quad (5)$$

where $u_{i:f}$ is an observed binary item score for an item in task-model family.

The second method is to compare the deviations between the task model ERF and the individual ERF for items within a family (Luecht & Burke, 2020):

$$\varepsilon_\pi = \pi_{i:f} - \pi_f \quad (6)$$

Luecht and Burke (2020) also suggest that evaluating residuals in terms of unconditional or conditional means, variances, and covariances can be useful in different ways. They state:

> Unconditional aggregations of the residuals can be useful for general flagging purposes relative to empirically based thresholds, tolerances, or expected values. For

example, many IRT-based item fit statistics compare the sum of squared residuals to an expected variance (Wright & Masters, 1982; Lord, 1980). Conditional statistics can further identify item model or item design flaws relative to a proficiency level or for one or more examinee or mode-based grouping variables—much like differential item functioning (DIF) analyses. (p. 42).

Finally, they discuss the importance of evaluating statistical isomorphism within the IRT framework in terms of person fit – residuals that are aggregated by a grouping variable and possibly conditioned on score intervals along the ability scale. They go on to say that issues concerning person fit "signal some type of differential interaction between the items and the examinees in the focal and referential groups." (p. 42). Delving further into these results would mean re-designing task and item models that may be contributing to the problematic results. They advocate for the use of graphical presentations of fit statistics that can effectively visualize aberrance or residuals that may be larger than anticipated. When task and/or item models deviate from tolerable limits of statistical isomorphism, Luecht and Burke (2020) recommend eliminating those items from scoring, but more importantly, emphasize proactivity in terms of rectifying the cognitive specifications that drive the task models or the fixed/variable elements of the item models.

Conclusions from the Literature on AIG and PAD. The literature on AIG and PAD describes procedures for generating item families while simultaneously attempting to predict their psychometric characteristics. These procedures offer several benefits: 1) mass producing items to support the demand for large item banks; 2) reducing the sample size needed to calibrate stable item parameter estimates; 3) rapid assembly of test forms for low and high stakes testing; and 4) calibrating item families, using family-level parameters for scoring purposes. The literature also highlights the importance of developing and using analytical and investigative mechanisms to

ensure that task and/or item models are working as intended. However, these studies do not delve into exactly how much variation can be tolerated within item families before the task or cognitive models on which they were based are rendered ineffective.

Bejar et al. (2003), Shu et al. (2010), Luecht (2012b) have addressed the topic of variation within item and task model families, but ultimately left an important question answered: how much variation matters? Specifying the magnitude of variation within item families and examining their impact on scores provides important information. First, it can help set exact tolerances or limits of variation within item families, which can support task and item model development. For example, if a variation of 0.2 within item families has negligible impact on scores versus a variation of 0.5, then task model and item families can be designed to maintain a variation of 0.2 or lower. It would still require QC mechanisms to ensure that the items within the family do not exceed the 0.2 limit, but understanding how much variation can be ignored provides concrete guidelines for QC. Second, examining the differential quality of item families on scores can locate where along the score scale would the error be at its maximum. In other words, would a variation of 0.5 impact scores in the middle of the proficiency scale or at the tail ends or both?

Understanding exactly how much variation within item families impacts scores and to what extent scores along the scale are affected, helps researchers more accurately anticipate the consequences of such variation as opposed to merely hypothesizing them. Moreover, using modern approaches as a substitute for traditional methods of item and test development proves useful not only when they overcome their limitations, but also maintain, if not, improve score precision. Since the literature thus far has not adequately answered these questions, the present study aimed to address this gap by examining the impact of specific variation within item families on scores.

The next section deviates from item families altogether to examine on-going item generation and estimation of item difficulty without the need for pretesting. These studies utilize large language models and machine learning techniques to achieve this form of test development.

On-the-fly Generation of Items with Predicted Item Characteristics

While AIG and PAD focus on generating families of items, the third and final modern approach towards item development involves on-the-fly item generation, producing new items with an estimated difficulty that completely bypasses the need for pretesting. Obtaining the estimated difficulty of items can be achieved by using the Linear Logistic Test Model (LLTM; Fischer, 1973), where difficulty is decomposed into a linear combination of item features that are determined to impact it. Several studies have used the LLTM and other regression-based approaches to predict item difficulty in the areas of language testing, mathematics reasoning, and cognitive abilities. They are briefly discussed below.

A Brief Overview of Research in Item Difficulty Modeling (IDM). Perhaps one of the earliest applications of Fischer's (1973) LLTM was Embretson and Wetzel's (1987) study on paragraph comprehension. The researchers expanded the work of Drum, et al. (1981) by proposing the component latent trait model for paragraph comprehension multiple-choice items, drawing from Kintsch's (1998) construction-integration (CI) model. In general, the component latent trait model involves two stages – text representation and response decision - and underlying cognitive processes of encoding, coherence, and integration for one or both stages. Text representation involves comprehension of the text while response decision involves comparing the question stem and alternatives to the text itself. Encoding and coherence in the text representation stage involve converting visual stimuli into meaningful representations and integrating interpretation of the text with other facts and inferences, respectively (Embretson &Wetzel, 1987). The decision process

also includes encoding and coherence, which concern the questions associated with the text, specifically, making meaningful representations of the alternatives, mapping the alternatives back to the text, and evaluating the truth behind the alternatives (Embretson & Wetzel, 1987). The researchers investigated the effects of surface variables examined in Drum et al.'s (1981) study as well as the component latent trait model on the difficulty of 75 reading comprehension items in the Armed Services Vocational Aptitude Battery (ASVAB). The Linear Logistic Latent Trait Model (LLTM; Fischer, 1973) was used to analyze several cognitive models of item difficulty. The results suggested that the item feature predictors over passage feature predictors influenced item difficulty at the text representation stage.

Gorin (2005) applied the LLTM to experimentally manipulate reading comprehension items generated based on ETS's Graduate Record Examination's (GRE) verbal measure. The experimental conditions under which the items were generated and tested were driven by Embretson and Wetzel's (1987) work. More specifically, the experimental manipulations were based on passage propositional density and syntax modification; passage passive voice and negative wording modification; passage order of information change; and response alternative-passage overlap change. The results revealed that manipulation of some passage features increased item difficulty while others did not. For example, an increased use of negative wording increased item difficulty in some cases while altering the order of information presented did not affect difficulty. While experimental manipulations of several cognitive features did not yield significant results, this study proved that algorithmic changes in cognitive features can impact item difficulty.

Several other studies used regression methods to explain variance in item difficulty of language assessment items. Drum, et al. (1981) proposed a framework that included passage and item features that could predict reading comprehension item difficulty. They identified four

structural components of reading comprehension assessments – passage, question stem, correct answer, and distractors. Using several stepwise multiple regression analyses, the study found that plausibility of the distractors explained about 70 percent of the variance in item difficulty. Freedle and Kostin (1993) argued that 70 percent of the variance in item difficulty is misleading because of the apparent flaws in the stepwise regression analyses. They state that two or three predictors instead of the ten identified in Drum et al.'s (1981) study would have been sufficient based on the small sample of items (20-36) from which they were extracted. Drum et al. (1981) also highlighted differences in item difficulty for younger versus older readers. They found that predictors such as word recognition and word meaning in passages made items more difficult for younger readers than older readers.

Freedle and Kostin (1993) explored the role of text and text-related variables in predicting the difficulties of reading comprehension items from the Test of English as a Foreign Language (TOEFL). The items sampled for this study pertained to main idea, inference, and supporting statements. The purpose of the study was to examine whether twelve categories of variables identified based on experimental literature affected reading item difficulty. The twelve categories are negations, referentials, fronted structures, vocabulary, rhetorical organization, location of relevant information, lexical overlap, sentence and paragraph length, number of paragraphs, and abstractness of text. The researchers used 213 items from 20 test forms of the TOEFL. The item difficulty measure used was the equated delta based on a sample of 2000 examinees. Results from stepwise regression analyses found that lexical overlap, sentence length, paragraph length, rhetorical organizers, negations, referentials, and passage length accounted for nearly 58 percent of the variance in item difficulty.

Kostin (2004) investigated the effects of 49 variables - nested within three categories (word-level, sentence-level, and task-processing variables) - on the difficulty of TOEFL dialogue items. Dialogue items are essentially listening comprehension items wherein examinees listen to a recorded conversation between two people and answer multiple-choice questions regarding the conversation they heard. The item difficulty measure used in this study was the equated delta, wherein higher values are associated with more difficult items and lower values with easier items. Results of the multiple regression analyses revealed that the variables identified explained about 40 percent of variance in the item difficulty of the dialogue items.

Gorin (2011) study also used regression analysis to evaluate the contribution of nine item features towards complexity of 33 sentence-recall tasks. The predictive value of the nine item features had an R squared of 0.56 with a sample of 151 students. Adjusted R squared of 0.33 were found when Gorin regressed the nine item features on the effect sizes for each of the 33 sentence-recall tasks. Although this study was conducted a few years after the studies on paragraph comprehension items, when advances in artificial intelligence could support the quantification of text analysis, the R squared or explained variance in item difficulty remained quite low. This seemed to be the case even though sound cognitive models were used to identify item features that affect complexity. These results raise two questions: are the cognitive models flawed in some capacity? Or are linear regression approaches inappropriate for analyzing cognitive complexity of language?

Susanti et al. (2017) investigated factors that affect item difficulty of multiple-choice English vocabulary questions. Three factors – reading passage difficulty, semantic similarity between correct answers and distractors, and distractor word difficulty level – were examined for their contributions to item difficulty of the vocabulary questions. The researchers automatically

used various combinations of the three factors to automatically generate 120 vocabulary items modelled after items used on the TOEFL. These items were administered to English language learners, grouped by their standardized English proficiency score. A one-way ANOVA was used to test for significant differences between the mean difficulties of items classified according to the three factors. The results showed that distractor word difficulty level had the most significant impact on item difficulty.

Mathematics and Quantitative Reasoning. Other extensions of the LLTM have been applied to the domain of mathematical problem solving and reading comprehension. Embretson and Daniel (2008) highlight the advantages of the using LLTM - "elaborating construct validity at the item level, defining variables for test design, predicting parameters of new items, item banking by sources of complexity and providing a basis for item design and item generation." (p. 328). Even though the LLTM was meant to serve an explanatory purpose – linking cognitive features to item difficulty - it has not been fully leveraged to explore this further (Embretson, 1999; Embretson and Daniel, 2008). Thus, in their study, Embretson and Daniel (2008) modified earlier cognitive models on mathematical problem solving, proposing five stages of cognitive processing – encoding, integration, solution planning, solution execution, and decision. They identified 12 variables across the five stages that affected cognitive complexity. They used both LLTM and regression models to investigate the impact of quantifying cognitive complexity in mathematical solving problems on item difficulty. The results suggested that two out of the 12 variables had a statistically significant impact on item difficulty using regression analysis, while the LLTM showed eight out of the 12 variables had a statistically significant impact on difficulty. Moreover, standard errors for the model coefficients were much smaller for the LLTM than the regression model coefficients.

Daniel and Embretson (2010) further investigated the impact of the cognitive variables identified in Embretson and Daniel (2008), specifically, source of equation and number of subgoals on item difficulty by intentionally controlling for these features in item design. The source of the equation and number of subgoals correspond to the encoding and solution planning stages of their earlier study. They used the LLTM and 2PL constrained model, and a linear mixed modelling procedure to examine the effects of these cognitive variables on item difficulty. Results from all three models suggest that when equations are provided in the item, the solution planning stage is straightforward, causing items to be less cognitively complex. When the equation is not provided in the item, the solution planning stage becomes more cognitively complex, involving more steps to arrive at the solution. Thus, items without the presentation of an equation are significantly more cognitively complex than those with the equations.

Several studies were also conducted on IDM using different items from the quantitative reasoning section on the GRE. For example, Sebrechts et al. (1996) conducted detailed analyses of solutions to 20 algebra word problems administered on the GRE quantitative reasoning section. The problems were classified based on their attributes – problem complexity, need to apply algebraic concepts, and problem content. Regression analyses revealed that these attributed accounted for 37 to 62 percent of variance in item difficulty. Moreover, four solution strategies were identified – equation formulation, ratio setup, simulation, and other approaches. It was observed that students with high math ability used more equation strategies and fewer unsystematic approaches than students with low math ability. Deane et al. (2006) studied items that assessed linear equations and simple rational equations. They found that two-equation problems were more difficult than one-equation problems. Enright & Sheehan (2002b) investigated variables that could be attributed to item difficulty of arithmetic and algebra problems. They found that rate problems

increased in complexity when an additional constraint was introduced. Moreover, with regards to problems that required less than three steps to solve them, those that had mathematical keys were more difficult than those whose keys were a quantity.

Cognitive Abilities. Hornke and Habon (1986) developed classification system of cognitive operations for abstract reasoning items derived from cognition research on matrix items. The primary objective of this study was to use 14 cognitive operations in item development and subsequently evaluate their impact on Rasch item difficulty. Using this classification, 616 items were generated and presented in 35 smaller tests of 24 items each. A multiple regression model was used, where the Rasch difficulty estimate as the dependent variable and the item design features as independent variables. The classification system predicted the difficulties for the generated items quite well (R = .65).

Bejar (1990) explored generative approach of psychometric modeling of three-dimensional spatial rotation items by manipulating the angular disparity between two figures. Angular disparity was linked to psychometric difficulty in this study as previous literature had found that it impacted the item response time and correctness. A three-dimensional rotation test consisting of 80 items was constructed and administered to high school students. Both response and response-time data were collected. A dichotomous IRT model was modified to model response latency, thus, expressing the response as a proportion of the total time taken to respond to an item. The findings did confirm that angular disparity contributed to item difficulty and go on to suggest that the modified measurement model used is practically feasible. In other words, the approach of incorporating information of how examinees are likely to solve items into the measurement model widespread applications to psychometric modeling. Moreover, empirically verifying item features

that contribute to its difficulty implies that a generative approach towards item development can produce items with predicted psychometric characteristics.

Embretson (1998) used Carpenter et al.'s (1990) processing theory to determine item stimulus properties that could be manipulated to control the item difficulty of abstract reasoning items. Based on the processing theory, which states that generating and evaluating relationships across rows and columns of matrix problems, utilizes working memory and abstraction capacity to solve items. An increase in working memory capacity is needed to correctly answer an item when the numbers and levels of relationships also increase. Item structures were developed to account for the different combinations of the number and level of relationships. They were then used to generate individual test items using 22 objects and seven attributes. A proportion of .773 explained variance in item difficulty was attributed to the item structural model. As far as the cognitive processing theory was concerned, the study found that working memory capacity contributed significantly to item difficulty and response time, with a proportion of .71 explained variance in item difficulty.

In summary, the findings from the IDM literature vary in their extent to predict the difficulty of items, depending on the content domain. However, the studies that found higher proportion of variance explained in difficulty did not go on to evaluate the impact of the predicted difficulties on test scores. The very purpose of controlling difficulty is to ensure that items or item families operate in predictable ways. If this has been successfully established through experimental design and hypothesis testing, then these predicted item parameters can be used for scoring purposes.

Recent Developments in the Field of Computational Linguistics. With the latest advancements in machine learning and language models, the field of computational linguistics has

produced some research concerning on-fly-generation of items with estimated difficulties. Two of these studies, reflecting research at Duolingo, investigate the automatic generation of items and tests using a combination machine learning and NLP models. The first study, Settles et al. (2020) discussed in Chapter I, attempted to solve the "cold start" problem in language test development by using machine learning and natural language processing to automatically produce reading, writing, listening, and speaking tasks. They also used linguistic models to estimate the difficulty of these tasks.

The second study by McCarthy et al. (2021), developed a multi-task generalized linear model with Bidirectional Encoder Representations from Transformers (BERT) to "jump-start" item difficulty estimates of newly developed items for a high-stakes English proficiency test. Although Settles et al. (2020) described the automatic estimation of item difficulty from linguistic features of the CEFR difficulty scale to solve the cold-start problem, McCarthy et al. (2021) sought to use direct supervised learning of item difficulties from test takers. They combined IRT approaches with the work of Settles et al. (2020) to develop a single model that produces a priori estimates of difficulty to guide pilot testing and "jump-start" difficulty estimates of new items that are similar to existing items. The study used "BERT-derived passage embeddings to facilitate strong generalization to new test items." (p. 884). They also created an item-split dataset, where they randomly sampled 3% of items, and held out on sessions where at least one of those items were administered - approximately 20% of sessions. The remaining sessions were used for training. In the evaluation phase of the item-split experiments, they only evaluated results on the held-out items and not the entire sessions. The results of their study demonstrated that 3,000 pilot item administrations were adequate for good performance with a large item bank, whereas the 2PL-IRT model required 200 times as many administrations to achieve similar performance. They

also showed that the BERT Linear Logistic Test Model (LLTM) item parameter estimates generalize well even for new items that have not yet been piloted. Thus, their model demonstrated that it improves upon the performance of Settles et al. (2020) by incorporating test-taker response data and matches the 2PL-IRT to a large extent.

Overall Conclusions from the Literature Review

The objective of this literature review was to synthesize the work that has been done thus far in both traditional and modern methods of item and test development while highlighting gaps in this literature, some of which the present study intended to address. The research conducted on AIG, PAD, and on-the-fly generation of items and their estimates of psychometric characteristics, has attempted to address the limitations of traditional item and test development. These studies usher in a new way of thinking of how items and tests can be developed. While these studies went to great lengths to tackle the limitations of traditional item and test development, they were also accompanied by their own challenges – challenges unique to these approaches that must be addressed if they are to become the mainstream ways of designing tests for the 21st century. The literature also prioritized research questions that focus on what modern approaches are and how they work over when and how they impact scores. A possible reason for this is because there is no consensus amongst experts in the field with regards to the best method for developing and calibrating item families while verifying their quality. The calibration method of choice largely depends on the goal of calibration and the willingness to accept the trade-offs that accompany it.

