View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by The University of North Carolina at Greensboro

BRUCIA, ROBERT CHARLES, Ph.D. Operationalizing Item Difficulty Modeling in a Medical Certification Context. (2020)
Directed by Dr. Richard Luecht. 134 pp.

This research study modeled item difficulty in general pediatric test items using content, cognitive complexity, linguistic, and text-based variables. The research first presents an introduction which addresses the current shortcomings found in item development and alternative methods such as principled assessment design which aim to address those shortcomings. Next, a review of the literature is presented which addresses traditional item development, item development using cognitive demands, item difficulty modeling, and the Coh-Metrix (Grasser et al., 2004) linguistic tool. The methods section outlines how content, cognitive, linguistic, and text-based variables were defined and coded using both subject matter experts (SMEs) and Coh-Metrix web-based software. The methods section goes on to outline the backward multiple regression analysis which was conducted to determine the proportion of variance in Rasch item difficulty accounted for by the defined variables and a study which can be used to demonstrate the impact of the current findings on examinee ability calibration. The results of the study demonstrate an operationalizable process for determining item difficulty variables. The results also found that Rasch item difficulty was significantly predicted by five item difficulty variables which accounted for .324 variance in Rasch item difficulty. The research concludes with a discussion of the findings, including steps that can be taken in future studies to build upon the current research and results.

OPERATIONALIZING ITEM DIFFICULTY MODELING

IN A MEDICAL CERTIFICATION CONTEXT

by

Robert Charles Brucia

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
2020

Approved by

_____

Committee Chair

To my wife, Eva, and three children, Sophia, Benny, and Vincent.

My world and motivation.

APPROVAL PAGE

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

Committee Chair _____

Committee Members _____

_____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I first would like to thank Dr. Richard Luecht for guiding me through this process with his vast knowledge of both technical and real-world information which allowed me to navigate the inevitable hurdles faced throughout this journey. I would also like to thank the entire faculty of the Educational and Research Methodology department, who helped me learn and grow in the many areas covered throughout this program.

I would also like to thank those serving on my plan of study and dissertation committee: Dr. Robert Furter; Dr. Robert Henson; Dr. Kyung Yong Kim; and Dr. John Willse. Each of the members of my committee have provided valuable advice and recommendations along the way which has made this research a success. A special thank you is necessary for Dr. Robert Furter, who works alongside me at the American Board of Pediatrics, and who has always had an open door and open ear for any questions or ideas that I have had.

I thank Dr. Linda Althouse, Dr. Andrew Dwyer, and Jared Riel, for supporting me in completing this program and with my dissertation research. Additionally, I owe a debt of gratitude to the test development staff at the American Board of Pediatrics, particularly Tracy Lorg, who helped fill in for me during the many times I could not be present when attending class.

Finally, I would like to acknowledge my loving family, who have supported me throughout this process and who have, and will continue to be, the motivation for my past, current, and future successes.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Traditional item development has remained largely unchanged for decades despite high costs and less than optimal item discard rates due to poor statistical performance. Three notable advantages which may be gained if test developers can discover what drives an item's statistical performance are: 1) a reduction in cost per item by lowering item discard rates and eliminating the need for costly pilot testing, 2) increased precision when making high-stakes pass or fail decisions by using items with difficulty levels which coincide with the difficulty level found at or around the cut score, and 3) producing a more robust validity argument by tying variables of the items to the ability elicited from a test taker when answering an item correctly. Despite these advantages, many testing organizations remain a prisoner of the traditional methods of item development where an "artistic" (Millman and Greene, 1989) approach is taken in which subject matter experts (SMEs) write items focused mainly on assigned content areas. While there are obvious obstacles for organizations in redesigning item development methods, including managing change and creating buy-in from stakeholders, the advantages to adopting a more evidence-based approach to item development should compel test developers to explore item difficulty modeling and ways to incorporate new item development approaches.

Traditional item development has been used by a majority of testing organizations for the past several decades. This method of item development is outlined in *The Handbook for Test Development* (2015) which provides twelve steps for effective test development. These steps include creating a test specification, item development, test design and assembly, and scoring test responses. The first step in traditional item development is developing a test specification (sometimes referred as a blueprint or test content outline), which outlines the content areas and the number of items from each area that will appear on a test. The content which is outlined addresses the areas of knowledge and skill that a testing organization would like to assess an examinee on. Once the test specification is developed, SMEs are trained on item writing guidelines and best practices, which have been developed over time. The SMEs are then assigned content areas in which to write items to. The table titled "A Revised Taxonomy of Multiple-Choice Item Writing Guidelines" is provided in *The Handbook for Test Development* (2015) and consists of 28 guidelines which are split into four categories which include content, formatting concerns, style concerns, and the options. The guidelines which specifically impact item difficulty are outlined in Table 1. While the failure to follow formatting and style guidelines can impact item difficulty, issues in these areas are often corrected with professional editing and thus not included in Table 1.