Another possible reason for a lack of guidelines that support statistical QC of these item families is that more research is needed to support a set of procedures to achieve this. Traditional methods have seven decades of research to support their procedures whereas modern approaches have at best three decades of fragmented research that individually approaches these issues instead

of more collective approaches. The literature is also limited in their discussions on how the various item generation and family calibration methods impact scores beyond correlations with other tests or operational item characteristics. The research thus far has also approached item family generation and isomorphicity, calibration of item families, and prediction of their psychometric characteristics as separate but related topics. The works of Bejar et al. (2003), Embretson (1998, 1999, 2006, 2008, 2010, 2018), Luecht (2009, 2012a, 2012b, 2013), and Shu et al. (2010) attempted to address these issues together but leave the question of variation within item families largely unanswered. The fact that these studies even exist suggest that further investigation of the quality of item families and how predicted item parameters can affect scores is worth exploring. Thus, more research is needed to: 1) determine appropriate calibration strategies of item families in different contexts; 2) examine variation of items within families and determining acceptable tolerances of variation in item characteristics; 3) assess the impact of using different variations of isomorphic items on scoring; and 4) evaluate the impact of on-fly-estimates of item characteristics on scoring. The present two-part study intended to take a step forward in this direction in an attempt to address all four of these issues.

CHAPTER III: METHODS

This chapter provides an overview of the study design conditions, data generation, and simulation process for a two-part study that extends the previous work of Shu et al. (2010) and Bejar (1983). As discussed in Chapter I, there are three approaches to modern item development – reverse engineering a parent item using AIG, cognitive modelling of item instances, and generating items on-the-fly using machine learning and NLP methods. The methods outlined in Part A of this study address the first two approaches by examining the quality control and impact of variation within task model families on test scores. Part B discusses methods that address the third approach of modern item development (on-the-fly item generation using NLP methods) by examining the quality of item difficulty predictions and the effects of prediction errors on test scores. An overview of the methods in Parts A and B is provided in Table 1.

Table 1. Overview of Methods used in Part A and Part B

Variables/Evaluation Measures	Part A	Part B
Test length	25, 50, 75 items	25, 50, 75 items
Calibration Strategy (CS)	CS_1 – calibrating the task models CS_2 – calibrating the individual items	-
R^2	-	0.9, 0.8, 0.7, 0.6, 0.5
Evaluation Measures	Root Mean Square Deviation (RMSD) Bias Root Mean Square Error (RMSE) Conditional residual analysis	Bias Root Mean Square Error (RMSE) Conditional residual analysis including person fit

Part A

The first part of this study was based on Shu et al.'s (2010) study, which examined the role of QC and QA (quality-assurance) in a multidimensional formative assessment framework. The study was a large-scale simulation that involved multiple sources of potential variation for the multidimensional formative assessments of two test lengths – 8 and 20 items per trait, levels of item discrimination or subtest reliability, the quality control over items written to AE task model and template specifications (no variance within task model, moderate variance of templates within task models, and variance of both items and templates within task models, and the quality of test forms across seven proportional mixtures of items drawn from various item pools). The present study serves as a replication and extension of Shu et al.'s (2010) QC study by examining exactly how much variation matters within task model families without impacting scores.

This study considered a unidimensional summative achievement test that would be typically administered at the end of the school year, to assess mainstream students' performance on learning goals, objectives, and grade-level competencies as determined by states across the United States. Dichotomous responses were generated using the three-parameter IRT logistic model. The task models, templates, and items that were used in the simulations, were represented by various distributions of item discrimination parameters and item difficulty parameters. Test forms of three test lengths, 25, 50, and 75 items, were assembled based on a systematic selection of items generated under three variation conditions, low, moderate, and high, in various mixtures. For each test length, there were three pools of items that vary in terms of the quality of item parameters. Assembling test forms from different mixtures of these variation conditions allowed them to represent a comprehensive range of quality.

Item Pool Generation for the Simulations

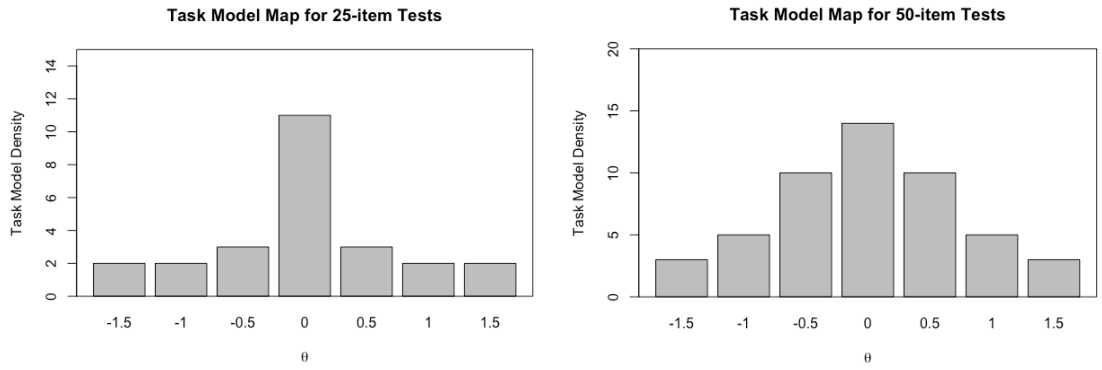
The item responses were simulated using the IRT three-parameter logistic model (3PLM) to generate dichotomous response data:

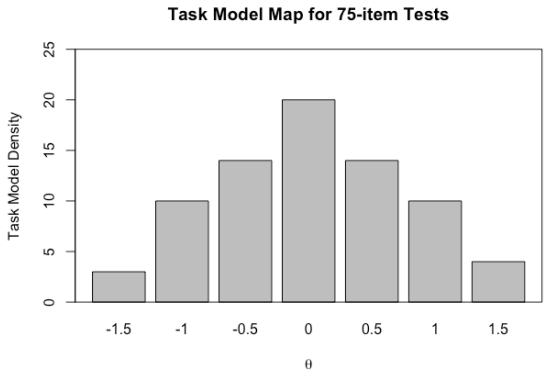
$$P(u_i = 1 | \theta_j, \xi_i) = c_i + \frac{1 - c_i}{1 + exp[-Da_i(\theta_j - b_i)]} \quad (7)$$

where a_i denotes the item discrimination, b_i is the item difficulty, c_i refers to the lower asymptote ($c_i = c = 0.15$), and θ is a proficiency score for j = 1,….., N examinees.

Three test lengths were considered in this study – 25, 50, and 75 items. The construct was measured by a well-defined cognitive test specification called a task model map (Luecht et al., 2010). The task model maps for the three tests are presented in Figure 2. Only one item was administered per task model (based on the test length). The test lengths were chosen to represent short, moderate, and long summative tests such that each would be long enough to provide acceptable validity evidence for the construct (Shu et al., 2010).

Figure 2. Task Model Maps for 25-, 50-, and 75-Item Tests





Item pools from three variation conditions were developed for each test length, representing low, moderate, and high variation, respectively. Thus, there were a total of nine item pools. Each pool consisted of $n \times 30$ items, where n is the number of task models (n = 25, 50, or 75 task models). The 30 replications of each task model consisted of three templates generated for each task model and ten item replications generated for each template. For the n = 25 task models, there were 750 items for each of the three variation conditions: i) 25 × 3 templates and ii) 10 replications or items from each template. Similarly, there were 1500 items for n = 50, and 2250 for n = 75 task models, respectively as presented in Table 2.

The differential variation within task model families was based on low, moderate, and large conditions used in previous literature (Bejar et al., 2003; Luecht, 2013c; Chen & Choi, 2023). However, the exact magnitude of variation for each condition was modeled after the standard errors of actual item parameter estimates from an end-of-grade (EOG) test, by varying the sample size for calibration. For example, when the 3PL model was used to calibrate a sample size of 300 examinees, the average standard error of estimates was found to be 0.5, thus providing the worst-case scenario. When the sample size increased to about 800 examinees, the average standard error of item parameter estimates was approximately 0.2, indicating relatively moderate variation compared to the large variation condition. Finally, a sample size of 3000 examinees or more

resulted in the average standard error of item parameter estimates being less than 0.1, forming the low variation condition. Varying the calibration sample size was carried out to reflect practical settings where obtaining large samples is not always possible, thereby causing the standard error of item parameters to fluctuate accordingly.

Since the AE framework involves a three-step item-family generation process, task models, item models or templates, and then the items themselves, the task models were generated first. The task models were sampled from target distributions of: i) item discriminations, $a_f \sim N(\mu_{af} = 1.3, \sigma_{af} = 0.3)$;; ii) item difficulties, $b_f \sim N(\mu_{bf} = 0, \sigma_{bf} = 1)$; and iii) the pseudo-guessing or lower asymptote parameters, $c_f = 0.15$. Each set of task model parameters were then used to generate three item models, using the same means as the task models, but with a standard deviation of 0.1, for both the item discriminations and item difficulties, thereby showing slight variation between the templates within a task model family. The item parameters for the 10 items generated from each of the three item templates were constrained to be equal to achieve the lowest possible variation within item families. Thus, the low variation condition contains item families generated from extremely well-constructed task models and templates, with minimal variation within them. This process was repeated for all three test lengths so that 25 (50, or 75) sets of task model parameters, 3 sets of item template parameters, and 10 items within each item template were generated.

The moderate variation condition generated task model parameters sampled from target distributions of: i) item discriminations, $a_f \sim N(\mu_{af} = 1.3, \sigma_{af} = 0.3)$;; ii) item difficulties, $b_f \sim N(\mu_{bf} = 0, \sigma_{bf} = 1)$; and iii) the pseudo-guessing or lower asymptote parameters, $c_f = 0.15$. Like the low variation condition, the task model parameters were used to generate the item model parameters, using a standard deviation of 0.2 for the item difficulty and item discrimination

parameters, introducing moderate variation in the item characteristics for each template associated with a given task model. This implies flaws in item template design perhaps in terms of cognitive skills or knowledge within a particular task model. The item template parameters were used to generate the item parameters for individual items within them, once again constraining them to be equal since the design flaws are restricted to the item template level for the three test lengths.

Finally, the high variation condition sampled from target distributions of: i) item discriminations, $a_f \sim N(\mu_{af} = 1.3, \sigma_{af} = 0.3)$;; ii) item difficulties, $b_f \sim N(\mu_{bf} = 0, \sigma_{bf} = 1)$; and iii) the pseudo-guessing or lower asymptote parameters, $c_f = 0.15$. This condition introduced variation in the item parameters for the templates and the items associated with each template. It also represented poorly constructed AE task models and templates, with very little quality control during item writing with respect to the item difficulty. For all three test lengths, the means of the task models and a standard deviation of 0.5 for the item difficulty and item discrimination parameters were used to generate the item templates. The resulting item template parameters were then used to generate significantly flawed item families with a variation of 0.5 in item difficulty and item discrimination distributions. Thus, the high variation condition represented the worst-case scenario – flawed task models, item templates, *and* items.

Table 2. General Item Pool Characteristics

Characteristics of Item Pools	Low Variation			Moderate Variation			High Variation		
Number of Task Models for each test	25	50	75	25	50	75	25	50	75
Number of Replications	30	30	30	30	30	30	30	30	30
Number of Templates	75	150	225	75	150	225	75	150	225
Number of Items	750	1500	2250	750	1500	2250	750	1500	2250
Within TM Variation in a parameters	0.10	0.10	0.10	0.20	0.20	0.20	0.50	0.50	0.50
Within TM Variation in b parameters	0.10	0.10	0.10	0.20	0.20	0.20	0.50	0.50	0.50
Within Template Variation in a parameters	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.50
Within Template Variation in b parameters	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.50

Test Form Assembly

There were seven different quality conditions representing different mixtures of items from drawn from the nine item pools of varying item quality. Table 3 presents a summary of the seven test-assembly conditions. Using the 25-item tests as an example, condition I required 100 percent of items to be selected from the low variation pool. This condition represents a best-case scenario wherein the item characteristics vary only in terms of intended task model difficulty. Conditions II required 75 percent of items to be selected from the low variation pool and 25 percent from the moderate variation pool. Conditions III to VI required a majority of items to be selected from the moderate and high-variation pools to assemble test forms representing sub-optimal conditions from an AE perspective. Finally, condition VII represented the worst condition, where task models and templates have been poorly designed from an AE perspective. There were 30 replications or test forms of 25, 50, and 75 items at each of the seven test assembly conditions. Thus, 210 25-item tests, 210 50-item tests, and 210 75-item tests from low, moderate, and high variation pools,

respectively. Furthermore, there were a total of six hundred and thirty test forms assembled and administered to examinees drawn randomly from a univariate normal distribution. Both the test form generation and assembly were developed using R-programming (R Core Team, 2022).

Table 3. Test Form Assembly Conditions

	Item Pools		
Condition	Low Variation	Moderate Variation	High Variation
I	100%	0	0
II	75%	25%	0
III	25%	75%	0
IV	0	100%	0
V	0	75%	25%
VI	0	25%	75%
VII	0	0	100%

Response Data Generation

The response data were generated by sampling the item parameters from a particular pool as described above to create a 25- or 50- or 75-item test form matching the task model specifications. The actual simulation of response data followed the routine procedures outlined in most unidimensional IRT studies. The item pools consisted of I items with parameters $\xi_i = (a_i, b_i, c_i, i=1,\ldots,I$. The item parameters were sampled from the different variation conditions as described above to create a 25- or 50- or 75-item test form matching the differential variation of task models. The target distributions from which the item parameters were sampled are: i) item discriminations, $a \sim N(\mu_a, \sigma_a)$; ii) item difficulties, $b \sim N(\mu_b, \sigma_b)$;); and iii) the pseudo-guessing or lower asymptote parameters, $c = 0.15$ (as defined in Shu et al., 2010), where N denotes the Gaussian normal distribution. These sample item pool parameters were used in equation 7 to generate dichotomous response data (0 = incorrect, 1 = correct) for the entire pool of items, given that the sample proficiency score $\theta \sim N((\mu_\theta = 0, \sigma_\theta = 1)$. Using (7), the "true score" for each item can

be obtained i.e., P_{ij}, using ξ_i and θ_j (i=1,…,I items and j=1,…,N examinees). Next, a uniform random real-valued number was generated within the interval $0 \leq \pi_{ij} \leq 1$. If $P_{ij} \geq \pi_{ij}$, the scored item response was set to $u_{ij} = 1$, otherwise $u_{ij} = 0$. Two data sets were generated for each test form. The sample calibration data sets had N = 1500 examinees completing each test form and one scoring data set was generated, N = 1000 response vectors for each test form. The calibration data set was used to estimate item parameter estimates under each of the two calibration strategies while the scoring data set was used to obtain estimates of θ_j using the item parameter estimates from the calibration data set.

Calibration Strategies (CS)

There were two calibration strategies that were used to calibrate each of the 630 datasets: i) CS_1 calibrating the 25, 50, and 75 individual task models; and ii) CS_2 calibrating the 750, 1500, or 2250 items individual items, ignoring the family of items per template. Only two calibration strategies were used in this study as opposed to the three in Shu et al.'s (2010) study, which included calibrations of the templates. Calibration of the templates were not considered because the results of their study found that calibration of templates and calibration of individual items performed equally well even for the worst test assembly conditions V, VI, and VII.

For the task model calibrations, the data consisted of 25, 50, or 75 response columns, one for each of the task models associated with the different test lengths. For the individual item calibrations, the datasets were calibrated separately, and the item-level statistics were used to score the simulated examinees.

Calibrations and Scoring

Calibrations using the strategies mentioned in the previous section was performed in flexMIRT 3.6.4 (Vector Psychometric Group, 2021) using the three-parameter logistic model (3 PLM) from Equation 7. Scoring was performed on the separate scoring data sets (where scoring depends on

previously calibrated IRT item statistics) using flexMIRT 3.6.4 (Vector Psychometric Group, 2021). While the flexMIRT software offers four scoring methods – maximum likelihood (ML) scoring, summed score to expected a posteriori (EAP) conversion table, Bayes expected a posteriori (EAP) score, and maximum a posteriori (MAP) scores, only the Bayes EAP method and its associated standard error was used in this study, using the 3 PLM item parameter estimates obtained from the calibration data.

Data Analysis

The focus of this study is to establish how much variation matters within task model families without impacting scores. Thus, this study used the indicators described in subsequent sections as quality control statistics within the context of AE.

QC Indicators using Variation of IRT Statistics

The variation of item statistics within task model families, across the seven test assembly conditions, and the two calibration strategies warrants an evaluation of item fit under these different conditions. This can be achieved by using the root mean square standard deviation (RMSD):

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{\xi}_{lh} - \hat{\xi}_{lh'}\right)^2}{n}} \quad (8)$$

where $\hat{\xi}_l$ refers to the IRT $a_i, b_i, and\ c_i$ parameter estimates from flexMIRT and h and h' denote the calibration strategies, CS_1 and CS_2.

QC Indicators for Assessing Impact of Item Quality on Examinee Scores

The calibrated three-parameter IRT estimates was used to score each of the 1000 examinees and the resulting score estimates was compared to the generated proficiencies, θ_j. The response vectors for these 1000 examinees were generated for each of the thirty test forms under the

different simulation conditions, and subsequently scored using the item parameter estimates from the two calibration strategies CS_1 and CS_2 for the same set of conditions. Thus, each simulated examinee was scored twice, using the two calibration strategies, having two sets of scores: i) the first set of scores was based on the task-model calibrated item statistics; and ii) the second set of scores was based on the individual item statistics. The individual residual error estimate for each simulated examinee, $j = 1,\ldots, N$, on a given test can be expressed as:

$$e_j = \hat{\theta}_j - \theta_j \tag{9}$$

where $\widehat{\theta_j}$ denotes the EAP estimate using the 3 PL model with item parameter estimates from each of the two calibration strategies i.e., calibration estimates from the task models and individual items. The residuals indicate the measurement of error, where the expected value of the residuals should ideally be 0. If the expected value of the residuals is non-zero, then $\widehat{\theta_j}$ will contain bias, expressed as:

$$bias = \frac{1}{N}\sum_{j=1}^{N} e_j \tag{10}$$

Negative and positive bias indicates an underestimate or overestimation of θ_j (the true proficiency score), respectively. Another QC indicator associated with residual error is the Root Mean Square Error (RMSE), defined in this study as:

$$RMSE = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{e}_{\mathrm{j}}^2} \tag{11}$$

RMSE is the square root of the mean square error or in the case of this study, the average standard error of the estimate, given that θ_j are known. The RMSE was used to compare the true

theta to the estimated thetas under the different study conditions, with smaller values being desirable.