Table 1. Handbook of Test Development (2006) A Revised Taxonomy of Multiple-Choice Item Writing Guidelines

Category	Guideline
Content	1. Every item should reflect specific content and a single specific mental behavior. 2. Avoid trivial content.

	3. Use novel material to test higher level learning. 4. Keep the content of each item independent. 5. Avoid overly specific and overly general content. 6. Avoid opinion-based items. 7. Avoid trick items. 8. Keep vocabulary simple and appropriate.
The options	1. Keep options homogenous. 2. Keep the length of options equal. 3. Avoid giving clues to the correct option. 4. Make all distractors plausible. 5. Use typical errors of students to create distractors.

After reviewing the cited guidelines, one can see that SMEs are still afforded a great deal of flexibility in crafting items due to the basic and general nature of the guidelines. Following item writing best practices such as those in Table 1 is seen as a means for reducing construct irrelevant variance, which is when an examinee may answer an item correctly or incorrectly, not due the knowledge or skill in which the item intends to assess, but due to a flaw in the test item itself. It is expected that between the SMEs' expertise of the content, and following best practice "rules of thumb," high quality test items will be produced. Although most organizations have SMEs code each item to a cognitive level (recall, application, analysis, etc.), test items developed using this method are typically written with a focus on content, with the cognitive level in many cases assigned after the item is written. It is a common belief among SMEs that items crafted using this approach provide unique, relevant, and realistic real-world scenarios for the test taker. In this less controlled environment, it is not uncommon for SMEs who are assigned the same content area to produce items that are very different in both "look and feel," and in statistical performance. These handcrafted items are viewed by SMEs as providing "face validity," or the appearance to the examinee and stakeholders that the test

and test items appear to measure what they are meant to. While "face validity" is not recognized or cited in modern validity literature, the concept remains important (to stakeholders, SMEs, and examinees) in high stakes testing where the costs for taking an exam can be substantial. Using this approach to item development, SMEs typically spend one to two hours crafting each item. When SMEs are paid for their work, this time expense proves costly, and even when SMEs are volunteers, this time expense can contribute to burnout. Once items are written using this method and are subsequently approved for use, they often require pilot testing, which is a costly process used to collect psychometric statistics. If the items perform outside of acceptable psychometric difficulty parameters, they must be either deleted, or revised. If an item is revised, it must then be pilot tested again to determine if the revisions had a positive impact on the item's statistical performance.

Two disadvantages of the traditional item development approach are high costs due to item discard rates and the inability to determine item difficulty without the use of pilot testing. Due to known item discard rates, a common rule used by testing organizations is to develop three times the number of items needed to develop an examination, which further adds to the cost of this item development method. The 2019 ABP General Pediatrics (GP) certification exam consisted of two 335-item forms and 732 unique items. Of the 732 items, 440 items were newly approved and had not been previously tested. Of the 440 new items, 159 (36 percent) were flagged for poor statistical performance. A breakdown of the flagging criteria of the 159 poor performing items is provided in Table 2.

Table 2. 2019 General Pediatrics Poor Performing Items

Number of New Items	Percent of New Items	Statistical Flagging Criteria
135	31%	Too easy (p-value > .95)
10	2%	Too difficult (p-value < .35)
14	3%	Negative discrimination
159 (Total)	36% (Total)	

Case et al. (2001) found that using the traditional item development approach allowed for only 55 percent of items to be kept after pilot testing. Their approach included the traditional activities of: an in-person training on item writing; review, revision, and approval of the items by SMEs; cleanup of the items by a professional editor; and retention of SMEs for a three-year period. The cost of each item approved for live use in the Case et al. (2001) study was $111. While the cost of items produced using the traditional approach may vary, more recent quotes that have been cited include $1400 per item (the National Registry of Emergency Medical Technicians) and $1500 to $2500 per item (Rudner, 2009). When items are deleted due to poor performance or flaws in the content, the testing organization essentially receives no end-product despite the time and resources which went into creating the item.