Finally, a conditional residual analysis (Luecht, 2016) was conducted to compare the two calibration strategies, using the conditional standard error of measurement (CSEM), which are proportional to the test information function (TIF):

$$\sigma\left(\hat{\theta}_X \middle| \theta_X\right) = CSEM\left(\hat{\theta} \middle| \theta\right) = [TIF(\theta)]^{-}1 \tag{12}$$

Based on Equation 12, it then makes sense to normalize the differences in calibration strategies 1 and 2 (CS_1 and CS_2) relative to the pooled CSEM. The conditional standardized difference function or the normalized difference becomes:

$$\delta_X = \frac{\mu\left(\hat{\theta}_{X,CS_1}\right) - \mu\left(\hat{\theta}_{X,CS_2}\right)}{\sqrt{\sigma^{-2}\left(\hat{\theta}_{X,CS_1} \middle| \theta_{X,CS_1}\right) + \sigma^{-2}\left(\hat{\theta}_{X,CS_2} \middle| \theta_{X,CS_2}\right)}} \tag{13}$$

where the denominator is the pooled conditional standard error of estimate, averaged over the response patterns. The proficiency scores, θs are subscripted as CS_1 and CS_2 to represent the two calibration strategies that were used in this study. Luecht (2016) describes conditioning on the raw scores, *X*, as serving two purposes:

> First, it provides a readily interpretable score metric that can be understood by both psychometricians and non-psychometricians. Second, it is directly applicable to various types of raw-score-to-scale score look-up tables used by many testing programs. [Note that these procedures also readily extend to transformations of the estimated θ scores.] (p. 1)

The conditional normalized difference functions and plots of the normalized difference along the raw-score scale depict a clear picture of where along the raw-score scale the differences in the calibration strategies would occur. The plots present the normalized differences for the

range, $|\delta_X| \leq 0.2$ Luecht (2016) explains, "empirical experience has shown that a criterion absolute value of less than 0.2 typically precludes nontrivial score or classification differences." (p. 1).

The conditional residual analysis indicates the extent of the differences between the two calibration strategies in terms of their impact on scoring. [Note: also see Luecht & Ackerman, 2018 for an expanded discussion of residual analytics from simulation studies.]

PART B

The second part of this study addressed the claim made by Bejar (1983) that there needs to be an explained variance of 0.8 ($R^2 = 0.8$) in predicted difficulties for these estimates to substitute the empirical difficulties in scoring examinees. The Part B study examined how good the predicted difficulties need to be for IDM research to provide useful enough estimates to tackle examinee scoring and potential person (mis)fit. Different degrees of prediction in item difficulty parameters were generated at R^2 of 0.9, 0.8, 0.7, 0.6, and 0.5 and used to score response data from simulated examinees generated from a target distribution of the three-parameter logistic model parameters.

Item Parameter Generation for the Simulations

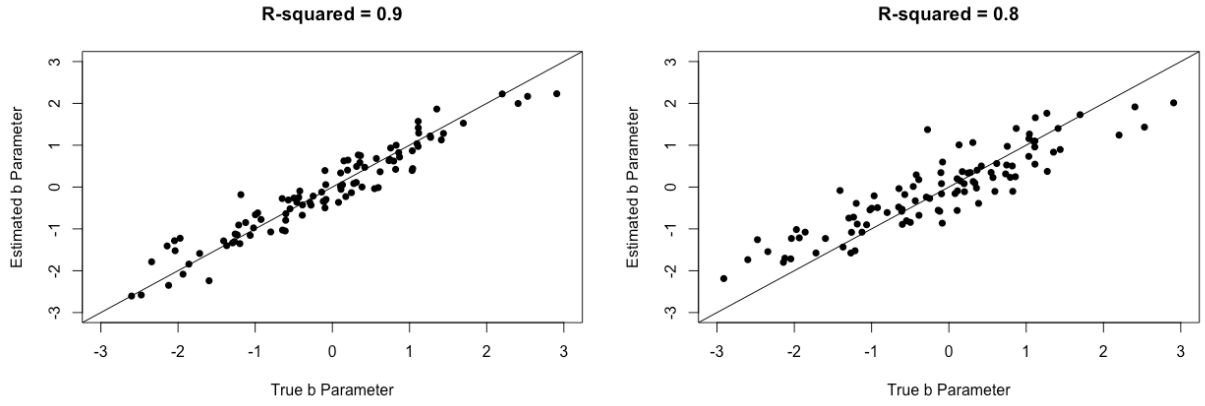
The item parameters generated for the simulations were based on the three-parameter logistic model from (7). Since the focus of this study was to assess the differential quality of predicted item difficulties on examinee scores, the "true" and predicted item difficulties, b and $\hat{b}$, were generated for different conditions of R^2. A correlation matrix was used to generate five variants of each item relative to a source set of $n = 200$ item difficulties to support test lengths of 25, 50, and 75 items. The correlations of the true item difficulties with their variants using the different R^2 (0.9, 0.8, 0.7, 0.6, and 0.5) are shown in Table 4. Scatterplots of the true and estimated item difficulties are presented in Figure 3.

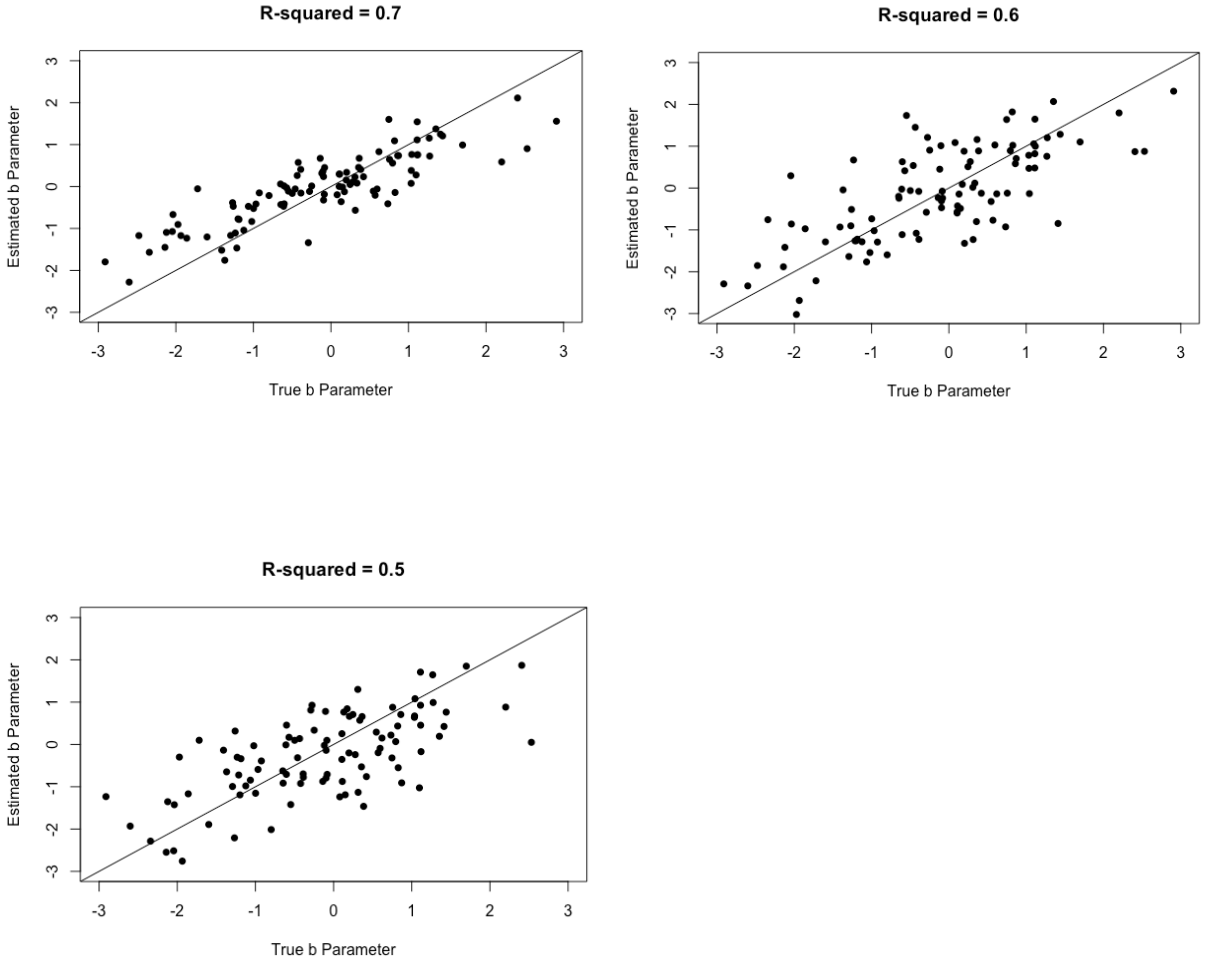
The item difficulties were sampled from a multivariate normal distribution with $\mu = [0,0,0,0,0,0]$ and the correlation matrix from the description above. Since the focus of this study was on the quality of the predicted item difficulty parameters, the item discrimination and lower asymptote parameters were set to 1.3 and 0.15, respectively, for each R^2 condition, thereby making it a modified one-parameter logistic model with a fixed pseudo-guessing parameter. Test forms of 25, 50, and 75 items were created to score 30 response datasets.

Table 4. Correlations of the True Item Difficulties with their Variants at R^2 Of 0.9, 0.8, 0.7, 0.6, And 0.5

1	0.948	0.894	0.837	0.775	0.707
0.948	1	0.847	0.793	0.735	0.670
0.894	0.847	1	0.748	0.693	0.688
0.837	0.793	0.748	1	0.649	0.591
0.775	0.735	0.693	0.649	1	0.548
0.707	0.670	0.688	0.591	0.548	1

Figure 3. Scatterplots of the True and Estimated Item Difficulties at Different Degrees of R^2





Response Data Generation

The true proficiency score, θ_j was sampled from a standard normal distribution, $\theta \sim N(0, 1)$. Using Equation 7, the true item difficulty parameter estimates (the first column of the six columns of multivariate data generated), item discrimination parameters randomly sampled from a normal distribution with low variance, $a \sim N(\mu_a = 1.3, \sigma_a = 0.05)$, $and c_i = c = 0.15$, the "true score" for each item can be obtained i.e., P_{ij}, for i=1,…,I items and j=1,…,N examinees. Next, a uniform random real-valued number was generated within the interval $0 \leq \pi_{ij} \leq 1$. If $P_{ij} \geq \pi_{ij}$, the scored item response was set to u_{ij} =1, otherwise u_{ij}= 0. This process as carried out for the three test lengths (25, 50, and 75 items) to generate 30 scoring data sets of N = 5000 response vectors.

Unlike Part A, this part of the study focused only on scoring the data with the predicted item parameter estimates and not comparing different calibration strategies. Thus, calibration data sets were not generated for Part B.

Scoring

Once the scoring data sets containing the item-level responses were generated, IRT scoring was performed in flexMIRT 3.6.4 (Vector Psychometric Group, 2021), to obtain the Bayes expected a posteriori (EAP) estimates along with their associated conditional standard errors (i.e. the standard deviations of the examinee-level posterior distributions) using the *predicted* item difficulty parameter estimates obtained under each of the item-difficulty prediction conditions.

Data Analysis

The objective of Part B was to examine the impact of the quality of the different degrees of predicted item parameter estimates on examinee scores and person fit. The QC indicators described in Part A relied on residuals calculated based on the EAP estimates of examinees. However, in this study, QC was evaluated in terms of residual-based fit statistics based on the difference between an examinee's observed item-score and predicted probability for endorsing an item for each R^2 condition. Thus, the individual residual error of estimate for a particular examinee, j = 1,…., N, on a particular test form was expressed as:

$$e_{ij} = u_{ij} - P(\theta_j, v_i) \tag{14}$$

where u_{ij} denotes an examinee's observed item score and $P(\theta_j, v_i)$ is the probability of an examinee endorsing an item based on the EAP estimates and predicted item parameters.

For each examinee, the mean, standard deviation, and RMSE were calculated across the items such that:

$$Mean_j = \frac{1}{N} \Sigma_{j=1}^{N} e_{ij} \quad (15)$$

$$SD_j = \sqrt{\Sigma \frac{(e_{ij} - \bar{e}_{ij})^2}{N-1}} \quad (16)$$

$$RMSE_j = \sqrt{\frac{1}{N} \Sigma_{j=1}^{N} e_{ij}^2} \quad (17)$$

Calculating the residuals using this approach addressed the possibility that IRT scores can be reasonably insulated from errors in item difficulty prediction. In other words, there may not be substantial differences in scores based on item parameter estimates that were of poor quality. However, differences may be noticeable in terms of fit.

The impact of the differential quality of predicted item difficulties on person fit was evaluated using the l_z index (Drasgow et al., 1985). The l_z or person fit index, derived from the log likelihood function, is computed for each examinee. It is then compared to some threshold, usually $|z| > z_\alpha$, where z_α = - 1.645(Magis et al., 2012). Using notations from van Krimpen-Stoop and Meijer (1999, as cited in Magis et al., 2012), the likelihood function for any response pattern for the 3 PL model described by Equation 7 is:

$$L(\theta) = \prod_{i=1}^{n} P_i(\theta)^{X_i} \; Q_i(\theta)^{1-X_i} \quad (18)$$

where $Q_i(\theta)^{1-X_i}$ = 1 - $P_i(\theta)$ or the probability of an incorrect response.

$$l_0 = \log L(\theta) = \sum_{i=1}^{n} \{X_i \log P_i(\theta) + (1 - X_i) \log Q_i(\theta)\} \quad (19)$$

Since l_0 is not standardized and its distribution depends on θ, Drasgow et al. (1985) proposed a standardized version of l_0:

$$l_z = \frac{l_0 - E(l_0)}{V(l_0)^{1/2}} \tag{20}$$

where $E(l_0)$ and $V(l_0)$ are the mean and variance of l_0, respectively (van Krimpen-Stoop & Meijer, 1999, as cited in Magis et al., 2012):

$$E(l_0) = \sum_{i=1}^{n} \{P_i(\theta)\ \log P_i(\theta) + \ Q_i(\theta)\ \log Q_i(\theta)\} \tag{21}$$

$$V(l_0) = \sum_{i=1}^{n} P_i(\theta)\ Q_i(\theta) \left(\log \frac{P_i(\theta)}{Q_i(\theta)}\right)^2 \tag{22}$$

Once l_z was calculated for each examinee, the $|z| > z_\alpha$ (z_α = - 1.645) was used to determine if the response pattern was a misfitting one. Means and standard deviations of the person fit statistics were reported each R^2 condition and test length. The distributions of the person fit statistics were plot using stacked histograms to compare differences, if any, across the R^2 conditions. All person fit analyses were conducted using the PerFit package in the R programming language (R Core Team, 2022; Tendeiro et al., 2016).

Review of Research Questions

The analyses described in this chapter were directed at answering the two research questions in this study.

1) How much variation matters in task model families without impacting examinee scores?

1a. Under which conditions would calibrating the task models and/or calibrating the individual items be most appropriate?

Plots of RMSD of the variation in the item parameter estimates for each test length were developed and patterns were examined across the seven test assembly conditions. In addition, plots of the means of the item difficulty estimates were developed, conditional on task model location, where b = (-1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5) for each test length and classified by calibration strategy. These plots intended to demonstrate whether the task model locations were (or were not) maintained under a particular test assembly condition.

1b. Which calibration strategy would have the most (or least) impact on examinee scores?

Bias statistics were grouped by the test assembly condition and the calibration strategy to examine whether increases occur at conditions III and higher for CS_1, where the quality control of the template and item variance progressively worsened. Such increases are likely to suggest the negative impact of poor-quality items on scoring. The results would also serve as an indicator suggesting that CS_1 would be inappropriate when there is a substandard quality of templates and items resulting from them.

Plots of the RMSE statistics for the three test lengths across the seven test assembly conditions and two calibration strategies were developed. RMSE values for proficiency score estimates using CS_1 item statistics would be expected to increase for conditions V-VII since these conditions required test forms to be assembled from a selection of moderate- to high-variation items. Should the plots confirm these results, it would suggest that using CS_1 may negatively impact scoring for these conditions. Using the conditional residual analysis to compute the normalized difference, plots were developed to depict whether the magnitude of differences between the two calibration strategies were large or

small using $|\delta_X| \leq 0.2$. If a majority of the normalized differences are large and negative, it would suggest that CS_1 results in lower scores than using CS_2. Thus, under such circumstances, using CS_2 may be more appropriate.

2. How do different degrees of explained variance in predicted item parameters impact scores and person fit?

The residual-based summary statistics may reveal larger errors, especially the RMSE statistics and standard deviation of the residuals as the correlations with the true item difficulty parameters decreases. The potential increase in error may be due to the poor estimation of the predicted probability of endorsing an item, i.e., $P(\theta_j, v_i)$. Moreover, the person fit analysis would suggest the proportion of misfit across the R^2 conditions. The stacked histograms presenting the distribution of the person fit statistics for each R^2 condition would show the extent of misfit as the correlations between true and predicted difficulties decreased.

Conclusion

This chapter detailed the methods used in this study to answer the research questions for parts A and B. The methods focus on some of the QC mechanisms that can be used to verify if task model families and predicted item difficulty parameters meet their design-specific assumptions. To date, little has been done to address the role of QC in modern item and test generation methods, specifically evaluating their impact on test scores. This study explored this by proposing some QC procedures to assess within-family variation and prediction error in terms of their impact on test scores.

CHAPTER IV: RESULTS

This chapter presents the results of the simulation study for Parts A and B. Each section is organized by re-stating the purpose and summarizing the results relative to each research question.

PART A

The purpose of Part A was twofold. First, Part A compared how the two calibration strategies performed across seven test-assembly conditions: (1) CS_1 collapsed the generated response data by item family, effectively calibrating a single set of item parameters for each task model family and (2) CS_2 treated the items as unique. (As a reminder, the within-family variation was introduced via the "true" item parameters. The resulting comparisons suggest what might happen when the process of creating items fails to adequately control the degree of statistical isomorphism within families.) From an estimation perspective, CS_1 used all the available item-response data per task model for calibration purposes producing more stable item-parameter estimates by ignoring the within-family variation. In contrast, CS_2 treated the items as unique to optimize data-model fit but at the cost of relatively less stable item parameter estimates since the sample size per item reduced. The second and, in some ways, more critical purpose of Part A was to evaluate the impact of each calibration strategy with respect to the test scores given differing magnitudes of simulated within-family variation in the item parameters.

Calibrating at the task-model level for test forms that were assembled from primarily the low variation pools would appear to be a preferred strategy when feasible since, per AE specifications, all items within each family should be isomorphic by design. In this case, any variation within task model families is expected to be well-controlled, inducing only minor amounts of additional random noise during the estimation process. That is, CS_1 should produce a

single set of highly stable item parameter estimates for each item family implying a more efficient calibration design. In addition, CS_1, if empirically supported as being effective, eliminates the need for pilot testing individual items. However, when the items used to generate test forms are sampled from item families with more varied statistical item characteristics, calibrating at the task model level could prove problematic. That is, if the variation matters and is simply ignored within families, the item parameter estimates could be unstable, lead to data-model misfit similar to extreme "item parameter drift", and ultimately lead to scoring accuracy issues.

The second purpose was to investigate the consequences of using task model calibrations for scoring purposes under different magnitudes of where within family variation was low to quite severe. The operating assumption under CS_1 is that the hierarchical structure of the task model family could be exploited regardless of the magnitude of within-family variation in the item parameters. And if scoring accuracy is impacted under the higher variation conditions, the obvious question becomes, "How much variation matters?". These issues and the results are discussed below. The relevant research question as follows.

Research Question #1. How much variation matters in task model families without impacting examinee scores? Corollary Question 1a. Under which conditions would calibrating the task models and/or calibrating the individual items be most appropriate?

Variation in IRT Statistics

As discussed in Chapter III, the variation of item statistics within task model families, across the seven test assembly conditions (item parameter variability within families), and the two calibration strategies warranted an evaluation of item fit under these different conditions. Two types of statistics can be used based on a score residual,

$$e_i = \hat{\rho}_i - \rho_i \tag{23}$$

where $\hat{\rho}_i$ is an estimated item parameter and ρ_i is the true [generated] item parameter. Bias is the simple average, $BIAS = n^{-1}\sum_i e_i$. The root mean standard deviation is $RMSD = \sqrt{n^{-1}\sum_i (e_i - \bar{e})^2}$.

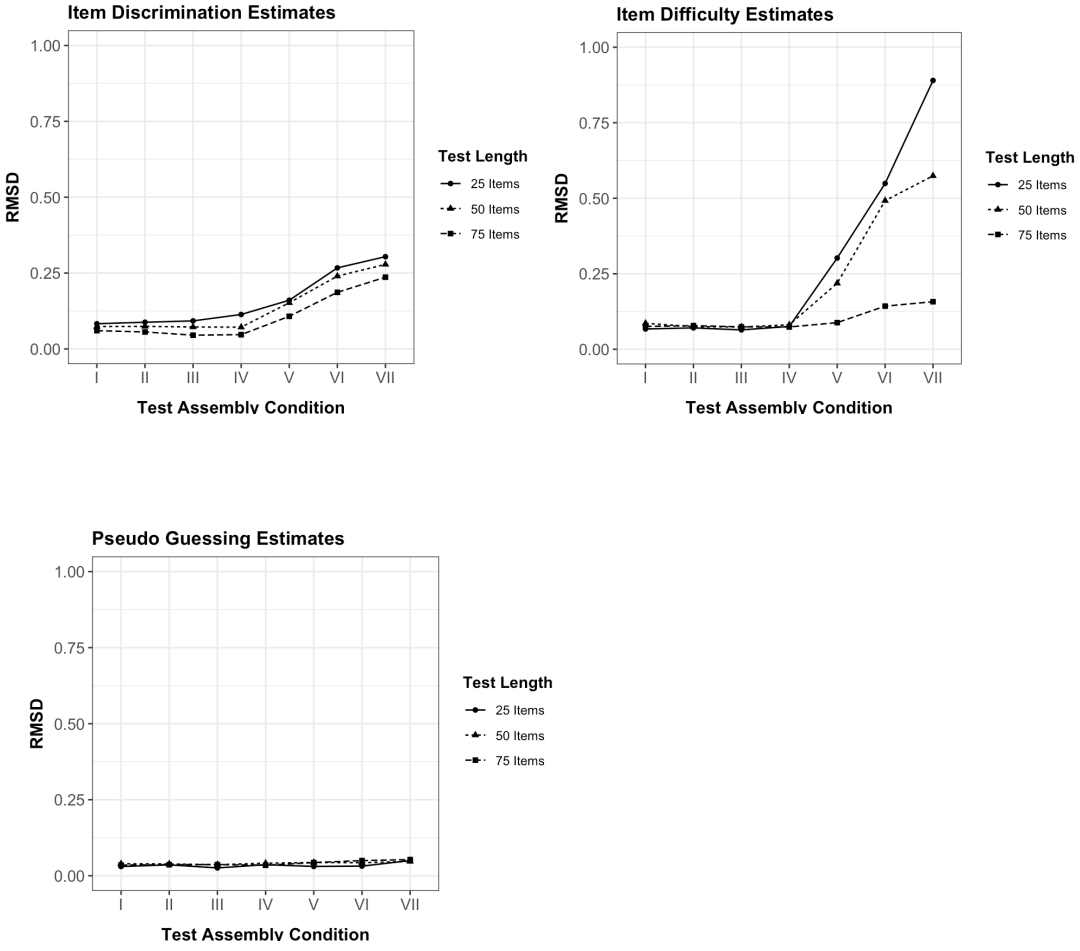
The RMSD statistics for each of the item parameter estimates to were used to compare CS_1 and CS_2 and determine which calibration strategy might yield the more appropriate item parameter estimates for test assembly purposes. RMSD patterns for CS_1 and CS_2 for the three test lengths are presented in Figure 4. The upward trajectory of the RMSD patterns for the item discrimination and item difficulty estimates across the seven test assembly conditions were expected as the variation within task model family increased from 0.1σ to 0.5σ. (Note that 3PL pseudo-guessing parameters were fixed to 0.15 during the item pool generation phase and therefore would not be expected to vary much across the test assembly conditions, calibration strategies, and test lengths.)

For condition I, where the within task model variation in difficulty and discrimination was low or at 0.1σ standard deviation, the RMSD ranged between 0.05-0.08, respectively. These estimates increased above 0.3 once variation within task model families and templates increased to 0.2σ (within-family standard deviation units) for conditions II, III, and IV. Conditions V, VI, and VII saw a sharp increase in RMSD, where the variation within task models, templates, and items was at 0.5σ, indicating that the item families were not functioning isomorphically as intended. This is especially true for the difficulty estimates for the 25- and 50-item tests.

Strictly considering these item parameter estimation results, calibrating task models when isomorphism is questionable seems ill-advised. This would especially be true when evaluating the psychometric properties of test forms (e.g., reliability or measurement information) or considering using those item parameters in IRT scale linking. Where empirical quality control procedures indicate large variation in the item operating characteristics for particular item families, the wise

strategy would be to implement CS_2 (i.e., calibrate the items in those families as unique rather than isomorphic instances from the family). Although the RMSD patterns continued to increase across the seven test assembly conditions for the three different test lengths, the magnitude of residual error appeared to decrease with increased test length.

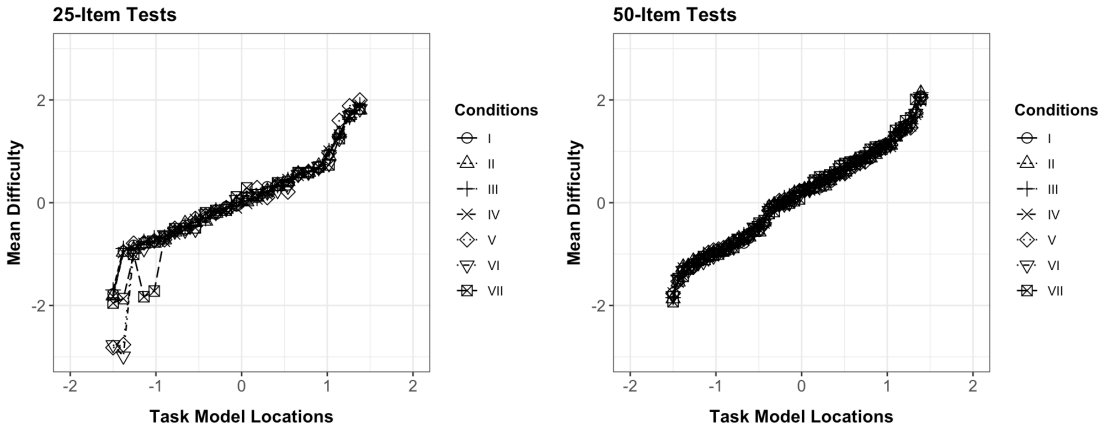
Figure 4. Variation in IRT Statistics across Calibration Strategies, Test Lengths, and Test Assembly Conditions



An important design assumption under the AE framework is that task models should be "located" on the proficiency scale as a function of their design-based cognitive complexity (task model difficulty). In the present context, task models were "designed" based on the task model

maps specified in Chapter III. In addition to comprising an isomorphic collection of items, each task model family should maintain its *intended* location on the task model map. That is, there should not be task model family drift in the parameter estimates. This assumption was verified by evaluating whether the mean difficulty conditioned on the task model location maintained the design-specified locations ($b_i = -1.5, -1.0, -0.5, 0, 0.5, 1.0, 1.5$). In other words, the task model difficulty was obtained by aggregating the difficulty parameters across the 30 replications for each task model for each test assembly condition and test length. These results are presented in Figures 5 and 6 for CS_1 and CS_2, respectively.

Figure 5. Mean Difficulty Estimates by Task Model Location for CS_1 across the Seven Test Assembly Conditions and Three Test Lengths



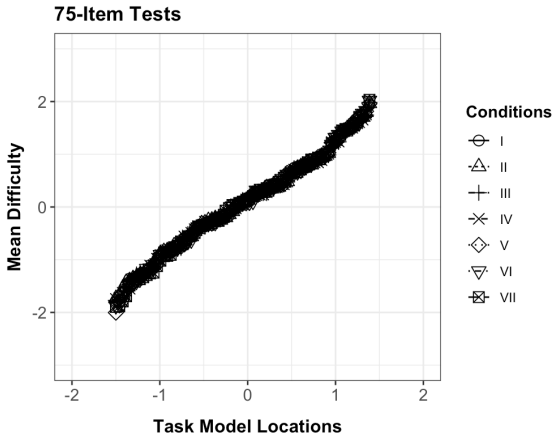
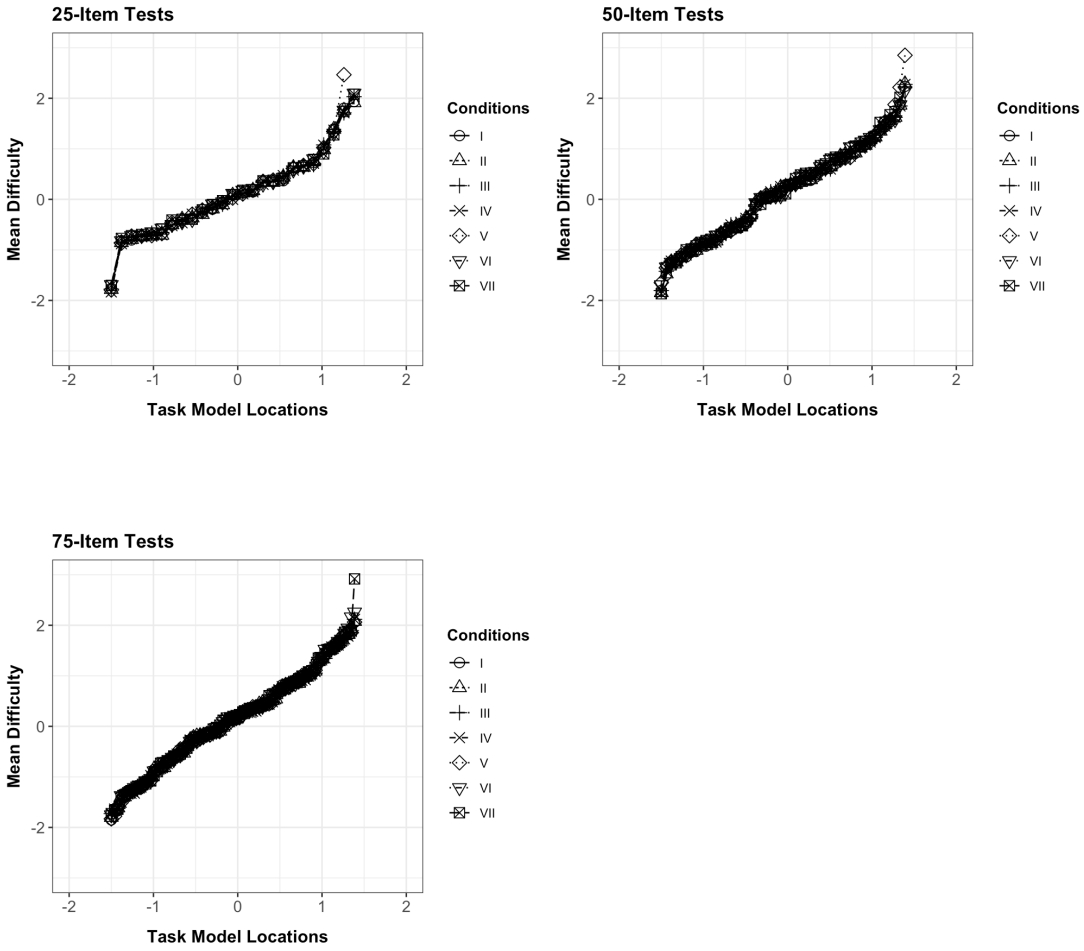


Figure 6. Mean Difficulty Estimates by Task Model Location for CS_2 across the Seven Test Assembly Conditions and Three Test Lengths



The plots in Figures 5 and 6 represent the mean difficulty estimates obtained from CS_1 and CS_2, respectively for the 25-, 50-, and 75-item tests. They illustrate that while the individual difficulty estimates may vary for a particular test assembly condition or calibration strategy, on average, they maintained their locations reasonably well. Thus, despite sampling from the three item pools using different mixtures of item parameter variation, the task model locations, on average, remained relatively unaffected.

Corollary Research Question 1b. Which calibration strategy would have the most (or least) impact on examinee scores?

Impact of Item Quality and Calibration Strategy on Examinee Scores

The item parameter estimates from CS_1 and CS_2 were used to score each simulated examinee that were generated as part of the scoring datasets (N = 1000). Thus, each simulated examinee was scored twice – using the task model family calibrated statistics, and the individual item statistics. These item parameter estimates were used for all 1000 examinees for each of the seven test assembly conditions and three test lengths. There were three types of QC statistics calculated for examinee scores – bias, RMSE, and the conditional standard error (CSEM) of normalized difference between CS_1 and CS_2.

Focusing on scores, the residual becomes

$$e_j = \hat{\theta}_j - \theta_j \ , \qquad (24)$$

where $\hat{\theta}_j$ is an expected a posteriori (EAP) proficiency score estimate for j=1,…,N examinees θ_j is the true [generated] proficiency score. Bias is the simple average, $BIAS = N^{-1}\sum_j e_j$. The more general root mean squared error was used here as $RMSE = \sqrt{N^{-1}\sum_j e_j^2}$, noting that the RMSE incorporates the squared bias.

The bias statistics for the seven test assembly conditions and three test lengths are presented in Table 5. They were further broken down by calibration strategy. Negative bias indicates that the true proficiency scores were underestimated, while positive bias indicates overestimation. Condition I, where the variation within the task model family was approximately 0.1 at the task model only, saw minimal bias between the two calibration strategies across the test lengths. This implies that when the variation is at 0.1σ or below, scoring examinees using the task model parameters may not be problematic since the task models, item models, and items adhered to the quality constraints in terms of their intended design. This finding also applies to condition II, where variation within the task model family increased to 0.2σ.

The bias progressively increases from condition III onwards, where scoring examinees using the task model calibration statistics, introduces relatively more error than conditions I and II. Conditions V-VII show more pronounced differences when using CS_1 parameter estimates for scoring purposes but do decrease with an increase in test length. Using item-level statistics to score examinees under these conditions seemed to significantly reduce the error between the true scores and EAPs. Even in these conditions, an alternative calibration method would be to use the item-model or template statistics to score examinees. For example, Shu et al. (2010) had found this hybrid method to be robust in providing accurate scores. Calibrating at the item-model (template) level would still bypass the need to pre-test every item and potentially reduce the high costs associated with item-level calibrations. It would also mean that less items are exposed during the pretesting phase, reducing threats to test security.