The problem of uncertainty regarding an items difficulty prior to pilot testing has been examined for decades, and several past studies have been conducted to determine if SMEs are able to determine item difficulty based only on item content. Tinkelman & Sherman (1947) and Lorge & Kruglov (1954) were the first experts who attempted to predict item difficulty using SMEs as judges of items. Both studies found that while the judges were able to predict relative item difficulty (ordering of items), the judges were

not able to accurately predict absolute item difficulty (percent of examinees correctly responding to an item). One can conclude from these studies that while SMEs are able to craft high-quality items from a content perspective, they are unable to craft items which account for item difficulty drivers. Thus, traditional item development relegates testing organizations to a "shotgun" type approach, with the statistical performance of items not being known until pilot testing is complete. Using this "shotgun" approach, it is not uncommon for items produced using traditional item development to come back with a wide range of item difficulties and discriminations. Due to the wide variance in item difficulty using traditional item development, acceptable item p-values often range from .35 (35 percent of examinees answering an item correctly) to .95. Essentially, organizations are faced with choosing between 1) expensive pilot testing or 2) accepting that the shotgun spread of difficulties may or may not provide sufficient item acceptance rates and precision at the cut score.

Due to the disadvantages of traditional item development and the motivation of test developers to improve assessments, the concept of evidence-based and cognitively focused item development approaches has recently emerged. Many of the new assessment design approaches fall under the umbrella of principled assessment design (PAD), with evidence-centered design (ECD) outlined by Mislevy et al. (2003) serving as the foundation for this family of assessment and item development. While traditional item development can be viewed as an art (Millman & Greene, 1989) that creates items which are unique entities (Luecht & Burke, 2019), PAD approaches use a scientific-based method for engineering items towards intended interpretations and uses (Nichols et al.,

2017). Nichols et al. (2017) describe five foundational elements for PAD item development which are: (1) clearly defined assessment targets; (2) a statement of intended score interpretations and uses; (3) a model of cognition, learning, or performance; (4) aligned measurement models and reporting scales; (5) and the manipulation of assessment activities to align with assessment targets and intended score interpretations and uses. These elements allow PAD to achieve more coherence than traditional item development, linking the different phases of development rather than treating each phase as a separate entity. PAD focuses on the construct being measured and score inferences, interpretations, and uses throughout the item development phase. Once the construct is clearly defined and intended inferences and score interpretations are clear, PAD focuses the development of items on theories of learning and cognition. This differs from traditional item development which mainly focuses on content.

While one may cite the various advantages of PAD methods across the spectrum of assessment development activities, the entirety of PAD is outside the scope of this paper, and this research primarily focuses on the activities of PAD which apply to item development. Mislevy et al. (2003) categorized item development activities into three parts: the student model, or "What we are measuring;" the evidence model, or "How we are measuring it;" and the task model, or "Where we measure it." Mislevy et al. (2003) defined the student model as "defining one or more variables related to the knowledge, skills, and abilities one wishes to measure," the evidence model as "providing detailed instructions on how one should update the student model variables given a performance in the form of the examinees' work products from tasks," and the task model as

"describing how to structure the kinds of situations one needs to obtain the kinds of evidence needed for the evidence models." Combined together, these elements should lead test developers and SMEs to produce more cohesive items which are better able to tie the performance on the items to the interpretations and uses of the assessment.

Another PAD method which offers an alternative approach to item development, is assessment engineering (AE), which uses task models and task model grammars (or item templates) to generate test items (Luecht, 2013). AE creates test items using four stages which are (1) construct mapping, (2) task modeling, (3) design and use of item templates, and (4) items (Luecht, 2013). Construct mapping allows the test developer to conceptualize (and visualize) the construct in an ordered fashion where proficiencies are placed along a complexity scale which aligns with the score scale to be used (Ferrara, et al. 2017). Prior to task model creation, the test developer creates descriptions of an examinee's abilities, proficiencies, knowledge, and skills at different points of the construct and score scale. Next, task models, or specifications of item families, are created which target the different descriptions of proficiency which were defined in the previous step. Within each task model, numerous items can be generated using task model grammars, which are "programmable specifications for generating items in the same family so that they are isomorphic in terms of cognitive complexity" (Ferrara, et al. 2017). This item development approach is much more controlled than the traditional approach, and items sharing the same content specification should not perform statistically different, which is commonly seen with traditionally developed items.

As one can see, the role of SMEs in PAD item development is reduced in quantity, but not in importance. SMEs are essential in assisting the test developer in the creation of the task models and task model grammars. While this is much different than their role in traditional item development, this responsibility should be viewed as of equal or greater importance since each task model and task model grammar is used to create large amounts of high-quality test items. Once the task model grammars are created, a SME or an automated item generation (AIG) software can use the grammars to create test items in a tightly controlled environment.

PAD methods such as ECD and AE both allow for the disadvantages of traditional item development, cost and item discard rates, to be addressed. Unlike traditional item development, which can take one to two hours to write each item, task models and templates allow for items to be generated rapidly, especially when AIG software is used. It should however be noted that the upfront time spent on creating the construct map and accurate task models and task model grammars neutralizes part of this perceived advantage. Additionally, item templates mostly eliminate the amount of time spent by SMEs and editorial staff to ensure items follow style and formatting guidelines. Task models and templates also allow for SMEs and test developers to reduce item discard rates, by providing more structured items which assess knowledge, skills, and abilities (KSAs) which are directly linked to the construct of interest and specifically those KSAs that are located on the construct map which are around the ability level that coincides with the exam cut score. While outside the scope of this paper, Furter (2015) demonstrated how PAD task models can be used in setting the standard for a certification

exam, which further demonstrates how PAD methods allow for more precise pass or fail decisions.