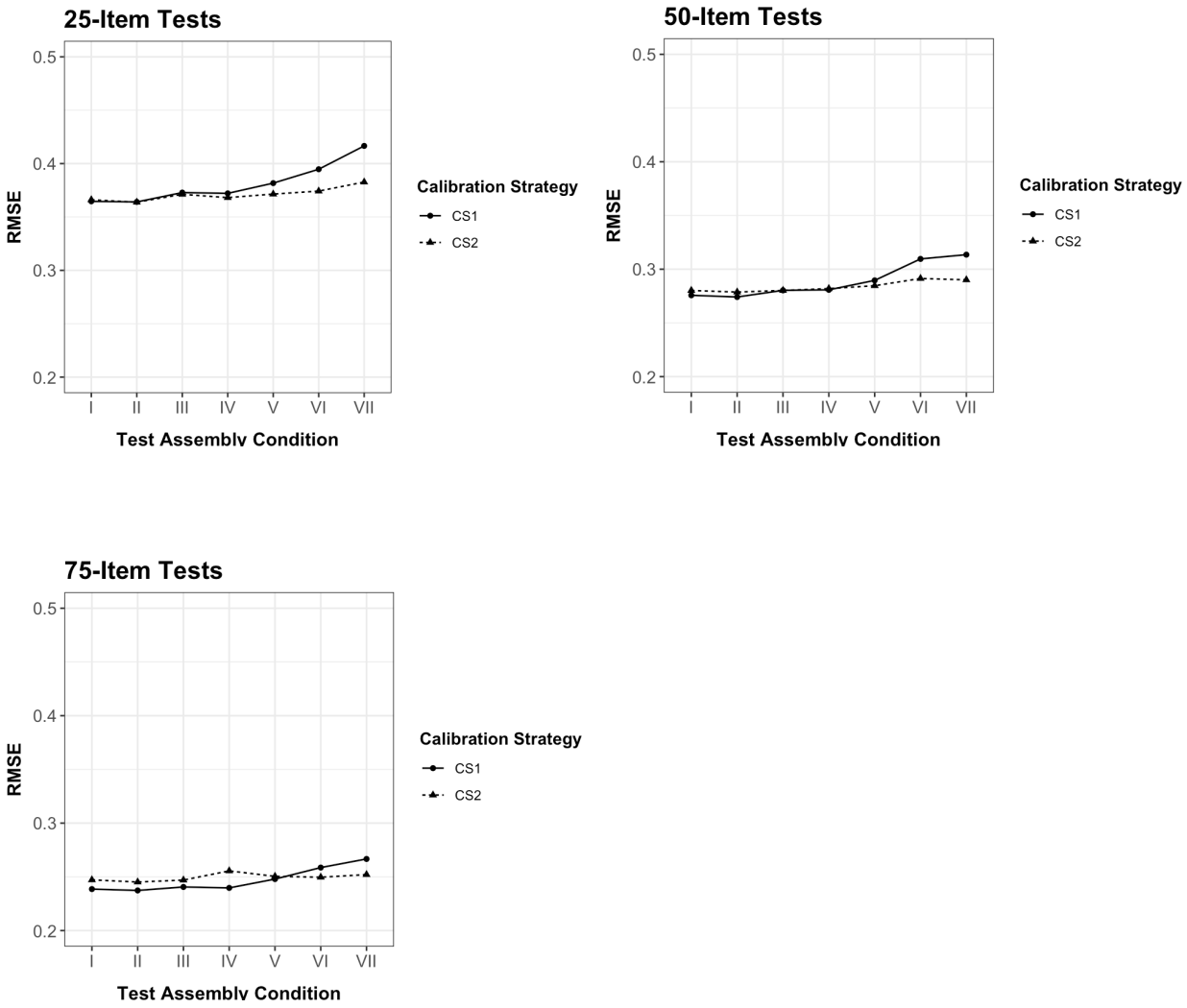
Table 5. Bias of θ Estimates for each Calibration Strategy, across Test Assembly Conditions and Test Lengths

		Test Length		
Test Assembly Condition	Calibration Strategy	25	50	75
I	CS_1	0.006	-0.003	0.005
	CS_2	0.006	-0.003	0.006
II	CS_1	-0.004	-0.005	0.007
	CS_2	-0.005	-0.006	0.008
III	CS_1	0.050	0.041	-0.016
	CS_2	0.001	0.001	-0.007
IV	CS_1	0.085	-0.074	0.021
	CS_2	0.003	-0.005	-0.019
V	CS_1	0.109	0.081	-0.052
	CS_2	-0.001	0.002	-0.002
VI	CS_1	-0.113	-0.081	-0.060
	CS_2	-0.005	-0.002	0.001
VII	CS_1	0.139	-0.090	-0.083
	CS_2	-0.002	-0.002	-0.003

The RMSE results are presented in Figure 7 for each test length. The plots indicate that the total magnitude of variation was quite similar for CS_1 and CS_2 up until we reach condition IV. As the variation in the item parameters within families increased toward 0.5σ (conditions V, VI, and VII), the RMSE values showed a clear increase. Despite the general upward trend for both CS_1 and CS_2 for the high variation conditions, unilaterally using the CS_1 family-level item parameter estimates to score examinees resulted in relatively higher RMSE values compared to CS_2. That being said, the difference between the two calibration strategies in terms of their impact on scores appears to be mitigated somewhat as the test length increased. In addition, on using CS_1 estimates for conditions I-IV in the case of the 75-item tests seemed to result in relatively less error variance compared to CS_2—although the difference is very subtle. Overall, these findings suggest that when

operating under the assumption of statistical isomorphism (CS_1)—even though it is clearly violated—there are scoring accuracy issues, particularly for the 25- and 50-item tests.

Figure 7. QC Variation in Examinee Scores across Calibration Strategies, Test Lengths, and Test Assembly Conditions using RMSE

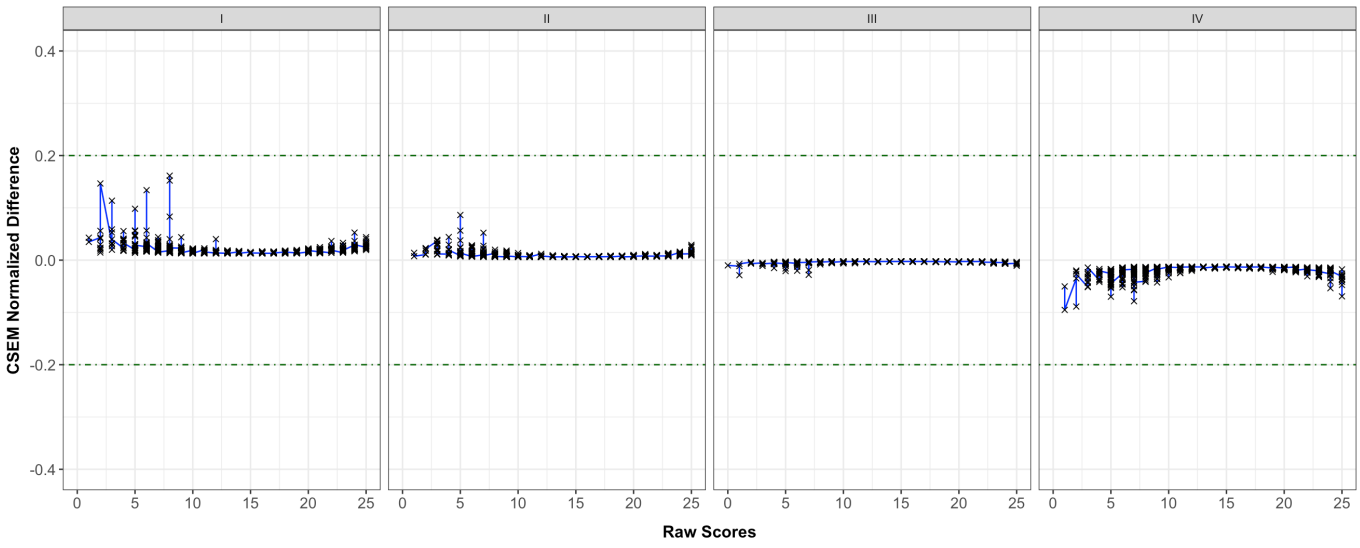


The third QC statistic used to assess the impact of calibration strategy on test scores was the normalized score difference (Luecht, 2016). This type of conditional statistic provided a direct comparison between CS_1 and CS_2 using the conditional standard errors of estimate (CSEMs) as the normalizing values. That is, the normalized difference is evaluated relative to the [pooled] asymptotic standard error of estimate—similar to an "effect size". This type of QC mechanism

was conditioned on the readily interpretable number-correct scores to demonstrate where along the score scale the error would be most evident when comparing the two calibration strategies.

The CSEM normalized differences for each test length and test assembly condition are presented in Figures 8 to 10. The plots show the raw scores (number-correct) on the horizontal axes and the normalized differences on the vertical axes for each test length. An absolute value of ≤ 0.2 was used as the threshold for detecting any sizeable differences in scores between the two calibration strategies (Luecht, 2016). This range is represented by the dark green dashed-dot lines on each plot. The plots illustrate that the majority of the normalized differences lie well within this "acceptable difference" range for each test assembly condition, and across test lengths. The exceptions are conditions V and VII for the 25-item tests. For these two conditions in particular, some of the normalized differences hovered near the $|0.2|$ acceptable effect-size range, but only for the lowest raw scores. Those differences dissipated at the longer test lengths (see Figures 9 and 10).

Figure 8. QC Variation in Scores between the Two Calibration Strategies Based on the Conditional Standard Error (CSEM) Normalized Differences for 25-Item Tests



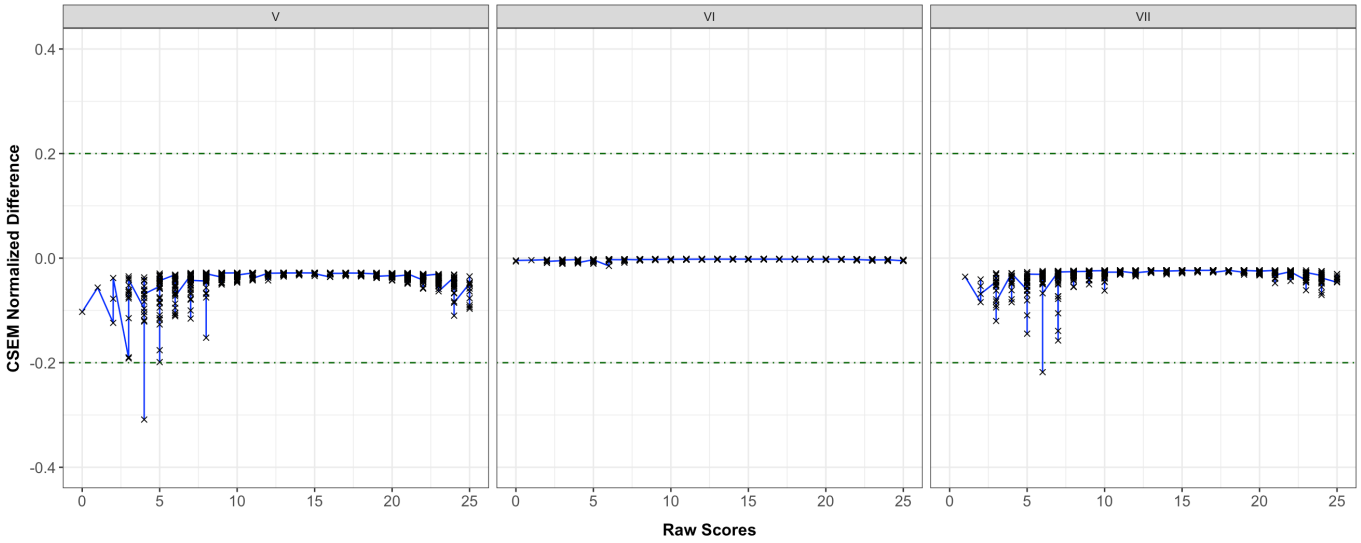
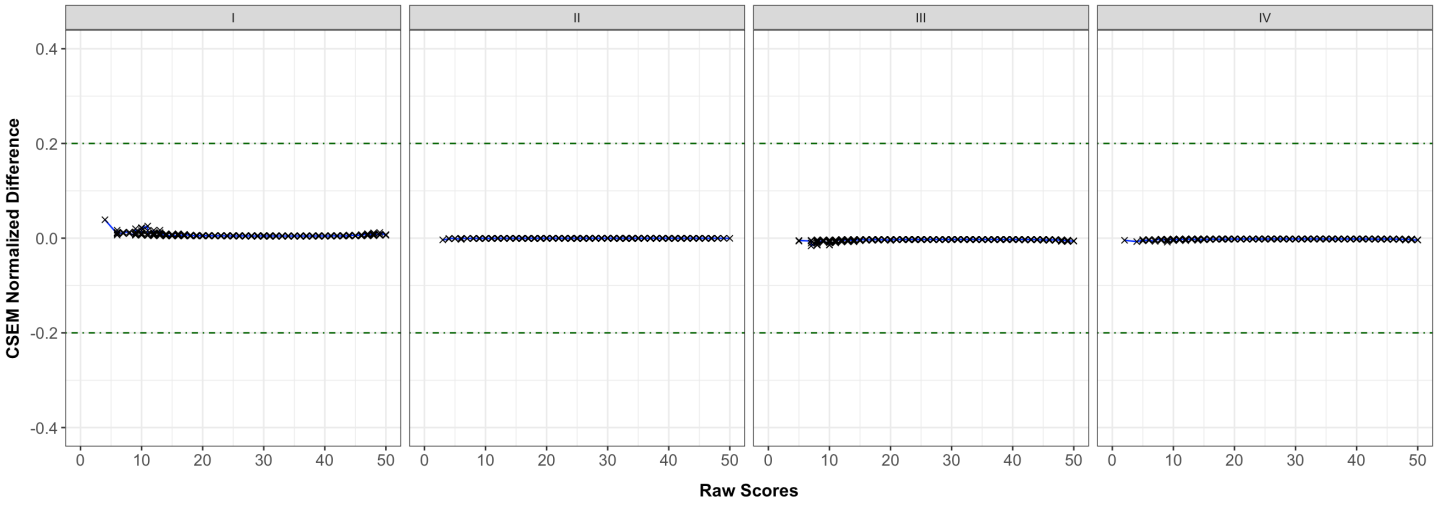


Figure 9. QC Variation in Scores between the Two Calibration Strategies Based on the Conditional Standard Error (CSEM) Normalized Differences for 50-Item Tests



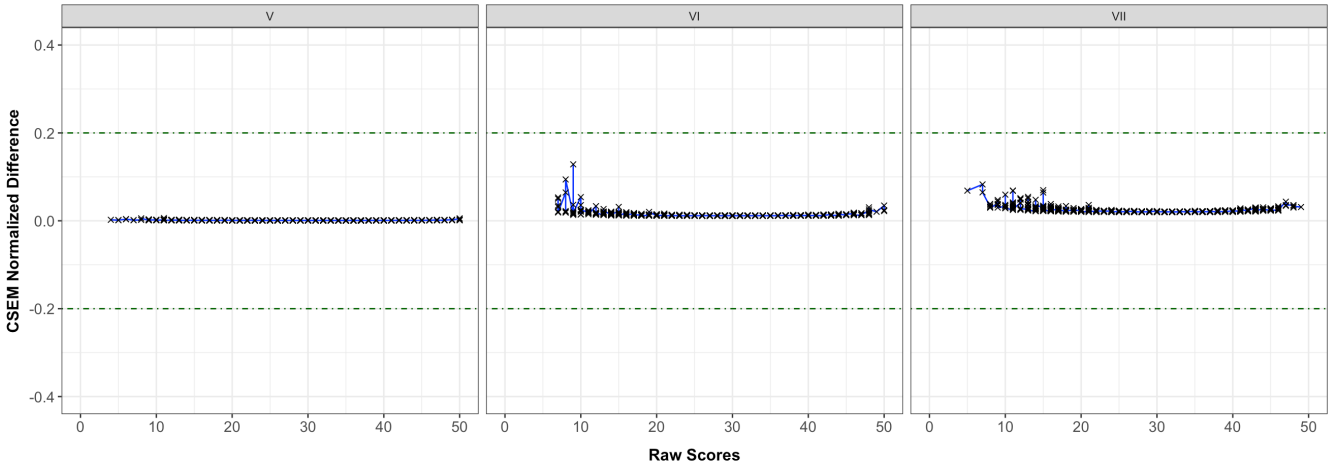
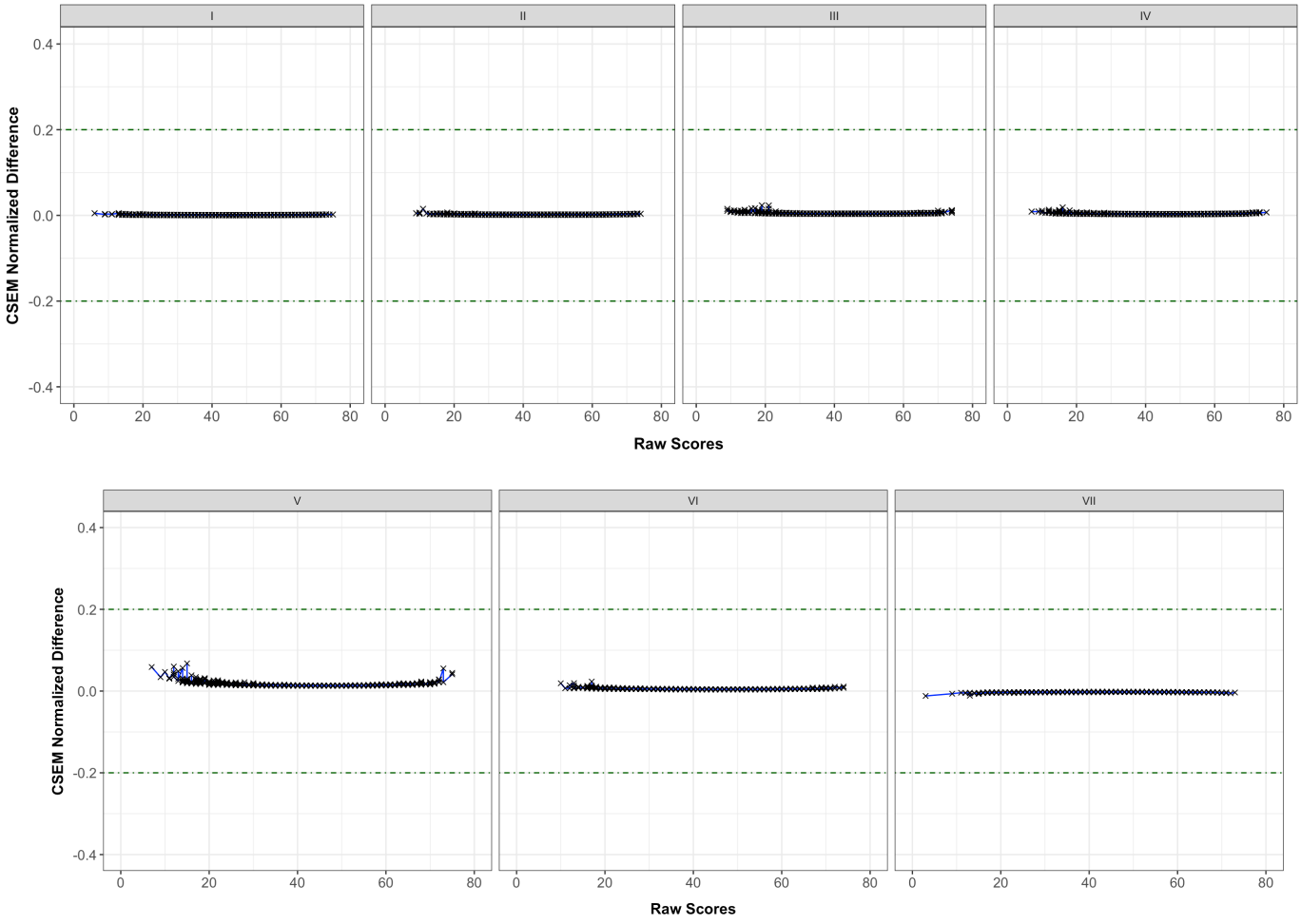


Figure 10. QC Variation in Scores between the Two Calibration Strategies Based on the Conditional Standard Error (CSEM) Normalized Differences for 75-Item Tests



Although it may appear that these results may slightly contradict the trends demonstrated for the bias and RMSE statistics the normalized difference plots shows that the error variations were, in a relative sense, well within ± 0.2 CSEMs, and negligible in a practical sense. In contrast, the bias statistics reported earlier for CS_1 tended to over-emphasize relatively small amounts of bias for conditions III-VII across the test. It should further be noted that the differences in bias across these test assembly conditions were more pronounced than the differences in RMSE between the two calibration strategies. When reconsidering those findings in light of the CSEM normalized difference results, the ultimate conclusion is that IRT models are robust enough to withstand the amount of [random] within-family item parameter variation introduced in this study. For example, even introducing high variation of 0.5σ within task- and item-model families did not have a serious impact on the scores. These findings are further supported by Luecht's (2024) study examining the impact of variation within item families on IRT linear equating parameters estimates and scores.

In conclusion, the results from Part A suggest the following. First, they provide sufficient evidence to support calibrating task model families when the within family variation is at 0.1σ or below. However, this tolerance limit can be relaxed to some extent, possibly up to the moderate variation of 0.2, with longer tests. Second even though within task model family variation of 0.2σ and 0.5σ showed increase trends in bias and RMSE for conditions III-VII under CS_1, this variation ultimately did not result in significant score differences between the two calibration strategies, especially for longer tests. (That said, all of error in this study was purely random. In practice, larger and systematic bias in the within-family item parameter estimates could have more dire consequences for operational scoring.)

PART B

Part B focused more on the quality of item difficulty predictions. This part of this study originated from the literature on item difficulty modeling (IDM) and/or computational linguistics. Thus, Part B aimed to specifically address how poorly predicted difficulty parameters could impact scores.

Research Question #2. How do different degrees of explained variance in predicted item parameters impact scores and person fit?

Quality of Predicted Item Difficulties and their Impact on Examinee Scores

Part B involved the generation of predicted item difficulty parameters of differential quality across R^2 conditions of 0.9, 0.8, 0.7, 0.6, and 0.5. These parameters, along with fixed item discrimination and guessing parameters, were used to score relatively large datasets (N = 5000) to obtain EAP estimates. The simulations were repeated for each R^2 condition and test lengths of 25, 50, and 75 items. As previously stated in Chapter III, although the goal of this study was to assess the impact of poor-quality item difficulty parameters on scores, it also sought to account for the possibility of scores being relatively insulated from errors in item difficulty prediction. Thus, once the EAP estimates were obtained for each R^2 condition, they were correlated with the true proficiency scores obtained during response data generation to examine whether the correlations did in fact decrease as the prediction of item difficulty progressively worsened. These correlations are presented in Tables 6-8 where each table has the results for a given test length.

Table 6. Correlations between True Proficiency Scores and EAP Estimates for each R^2 Condition for 25-Item Tests

	1	2	3	4	5	6
1. True EAP	--					
2. EAPs at 0.9	0.994	--				
3. EAPs at 0.8	0.992	0.999	--			
4. EAPs at 0.7	0.993	0.998	0.997	--		
5. EAPs at 0.6	0.992	0.997	0.998	0.996	--	
6. EAPs at 0.5	0.990	0.997	0.997	0.995	0.995	--

Table 7. Correlations between True Proficiency Scores and EAP Estimates for each R^2 Condition for 50-Item Tests

	1	2	3	4	5	6
1. True EAP	--					
2. EAPs at 0.9	0.996	--				
3. EAPs at 0.8	0.995	0.999	--			
4. EAPs at 0.7	0.995	0.999	0.999	--		
5. EAPs at 0.6	0.994	0.999	0.998	0.998	--	
6. EAPs at 0.5	0.995	0.998	0.998	0.997	0.996	--

Table 8. Correlations between True Proficiency Scores and EAP Estimates for each R^2 Condition for 75-Item Tests

	1	2	3	4	5	6
1. True EAP	--					
2. EAPs at 0.9	0.998	--				
3. EAPs at 0.8	0.997	0.999	--			
4. EAPs at 0.7	0.998	0.999	0.998	--		
5. EAPs at 0.6	0.996	0.999	0.999	0.998	--	
6. EAPs at 0.5	0.997	0.999	0.999	0.998	0.998	--

The correlations between the true scores and the EAP estimates for each R^2 condition were consistently above 0.99, indicating a near perfect correlation between the them, contradicting expectations regarding decreasing correlations as the quality of difficulty prediction degenerated. These initial results suggest that IRT scoring is and should be insulated from item difficulty prediction errors that occur with decreasing correlations between the true difficulty parameters and the estimates sampled with varied degrees of correlations with those parameters. In other words, even with poorly predicted item parameter estimates, the estimated scores are likely to be highly correlated with the true proficiency scores.

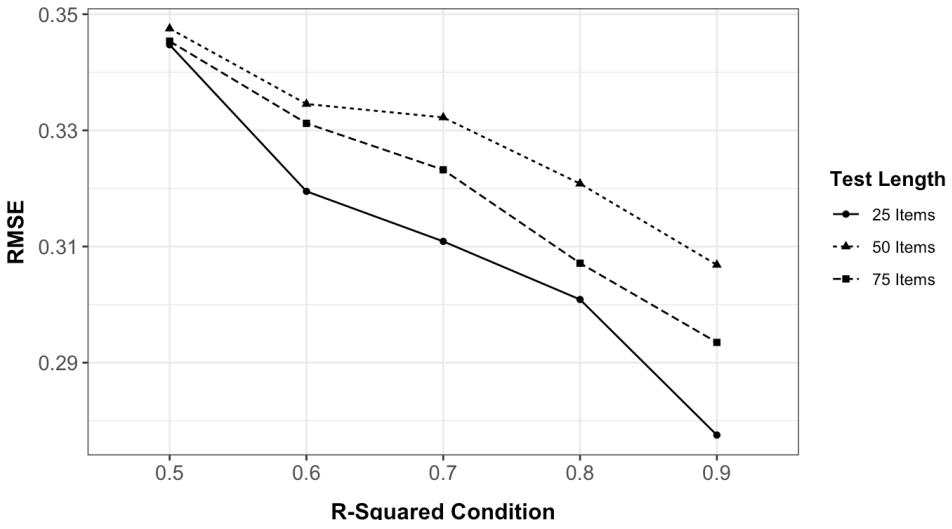
This finding could be due to the monotonic relationship between raw scores and EAPs—a finding that holds even for more complicated IRT models. That is, as raw scores increase, so do the IRT proficiency estimates. Moreover, this relationship is explicit for the Rasch model, given the sufficiency of raw scores for unconditional maximum likelihood estimation. It would certainly hold in this study given the modified 1PL model used (i.e. with fixed discrimination and pseudo-guessing parameters). The well-known monotone relationship between raw scores and IRT estimates implies that we should probably not expect to see any pronounced differences in the rank ordering of the estimated scores regardless of how poorly the item difficulties are "predicted".

Given these initial results, residuals were calculated slightly differently from Part A. Since there appeared to be no direct impact of the poorly predicted difficulty parameters on test scores, residual-based fit statistics were calculated based on the observed score for each examinee and their predicted score to discover any noticeable differences. Table 9 presents the mean and standard deviation of these residuals along with the RMSE across the 25-, 50-, and 75-item tests. Figure 9 depicts the RMSE values plot across R^2 conditions and test length to observe any patterns in the residual error.

Table 9. Residual-based Fit Statistics for the different R^2 Conditions across Test Lengths

	25-Item Test			50-Item Test			75-Item Test		
R-squared Condition	M	SD	RMSE	M	SD	RMSE	M	SD	RMSE
0.9	-0.004	0.139	0.278	-0.005	0.151	0.307	-0.012	0.157	0.293
0.8	-0.009	0.164	0.301	0.000	0.174	0.321	-0.011	0.174	0.306
0.7	0.001	0.189	0.311	0.002	0.185	0.332	-0.005	0.193	0.322
0.6	0.006	0.193	0.320	-0.001	0.188	0.335	-0.005	0.205	0.333
0.5	0.008	0.221	0.345	0.006	0.204	0.348	-0.004	0.215	0.341

Figure 11. RMSE for the Predicted Probabilities, $P(\theta_j, v_i)$ across Test Lengths



The residual-based fit statistics showed larger errors in terms of the standard deviations and RMSE as the correlations between the true and predicted item difficulty parameters decreased. However, these results seemed relatively invariant to test length in this study (see Figure 9). The larger errors could possibly be attributed to the poorly predicted $P(\theta_j, v_i)$ or predicted probabilities of endorsing an item because of the predicted item difficulty parameters for each R^2 condition. In any case, there did not appear to be any directly observable, serious impact of the poorly predicted

item difficulty estimates on the scores. However, there were subtle differences observed in terms of the RMSEs.

Quality of Predicted Item Difficulties and their Impact on Person Fit

The l_z person fit statistic (Drasgow et al., 1985) was computed for each examinee under the different R^2 conditions across the three test lengths. The means and standard deviations for the person fit statistics across these conditions are presented Table 10. Differences between the average l_z for each R^2 and true proficiency score conditions were more noticeable as the correlations between the true and predicted item difficulties decreased.

Table 10. Descriptive Statistics of the l_z Person Fit Statistic for each R^2 Condition and Test Length

R^2Condition/Test Length	25-Item Test		50-Item Test		75-Item Test	
	M	SD	M	SD	M	SD
True EAP	0.967	0.783	1.149	0.775	1.335	0.787
0.9	0.882	0.705	1.101	0.680	1.036	0.722
0.8	0.781	0.694	0.447	0.766	0.797	0.748
0.7	0.445	0.808	0.386	0.760	0.485	0.769
0.6	0.494	0.762	0.238	0.782	0.372	0.757
0.5	0.222	0.840	-0.062	0.896	0.152	0.807

Figures 12-14 display the density estimates of the distributions of l_z statistic for each R^2 condition for 25-, 50-, and 75-item tests. These plots are intended to help visualize the differences in person fit across the different conditions for each test length. The plot at the very bottom of each figure represents the distribution of the person fit statistics for the true proficiency scores. The plots above it represent R^2 conditions 0.9, 0.8, 0.7, 0.6, and 0.5, respectively. For each plot, the mean l_z is indicated using the colored dashed lines that are superimposed on each distribution.

Figure 12. Distribution of the Person Fit Statistics for each R^2 Condition for the 25-Item Tests

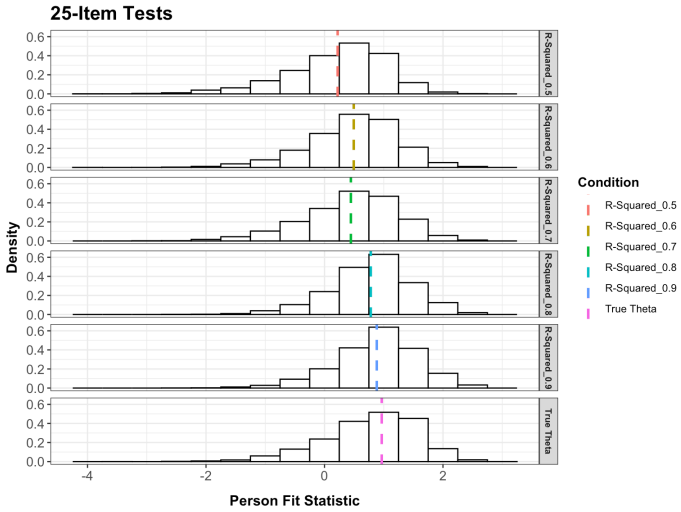


Figure 13. Distribution of the Person Fit Statistics for each R^2 Condition for the 50-Item Tests

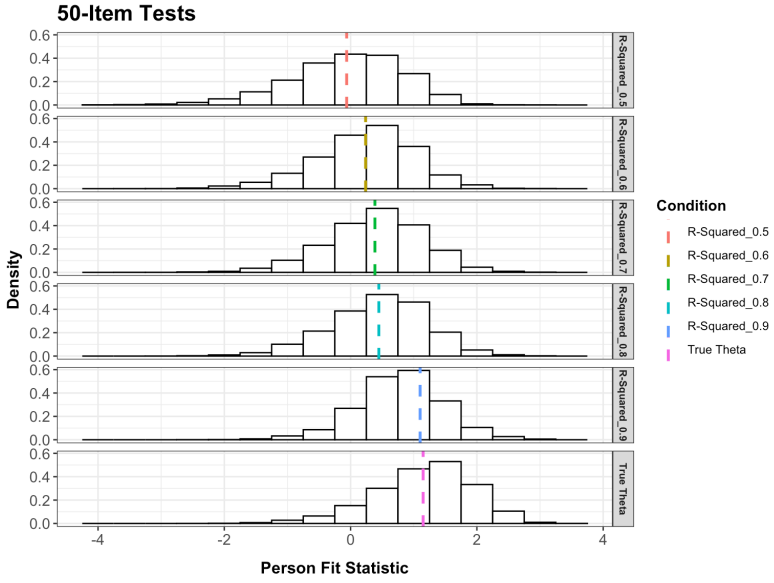
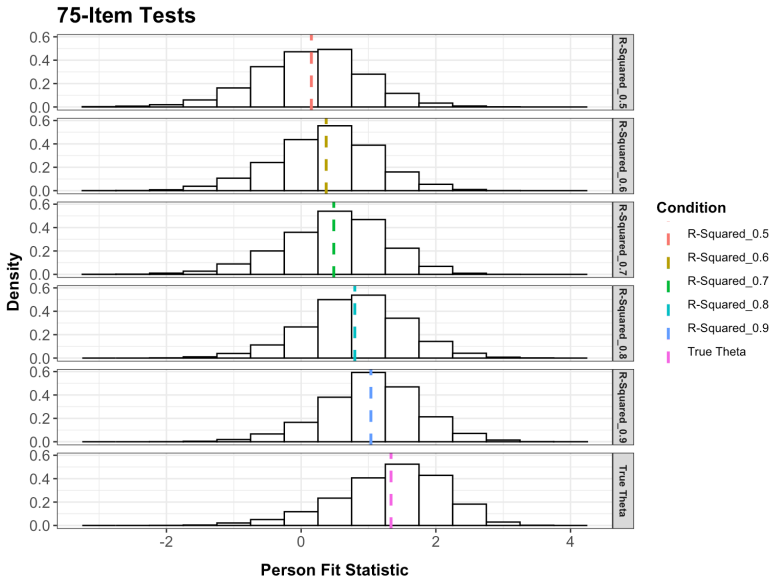


Figure 14. Distribution of the Person Fit Statistics for each R^2 Condition for the 75-Item Tests



The distributions of the l_z statistics progressively shifted towards the left of the true distribution, suggesting that person fit did get worse with a decline in correlations between the true and predicted item difficulties. These differences in distributions and potential misfit for the lower R^2 conditions can possibly be attributed to the poorly predicted $P(\theta_j, v_i)$ resulting from increased prediction error in item difficulty for these conditions. There also did not appear to be any trends in misfit across the different test lengths. Given that there are appreciable differences in potential misfit as the R^2 falls below 0.8, it can be argued that Bejar's (1983) proposed requirement of an explained variance of 0.8 in predicted difficulties is quite reasonable.

Summary

This chapter presented the results for Parts A and B, directly addressing their respective research questions from Chapter I. The results from Part A highlighted differences between the two calibration strategies across the seven test assembly conditions where CS_1 were found to be more appropriate for lower variation conditions and CS_2 for higher variation conditions. In terms

of addressing just how much variation matters, as the variation within task and item model families increased from 0.1 to 0.5 standard deviations, bias statistics and RMSE increased for test assembly conditions III-VII. However, when examining the normalized score differences relative to the average CSEM for the two calibration strategies across the seven test assembly conditions, little difference was found between the two sets of estimated scores.

The results from Part B provided insight into how robust IRT models can be even when using poorly predicted difficulty parameters to score examinees. While the estimated scores from each R^2 condition highly correlated with the true proficiency scores, the residual-based fit statistics for $P(\theta_j, v_i)$ revealed larger errors as the correlations between the true and predicted item difficulties decreased. Results from the person fit analyses revealed misfit is more likely to be observed for the lower R^2 conditions. Thus, IDM research should ideally continue its investigation of the impact of predicted difficulties on scores as well as person fit. Overall, the results from both parts A and B clearly showed how and when test scores may or may not be affected as test design conditions vary from optimal to inadequate.

CHAPTER V: CONCLUSIONS

The overarching purpose of this study was to address gaps in the literature on modern approaches toward item and test development, stated in Chapter II. These were to: 1) determine appropriate calibration strategies for item families in different contexts; 2) examine variation of items within families and determining acceptable tolerances of variation in item characteristics; 3) assess the impact of using differential quality of item families on scoring; and 4) evaluate the impact of on-fly-estimates of item characteristics on scoring. Each of these limitations will be addressed based on the findings from Parts A and B in the sections that follow. This chapter finally concludes with study implications, limitations, and future opportunities for research.

Summary of Findings

Determine Appropriate Calibration Strategies for Item Families in Different Contexts

One of the anticipated benefits of implementing a principled design framework like AE—a framework that focuses on designing task-model families rather than writing unique items—is that overall item generation process can become more systematic, predictable and scalable (Luecht, 2012; Luecht & Burke, 2020). A second benefit is that AE introduces detailed cognitive task specifications when designing each task model with the explicit expectation that each family of items will function isomorphically while maintaining the task model's location on the proficiency scale. That is not just wishful thinking. AE expects strong quality control (QC) procedures to be implemented to monitor and verify the psychometric integrity of each task model. When problems are identified, we can modify cognitive task constraints to mitigate them. Both these benefits suggest that if implemented correctly, an AE-like design strategy supports the statistical control of difficulty, such that item families within a task model operate in psychometrically predictable ways (Luecht & Burke, 2020). If items within a task model family

are truly isomorphic, family-level calibrations can be used provided that the difficulty of these items met QC specified tolerances.

Part A sought to determine what these tolerances might be as well as which calibration strategy (calibrating at the family-level or individual items) would be appropriate in both ideal and non-ideal conditions. The two calibration strategies were compared across low, medium, and high variation test assembly conditions of 0.1, 0.2, and 0.5 standard deviations within task and item-model families across the three test lengths of 25-, 50-, and 75-item tests. The RMSD plots for the item discrimination and difficulty estimates showed increasing values across the seven test assembly conditions, where calibrating at the task model level was more suitable for the low and medium variation conditions than the high variation condition.

For calibration purposes, a QC tolerance limit of 0.2σ within task model families might be recommended as the maximum acceptable variation allowable to collapse the data and calibrate at the task model level. For example, if a task model was designed to target a *b*-parameter (item difficulty) of 0.6 and the standard deviation of the estimated item difficulty parameter estimates was $s(\hat{b}_f) = 0.19$, then calibrating at the family (task model) level might be suitable. Those task model item parameter estimates could continue to be used but monitored to ensure that the acceptability threshold (tolerance) in the within-family item parameter estimates was not exceeded. Conversely, for item families where the tolerances were exceeded, we might need to adopt a different calibration strategy similar to CS_2 (also see Shu et al., 2010). That is, from a QC perspective, we assumed that the affected item families are no longer functioning according to their design specifications. Therefore, we will need to continue to pilot test those items and calibrate them as independent instances until we can solve the task model family design issues needed for empirically substantiated isomorphism.

Examine Within-Family Variation and Determine Acceptable Tolerances of Variation in Item Characteristics

On-going quality control (QC) plays an essential role within AE framework to monitor and empirically verify statistical isomorphism—especially with respect to item difficulty (task model family location). Variation within item families was induced at 0.1σ, 0.2σ, and 0.5σ standard deviation units to attempt to determine the breaking point at which the accuracy of IRT item parameters was compromised or significantly deviated from the intended locations. While there was evidence from Part A to suggest that a QC tolerance of 0.2σ could be recommended for using family-level calibrations over individual item calibrations, establishing a tolerance limit within task and item model families for scoring was comparatively less straightforward. This was possibly because the CSEM normalized difference between the score estimates from the two calibration strategies was minimal across the low, medium, and high variation conditions. Moreover, it suggested that IRT scores are robust enough to withstand error introduced by using family-level calibrations to score examinees even for the high variation conditions. While these results are reassuring in so far as that test scores seemingly remain unaffected despite high within-family variation, it becomes difficult to specify and justify a tolerance limit based on this evidence. Thus, it may be a matter of policy, based on empirical research to establish acceptable tolerances for within-family variation for scoring examinees. From a strictly AE QC standpoint, however, variation should ideally be as low as possible.

Assess the Impact of Using Differential Quality of Item Families on Scoring

Part A addressed this issue by assessing the combined impact of using each calibration strategy across the different conditions of variation within task and item model families using three types of QC statistics – bias, RMSE, and the CSEM of normalized differences. The bias statistics

did see an increase in magnitude for family-level calibrations for moderate and high variation conditions. In this study, the within family variance was uniform across the range of difficulty based on which the task models were designed. The bias of the EAP estimates indicated that they were overestimated for the moderate and high variation conditions. This could be because this study used a combination of the 3PL model and EAP scoring method, since using EAP estimates alone would not necessarily result in an increase in average bias. Given that there is less information for the theta estimates at the lower end of the ability scale, further reduced by the pseudo-guessing parameters, the bias for the EAP estimates at the lower end of the ability scale is relatively higher than those at the higher end of the scale. This may be reflected in the relatively large average bias for those moderate and high variation conditions when calibrating at the task model level. Alternatively, using a 2 PL model with MLE estimates may reduce the average bias of score estimates.

The RMSE values also increased for the moderate and high variation conditions for both calibration strategies, with family-level calibrations showing relatively larger values at high variation conditions than item level calibrations. Yet, these differences between the two calibration strategies tended to decrease with an increase in test length. The score estimates obtained by using each calibration strategy were compared by normalizing the differences between these score estimates relative to the average CSEM for the two calibration strategies. Despite using item parameter estimates from the moderate to high variation conditions to score examinees, there was no evidence to suggest any substantial differences between the two sets of scores. The test scores for all three test lengths were not severely impacted because of using item families from moderate to high variation conditions.

Evaluate the Impact of On-the-fly Estimates of Item Characteristics on Scoring and Person Fit

Part B attempted to verify Bejar's (1983) proposition that there needs to be an explained variance of 0.8 (R^2 = 0.8) in predicted difficulties for these estimates to substitute the empirical difficulties in scoring examinees. Since the estimated scores from the different R^2 conditions correlated highly with the true proficiency scores, the residual-based fit statistics were analyzed to detect larger error for the R^2 conditions less than 0.8. The results did show subtle increases in the magnitude of error as the correlations between the true and predicted item difficulty parameters decreased.

Person fit analyses also showed differences between the distributions of the l_z statistics for the true score condition and the lower R^2 conditions of 0.7 and below, suggesting that misfit may more likely occur at these conditions. In terms of how good the predicted difficulties need to be for IDM research to provide useful enough estimates to tackle examinee scoring and potential person (mis)fit, Bejar's (1983) proposition seems to hold. Even though the results suggest that scores are not directly impacted when using poorly predicted difficulty parameters, the risk of introducing error in terms of person (mis)fit could be relatively greater than using predicted difficulties from higher R^2 conditions. If examinees' expected item-score patterns are compromised as a result of using poorly predicted item difficulty parameters, it would suggest that the response data does not fit the IRT model or that the responses to the items deviate from the expected responses for the model (Ames & Penfield, 2015; Felt et al., 2017; Meijer & Sijtsma, 2001). It could also indicate that these unlikely item-score patterns are being driven by factors other than the test construct being measured (Meijer, 2002). Therefore, despite scores being

unaffected, the indirect impact in terms of person fit could bring into question the decisions made based on these test scores.

Implications

This study presented various scenarios of calibration and scoring using differential quality of item families and predicted item difficulty estimates. It highlighted potential concerns of calibrating at the family-level under high variation conditions and issues regarding person fit when utilizing poorly predicted difficulty parameters. Therefore, the most important implication of this study is that it shows that developing a QC system for item families and item difficulty modeling (IDM), as well as setting specific acceptable tolerances is entirely possible, and even necessary. It also implies that assumptions made during item and test design require constant verification, supporting an iterative process towards item and test development. In other words, when calibrated items and/or item families fall outside the acceptable limits, it demonstrates the need for further revision before they can be made operational.

Second, even though other complex calibration procedures could have been used (Geerlings et al., 2011) at the task model level, using a hierarchical calibration system, when items within a family were indeed isomorphic, implied that a relatively straightforward calibration procedure could be implemented in practice. Thus, when items within a family are found to be acceptable based on QC criteria, response data can be re-organized to support calibration at the family-level without the use of specialized software or complex statistical models.

Third, findings from most of the IDM research across content domains typically stop at correlating predicted item parameters and scores with their empirical or operational counterparts (Embretson & Kingston, 2018; McCarthy et al., 2021; Settles et al., 2018). This study has demonstrated why those correlations are likely to be high, justifying the need for IDM research to

look beyond the correlations to study impact on person fit. Thus, the results from this study highlight the need for QC in IDM work for evaluating the impact of resulting predictions on scores and person fit.

Limitations

Although this two-part simulation study attempted to be comprehensive in its approach by demonstrating the possibilities across a variety of optimal and sub-optimal conditions, the findings reflect a highly controlled set of simulations which may differ from operational practice. For example, in practice, systematic [non-random] errors in the item-family operating characteristics could evolve, especially if item families generated via artificial intelligence (AI) were based on flawed neural nets that led to subtle biases. In any case, these findings are limited in generalizability to the rather simplistic assumptions, simulated contexts, study design conditions, and constraints used in this study.

Three rather convenient assumptions were made in this study. First, all noise within task model families was random in nature. In that sense, the error introduced here was effectively indistinguishable from random sampling or measurement errors typically encountered during IRT parameter estimation. The introduction of plausible non-random errors should be investigated in future research. The second assumption was that items within families were assigned to randomly equivalent groups. This assumption eliminated the need to consider the impact of variation in the item parameter estimates within families on equating or linking procedures with non-equivalent groups (e.g., linking IRT scales over time and conditions of measurement). Luecht (2024) offered a small-scale look at this issue by simulating a non-equivalent groups linking design to explore the impact of variability within item families on IRT linear equating parameters estimates and scores. Similar to the findings of the present study, his simulation study found that scoring bias,

estimation, and equating errors increased relative to traditional IRT equating results but not substantially enough to attribute its cause to variant item families. In any case, more complex simulations could be used to help push on the empirical boundaries of acceptable variability in the item parameter estimations when examinee groups are not randomly equivalent.

The final assumption was that the item families and predicted difficulty estimates (Part B) were established specifically to meet the variation conditions in this study, meaning that they were generated to be almost exactly at that level of within-family variability as well as the variability in the predicted difficulties. For example, variation was also kept uniform at the task and item model levels as well as for the item discrimination and difficulties using the strict 0.1σ, 0.2σ, and 0.5σ specifications. It is entirely possible that item families developed in practice may not vary in this uniform and tightly controlled manner. Thus, more empirical research is needed to support the results of this study in terms of the magnitude of variation, setting QC tolerance limits, and exploring their impact in the context of a non-equivalent groups design.

The calibration procedures used in this study represent only one method of calibrating item families. A more comprehensive comparison of the different item family calibration methods at the different variation conditions is needed to understand if the results would vary based on methodology (also see Geerlings, Glas & van der Linden, 2011; Sinharay & Johnson, 2012). The same applies to the QC statistics used in this study, which are only some of the ways of evaluating whether item families and predicted difficulty parameters are functioning as intended.

Future Work

This study outlines the possible and potential impact of poorly designed item families and predicted difficulties on scores under different conditions. There is still more empirical research to be done in this area that can support the design of a larger QC framework for modern item and

test development approaches. First, developing families of items and/or items with predicted item parameter estimates using Large Language Models (LLMs) can accelerate the development and implementation of a QC system, if these models are trained to do so. For example, LLMs can be trained to develop *and* review items, similar to the training given to content experts. Some of this work in item development is currently being done in the medical field (Benoit, 2023; Gilson et al., 2023; von Davier, 2023; Walker et al., 2023). Moreover, recent research using AE in IDM to enhance score interpretation, trained content experts to assess items for their complexity on very specific criteria and found that difficulty could be predicted at R^2 of greater than 0.77 (Hu, 2024; Li, 2024; Whitney et al., 2024; Wilmurth et al., 2024). With more sophisticated generative pre-trained transformer models (GPT) such as the GPT 4.0 (OpenAI, 2023), it is reasonable to suggest that these models can be trained to support at least some of the QC work. It may be possible to achieve this using Wei et al. ‘s (2022) chain-of-thought prompting strategy, which breaks down instructions for GPT models into sequential, interrelated steps, thereby enabling them to engage in complex reasoning.

Second, subsequent studies can explore variation in different innovative item types that require different scoring protocols, relying on polytomous item response models. In this case, a separate QC system may be needed using residual-based statistics that are more appropriate for these conditions.

Third, another area of research could be extending this study’s simulation design to include tests in certification and licensure, where score precision is needed at a specific cut-score. An in-depth analysis of item and person fit statistics for certification and licensure exams would also help in assessing their quality and impact on test scores.

Fourth, this study explored the use of calibration strategies, scoring procedures, and residuals analyses in different statistical contexts. Further research is needed in analyzing substantive isomorphism, where items within a family should ideally be comparable in terms of content standards and cognitive complexity (Luecht & Burke, 2020). Moreover, examining the same study conditions in terms of their impact on different student groups such as English Language Learners (ELLs) or students with disabilities who require testing accommodations could also be explored. For example, residual analyses can be grouped by the different student subgroups.

Fifth, Part B explored the impact of errors introduced in item difficulty on test scores. In addition, more empirical research needed to explore how the covariance between item discrimination and item difficulty parameters can impact examinee scores under the different R^2 conditions.

Finally, future work can expand Luecht's (2024) study to include a comparison of different proportion of common items amongst test forms that come from item families with different within-family variation. Such a comparison can be informative in establishing conditions under which item families can be used to serve as anchor items or least highlight conditions under which they would be inappropriate. Thus, there are several possibilities for future research to explore, eventually working towards developing guidelines and QC systems that support the implementation of modern item and test development approaches to suit the needs of different assessment programs.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Brooks/Cole Publishing Company. https://books.google.com/books?id=cgElAQAAIAAJ

Alsubait, T., Parsia, B., & Sattler, U. (2013). Next generation of E-assEssmEnt: Automatic generation of questions. *International Journal of Technology Enhanced Learning*, *4*(3-4), 156. https://doi.org/10.1504/ijtel.2012.051580

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*.

Ames A. J., Penfield R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, 34, 39-48. doi: 10.1111/emip.12067

Arendasy, M., & Sommer, M. (2005). The effect of different types of perceptual manipulations on the dimensionality of automatically generated figural matrices. *Intelligence*, *33*, 307-324. doi: 10.1016/j.intell.2005.02.002

Arendasy, M., & Sommer, M. (2010). Evaluating the contribution of different item features to the effect size of the gender difference in three-dimensional mental rotation using automatic item generation. *Intelligence*, *38*, 574-581. doi:10.1016/j .intell.2010.06.004

Bai, Y. (2019). *Cognitive diagnostic models-based automatic item generation: Item feature exploration and calibration model selection*. [Unpublished doctoral dissertation]. Columbia University, NY.

Bejar, 1.1. (1990). A generative analysis o f a three-dimensional spatial task. *Applied Psychological Measurement, 14*, 237-245. doi: 10.1177/014662169001400302

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*(3), 303-310.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–357). Lawrence Erlbaum Associates, Inc.

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (ETS RR-96-13). Princeton, NJ: ETS.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-218). Lawrence Erlbaum Associates.

Bejar, I. I. (2012). Item generation: Implications for a validity argument. In M. J. Gierl & T. M. Haladyna (Eds*.), Automatic item generation: Theory and practice* (pp. 40–55). New York, NY: Routledge.

Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement, 15*(2)*,* 129-137. https://doi.org/10.1177/014662169101500202

Bejar, I. I., Lawless, R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment, 2*(3). Retrieved from http://www.jtla.org

Benoit, J.R.A. (2023). *ChatGPT for Clinical Vignette Generation, Revision, and Evaluation.* MedRxiv[Preprint]. https://doi.org/10.1101/2023.02.04.23285478

Bormuth, J. R. (1970). *On the theory of achievement test items.* University of Chicago Press.

Brucia, R. C. (2020). *Operationalizing item difficulty modeling in a medical certification context*. [Unpublished doctoral dissertation]. University of North Carolina at Greensboro, NC.

Burke, M., Pitkin, N., Durning, S., Carr, J. (2020). Leveraging scientific advances to inform and enhance item development. *CLEAR Annual Exam Review*, *30*(2), 1-15.

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 195-217). Cambridge University Press.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review, 97*(3), 404-431. doi: 10.1037/0033-295X.97.3.404

Case, S. M., Holtzman, K., & Ripkey, D. R. (2001). Developing an item pool for CBT: A practical comparison of three models of item writing. *Academic Medicine, 76*(10), S111-S113.

Cho, S. J., De Boeck, P., Embretson, S., & Rabe-Hesketh, S. (2014). Additive multilevel item structure models with random residuals: Item modeling for explanation and item generation. *Psychometrika, 79,* 84-104.

Choi, J., Kim, H., & Pak, S. (2018). Evaluation of automatic item generation utilities in formative assessment application for Korean high school students. *Journal of Educational Issues*, *4*(1), 68. https://doi.org/10.5296/jei.v4i1.12630

Choi, J., Kim, S., & Yoon, K. (2016). K-Math Workbook Grade 6. Clarksville, MD: CAFA Lab, Inc.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Wadsworth Publishing Company.

Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/bf02310555

Daniel, R.C., & Embretson, S.E. (2010). Designing cognitive complexity in mathematic problem-solving items. *Applied Psychological Measurement, 34,* 348-364.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* Guilford Press.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. Psicología Educativa, 20(2), 89–97.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Drasgow, F., Luecht, R.M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–516). Washington, DC: American Council on Education.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of sentence structure variables on performance in reading comprehension tests. *Reading Research Quarterly, 16*, 486-514.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407–433. https://doi.org/10.1007/BF02294564

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory items. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219-250). Lawrence Erlbaum Associates.

Embretson, S. E., & Kingston, N. M. (2018). Automatic item generation: A more efficient process for developing mathematics achievement items? *Journal of Educational Measurement*, *55*(1), 112-131. https://doi.org/10.1111/jedm.12166

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*(2), 175 193. https://doi.org/10.1177/014662168701100207

Embretson, S. E., & Yang, X. (2007). Automatic item generation and cognitive psychology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 747–768). North Holland, UK: Elsevier.

Embretson, S.E., & Daniel, R.C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem-solving items. *Psychological Science Quarterly, 50,* 328-344.

Enright, M.K., Morley, M., & Sheehan, K.M. (2002a). Items by design: The impact of systematic feature variation of item statistical characteristics. *Applied Measurement in Education, 15*(1), 49-74.

Enright, M.K., & Sheehan, K.M. (2002b). Modeling difficulty of quantitative reasoning items: Implications for item generation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-158). Lawrence Erlbaum Associates.

Fay, D., Levy, R., & Mehta, V. (2018). Investigating psychometric isomorphism for traditional and performance-based assessment: Investigating psychometric isomorphism. *Journal of Educational Measurement, 55*(1)*,* 52-77. 10.1111/jedm.12163.

Finn, P.J. (1975). A question writing algorithm. *Journal of Reading Behavior, 4*, 341-367.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359-374. https://doi.org/10.1016/0001-6918(73)90003-6

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: main idea, inference, and supporting idea items. *ETS Research Report Series*, *1993*(1), i–48. https://doi.org/10.1002/j.2333-8504.1993.tb01524.x

Geerlings, H., Glas, C. A. W., & van der Linden, W. J. (2011). Modeling Rule-Based Item Generation. *Psychometrika*, *76*(2), 337–359. https://doi.org/10.1007/s11336-011-9204-x

Gierl, M. & Lai, H. (2016). A Process for Reviewing and Evaluating Generated Test Items. *Educational Measurement: Issues and Practice 35*(4), 6-20.

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge.

Gierl, M. J., & Lai, H. (2006). Automatic item generation. In S. Lane, M.R. Raymond, & T.M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 410–429). New York, NY: Routledge.

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing*, *72*(3), 273-298. doi:10.1080/15305058.2011.635830

Gierl, M. J., & Lai, H. (2013a). Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, 32, 36–50. doi:10.1111/emip.12018

Gierl, M. J., & Lai, H. (2013b). Evaluating the quality of medical multiple-choice items created with automated processes. *Medical Education*, 47, 726–733. doi:10.1111/medu.12202

Gierl, M. J., & Lai, H. (2017). Using automatic item generation to create solutions and rationales for computerized formative testing. *Applied Psychological Measurement*, *42*(1), 42-57. https://doi.org/10.1177/0146621617726788

Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge.

Gierl, M. J., Lai, H., & Turner, S. (2012). Using automatic item generation to create multiple-choice items for assessments in medical education. *Medical Education*, 46, 757–765. doi:10.1111/j.1365-2923.2012.04289.x

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, *9*, e45312. https://doi.org/10.2196/45312

Gitomer, D. & Bennett, R. (2003). Unmasking constructs through new technology, measurement theory and cognitive sciences, *in Technology and Assessment: Thinking Ahead: Proceedings of a Workshop*. Washington, DC: Board on Testing and Assessment, Center for Education.

Glas, C. A. W. & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. Applied Psychological Measurement, 27(4), 247–261.

Glas, C.A.W., & van der Linden,W.J. (2001). *Modeling variability in item parameters in item response models* (Research Report 01-11). Enschede, The Netherlands: Department of Educational Measurement and Data Analysis, University of Twente.

Gorin, J. S. (2005). Manipulating Processing Difficulty of Reading Comprehension Questions: The Feasibility of Verbal Item Generation. *Journal of Educational Measurement*, *42*(4), 351–373.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational measurement: Issues and practice, 25*(4), 21-35.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394-411.

Gorin, J. S., & Embretson, S. E. (2013). Using cognitive psychology to generate items and predict item characteristics. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 136-156). Routledge, New York, NY.

Gorin, J.S. (2011, April). *Novel IDM applications: Special populations and testing uses.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME). New Orleans, LA, USA.

Graf, E. A., & Fife, J. H. (2012). Difficulty modeling and automatic item generation of quantitative items: Recent advances and possible next steps. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 157-179). Routledge.

Ha, L. A., & Yaneva, V. (2018). Automatic Distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. *Proceedings of the Thirteenth Workshop*

on Innovative Use of NLP for Building Educational Applications, 389–398. https://doi.org/10.18653/v1/w18-0548

Haladyna, T. M. (2013). Automatic item generation: A historical perspective. In M. J. Gierl, & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 13-23). Routledge.

Haladyna, T. M., & Gierl, M. J. (2013). Obstacles for automatic item generation. In M. J. Gierl, & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 231-239). Routledge.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied measurement in education, 2*(1), 37-78.

Haladyna, T. M., & Shindoll, R. R. (1989). Item shells: A method for writing effective multiple-choice test items. *Evaluation & the Health Professions*, *12*(1), 97 - 104. https://doi.org/10.1177/016327878901200106

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer Science & Business Media.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE.

Hively, W., Patterson, H.L., & Page, S.H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*, 275–290.

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10*(4), 369–380. https://doi-org.libproxy.uncg.edu/10.1177/014662168601000405

Hu, A. (2024, April). *Reading item difficulty modeling with various machine learning methods* [Paper presentation]. National Council on Measurement in Education, Philadelphia

Irvine, S.H. (2002). The foundations of item generation for mass testing. In S.H. Irvine & P.C. Kyllonen (Eds.) *Item generation for test development* (pp. 3–34). Mahwah: Lawrence Erlbaum Associates.

Johnson, M. S., & Sinharay, S. (2005). Calibration of Polytomous Item Families Using Bayesian Hierarchical Modeling. *Applied Psychological Measurement*, *29*(5), 369-400. https://doi-org.libproxy.uncg.edu/10.1177/0146621605276675

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163-182.

Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A cost–benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice*, *38*(1), 48-53. https://doi.org/10.1111/emip.12237

Kostin, I. (2004). Exploring Item Characteristics That Are Related to the Difficulty of Toefl Dialogue Items. *ETS Research Report Series*, *2004*(1), i–59. https://doi.org/10.1002/j.2333-8504.2004.tb01938.x

Kurdi, G., Parsia, B., & Sattler, U. (2017). An Experimental Evaluation of Automatically Generated Multiple Choice Questions from Ontologies. In M. Dragoni, M. Poveda-Villalón, & E. Jimenez-Ruiz (Eds.), *OWL: Experiences and Directions – Reasoner Evaluation* (pp. 24–39). Springer International Publishing.

LaDuca A., Staples W. I., Templeton B., Holzman G. B. (1986). Item modeling procedures for constructing content-equivalent multiple-choice questions. *Medical Education*, *20,* 53-56.

Lai, H., Alves, C., & Gierl, M. J. (2009). Using automatic item generation to address item demands for CAT. In Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.

Lai, H., Gierl, M. J. (2012). Generating items under the assessment engineering framework. In M.J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 77-101). Routledge, New York, NY.

Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A., & De Champlain, A. (2016). Using automatic item generation to improve the quality of MCQ Distractors. *Teaching and Learning in Medicine*, *28*(2), 166-173. https://doi.org/10.1080/10401334.2016.1146608

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2015). *Handbook of test development*. Routledge.

Lathrop, Q. N., & Cheng, Y. (2017). Item cloning variation and the impact on the parameters of response models. *Psychometrika*, *82*(1), 245-263. https://doi.org/10.1007/s11336-016-9513-1

Li, X. (2024, April). *Math item difficulty modeling using complexity design layers and item metadata* [Paper presentation]. National Council on Measurement in Education, Philadelphia.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*.

Lorge, I., & Diamond, L. K. (1954). The value of information to good and poor judges of item difficulty. *Educational and Psychological Measurement, 14*(1), 29-33.

Luecht, R. M. (2006, May). *Engineering the test: From principled item design to automated test assembly.* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Luecht, R. M. (2007, April). *Assessment Engineering in Language Testing: From Data Models and Templates to Psychometrics*. Invited paper presented at the Invited paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M. (2008, October). Assessment Engineering in Test Design, Development, Assembly, and Scoring. Invited keynote presented at the Annual Meeting of East Coast Language Testing Organizations (ECOLT), Washington, DC.

Luecht, R. M. (2009, June). *Adaptive Computer-Based Tasks Under an Assessment Engineering Paradigm*. Paper presented at the 2009 GMAC CAT Conference, Minneapolis, MN (Proceedings of the 2009 GMAC CAT Conference).

Luecht, R. M. (2012a). An introduction to assessment engineering for automatic item generation. In *Automatic item generation: Theory and practice* (pp. 196-216). Routledge, New York, NY.

Luecht, R. M. (2012b). Automatic item generation for computerized adaptive testing. In *Automatic item generation: Theory and practice* (pp. 196-216). Routledge, New York, NY.

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology*, *14*, 2-38. https://eric.ed.gov/?id=EJ1013205

Luecht, R. M. (2014a). Item Analysis. In B. Everitt & D. Howell (Eds.). *Encyclopedia Statistics in Behavioral Science* (pp. 958-967). John Wiley and Sons Ltd.

Luecht, R. M. (2014b). Luecht, R. (2014). In S. Davis-Becker & C. Buckendahl (Eds). *Testing in the Professions*. New York: Wiley & Sons.

Luecht, R. M. (2016). *Normalized scale score differences*. Greensboro, NC: University of North Carolina at Greensboro. Unpublished technical note (retrieved 16-Feb 2024,

https://www.dropbox.com/s/8yn4edfkhkgzlpk/NormalizedDifferencePlots_LuechtApr2016.pdf?dl=0).

Luecht, R. M. & Ackerman, T. A. (2018). A technical note on IRT simulation studies: Dealing with truth, estimates, observed data, and residuals. *Educational Measurement: Issues and Practice, 37(3)*, pp. 65-76.

Luecht, R. M., Burke, M., & Devore, R. (2009, April). *Task Modeling of Complex Computer-Based Performance Exercises.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Luecht, R. M., Dallas, A., & Steed, T. (2010, April). *Developing Assessment Engineering Task Models: A New Way to Develop Test Specifications*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Luecht, R., & Burke, M. (2020). Reconceptualizing items: From clones and automatic item generation to task model families. In H. Jiao & R. W. Lissitz (Eds.), *Application of artificial intelligence to assessment* (pp. 25–49). Information Age Publishing, Inc.

Luecht, R.M. (2024, April). *Maintaining Score Scales Using Item Families* [Paper presentation]. National Council on Measurement in Education, Philadelphia.

Magis, D., Raiche, G., & Béland, S. (2012). A didactic presentation of Snijders's l(z)* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37*, 57-81.

Markle S., M., & Tiemann, P. W. (1970). *Really understanding concepts.* Champaign, IL: Stipes.

Masters, J. S. (2010). *A comparison of traditional test blueprint and item development to assessment engineering in a licensure context*. Unpublished doctoral dissertation. University of North Carolina at Greensboro, NC.

Masters, J. S. (2010*). A Comparison of Traditional Test Blueprinting and Item Development to Assessment Engineering in a Licensure Context.* [Unpublished doctoral dissertation]. University of North Carolina at Greensboro.

McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., & Settles, B. (2021). Jump-Starting Item Parameters for Adaptive Language Tests. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 883–899. https://doi.org/10.18653/v1/2021.emnlp-main.67

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement*, 4th edition (pp. 257-306). Washington, DC: American Council on Education.

Mislevy, R. J., & Haertel, G. D. (2007). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6-20. https://doi.org/10.1111/j.1745-3992.2006.00075.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Item Generation for Test Development* (pp. 97-128). Mahwah: Lawrence Erlbaum Associates.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, *1*(1), 3-62.

Mislevy, R. J.; & Riconscente, M. M. (2006). Evidence-centered assessment design. In S.M. Downing & T. M. Haladyna (Eds.). *Handbook of Test Development* (pp. 61-90). Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R.J., Wingersky, M.S. and Sheehan, K.M. (1994). Dealing with uncertainty about item parameters: expected response functions. *ETS Research Report Series*, i-20. https://doi-org.libproxy.uncg.edu/10.1002/j.2333-8504.1994.tb01601.x

Mitchell, K. (1983). *Cognitive processing determinants of item difficulty on the verbal subtests of the Armed Services Vocational Aptitude Battery and their relationship to success in army training.* Unpublished doctoral dissertation. Cornell University.

Mosier, C. I., Myers, M. C., & Price, H. G. (1945). Suggestions for the construction of multiple-choice test items. *Educational and Psychological Measurement, 5*(3), 261-271.

Nathan, M. J., & Koedinger, K. R. (2000). An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction, 18*(2), 209–237.

Nathan, M. J., Koedinger, K. R., & Alibali, M. W. (August, 2001). Expert blind spot: When content knowledge eclipses pedagogical content knowledge . In L. Chen (Ed.), *Proceedings of the Third International Conference on Cognitive Science* (pp. 644 – 648). Beijing: University of Science and Technology of China Press.

OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Osburn, H. (1968). Item sampling for achievement testing. *Educational and Psychological Measurement*, *28*(1), 95-104. https://doi.org/10.1177/001316446802800109

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Practical considerations in computer-based testing. New York: Springer-Verlag.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. In *Elements of adaptive testing* (pp. 151-165). Springer, New York, NY.

Schmeiser C. B., Welch C. J. (2006). Test development. *Education Measurement*, 4, 307–353.

Sebrechts, M. M., Enright, M., Bennett, R. E., & Martin, K. (1996). Using Algebra Word Problems to Assess Quantitative Ability: Attributes, Strategies, and Errors. *Cognition and Instruction*, *14*(3), 285–343. http://www.jstor.org/stable/3233651

Settles, B., T. LaFlair, G., & Hagiwara, M. (2020). Machine Learning–Driven Language Assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247–263. https://doi.org/10.1162/tacl_a_00310

Sheehan, K. M., & Mislevy, R. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27*, 255–272

Shu, Z., Burke, M., & Luecht, R. M. (2010, April). *Some Quality Control Results of Using a Hierarchical Bayesian Calibration System for Assessment Engineering Task Models, Templates, and Items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Singley, M. & Bennett, R. (2002). Item generation and beyond: applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp.361-384). Lawrence Erlbaum Associates.

Sinharay, S., & Johnson, M. S. (2012). Statistical modeling of automatically generated items. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 183–195). New York, NY: Routledge.

Sinharay, S., Johnson, M. S. & Williamson, D. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics, 28*(4), 295–313.

Susanti, Y., Tokunaga, T., Nishikawa, H., & Obari, H. (2017). Controlling item difficulty for automatic vocabulary question generation. *Research and Practice in Technology Enhanced Learning*, *12*(1), 25. https://doi.org/10.1186/s41039-017-0065-5

T. Raykov, & G.A. Marcoulides (2010). *Introduction to Psychometric Theory*. New York, NY: Taylor & Francis.

Tendeiro, J.N., Meijer, R.R., & Niessen, A.S.M. (2016). PerFit: An R Package for Person-Fit Analysis in IRT. *Journal of Statistical Software, 74*(5), 1-27. doi:10.18637/jss.v074.i05

Tian, C., & Choi, J. (2023). The impact of item model parameter variations on person parameter estimation in computerized adaptive testing with automatically generated items. *Applied Psychological Measurement*, *47*(4), 275-290 https://doi.org/10.1177/01466216231165313

Tinkelman, S. (1947). Difficulty prediction of test items. *Teachers College Contributions to Education.*

van Krimpen- Stoop, E., & Meijer, R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327-344

Vector Psychometric Group (2021). FlexMIRT® 3.6.4. [Computer software]. Chapel Hill, NC: [Author].

Von Davier, M. (2023). Training optimus prime, M.D.: Generating medical certification items by fine-tuning OpenAI's gpt2 transformer model. *Advancing Natural Language Processing in Educational Assessment*, 90-106. https://doi.org/10.4324/9781003278658-8

Wainer, H. (2002). On the automatic generation of test items: Some whens, whys and hows. In S. H. Irvine, & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 287-316). New Jersey: Lawrence Erlbaum Associates.

Walker, H. L., Ghani, S., Kuemmerli, C., Nebiker, C. A., Müller, B. P., Raptis, D. A., & Staubli, S. M. (2023). Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *Journal of medical Internet research*, *25*, e47479. https://doi.org/10.2196/47479

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models [Preprint]. Retrieved from https://arxiv.org/abs/2201.11903

Whitney, S., Ferland, A., Lawler, M., Spikings, D., & Resanovich, M. (2024, April). *Developing complexity design layers and scoring protocols for math items* [Paper presentation]. National Council on Measurement in Education, Philadelphia.

Wilmurth, G., Richardson, J., Withycombe, A., & Johnson, J.L. (2024, April). *Defining text complexity, item complexity, and the text-by-item interaction for reading assessment* [Paper presentation]. National Council on Measurement in Education, Philadelphia.

Wright, B. & Stone, M. (1979). Best test design. MESA Press: Chicago, IL