While PAD methods provide improvements for the noted shortcomings of traditional item development, various issues with these methods still require further research to achieve effective implementation which will allow for the full spectrum of the conceptualized PAD benefits to be realized. A current disadvantage of PAD item development is that it may not lend itself to content areas or proficiencies which are more difficult to write items to, or at the very least, require a lot of time and resources to create a quality task model grammar which can account for the different variables present in these items (an example may be a clinical pediatric test item). Additionally, the templated items lose the unique scenarios that traditional item development provides. Finally, while PAD eliminates much of the time that SMEs spend on item writing (and any reimbursement that goes along with that time), and creates more items while simultaneously reducing item discard rates, SMEs are still needed to review items generated by the task model grammars (cost of travel, lodging, etc. for item review meetings would remain the same) and editorial staff are still required to edit the items.

Item difficulty modeling can be seen as a bridge to PAD for testing organizations currently using traditional item development methods. Item difficulty modeling research has been conducted going back to the work completed by Tinkelman & Sherman (1947) and Lorge & Kruglov (1954) where SMEs were used as the source of prediction. The notion of construct representative research and task decomposition was introduced in the late 1970's in an effort to tie performance on test items to underlying theoretical variables

using the linear logistic trait model (LLTM) and the multicomponent latent trait model (MLTM). Pellegrino et al. (1980) and Whitley et al. (1981) demonstrated how the LLTM could account for theoretical "complexity factors" in an item and how the scoring of an item could account for these factors. Whitley (1980) put forth the MLTM which built on the LLTM by introducing latent trait models which account for subtasks and alternative methods for solving an item such as guessing or cluing in the item stem. Embretson (1984) unified the LLTM and MLTM when proposing the general multicomponent latent trait model (GLTM) to understand item tasks with "multiple information outcomes and processing complexity factors" and allowing "complexity factors to have effects on component information outcomes rather than the total item response."

While the Embretson (1984) research was a step forward in item difficulty modeling, the results showed that additional complexity factors within items were needed to be defined. These early works served to demonstrate the importance of linking item difficulty to underlying variables which explain test taker's responses to the items and served as a foundational element for the more recent PAD element of task modeling. Since those early studies, numerous researchers have attempted to model item difficulty by accounting for theoretical variables which explain item responses, and a shift towards using ordinary least squares regression (OLS) and tree-based regression (CART) to explain the amount of variance in item difficulty that can be attributed to the theorized variables has been made (see Kirsch& and Mosenthal, 1991; Sheehan & Mislevy, 1994; Sheehan, 1997; and Gorin & Embretson, 2006). Recent studies have had varying success reporting R-squares (explained variance in item difficulties) ranging from .07 to .90 (see

Shaftel et al., 2006; and Enright et al., 2002). Most recently, Qunbar (2019) used text complexity and machine learning to model item difficulty in a medical certification context. The results of this study revealed that a prediction model using text complexity as an independent variable was not able to provide above chance predictions about an item's difficulty. In summary, while the research into item difficulty has been robust, test developers and researchers are still in search of an operationalizable process to uncover the theoretical variables, tasks, and underlying components of an item which can then be generalized across testing populations. Further, research is still needed to demonstrate not only how item difficulty can be modeled, but also how that information can be used by test developers for future item development. While nearly all of the item difficulty modeling studies thus far have approached the subject in a post-hoc manor (on items which were developed using traditional methods), a gap still remains on how understanding what drives item difficulty can be implemented at the item writing stage of test development to produce higher quality items and exams.

The purpose of this study was to create a workable and repeatable process that allows test developers to determine the variables within a test item which drive the item's difficulty so that the variables may be accounted for in future item development with the use of a regression equation where item difficulty is the dependent variable and the difficulty drivers are the independent variables. The variables that will be explored will address both the linguistic features and the content and cognitive aspects of an item. The linguistic variables will be determined using Coh-Metrix (Grasser et al., 2004) software which is an online platform that analyzes text and codes 108 different linguistic variables

which range from basic word, sentence, and paragraph counts to more complex features such as lexical diversity and temporal cohesion. The content and cognitive variables will be determined using SMEs and "think aloud" discussions to explore and define variables such as content difficulty and the number of cognitive steps required to arrive at the correct answer. The goals for defining significant difficulty drivers in test items and creating a regression equation which accounts for those difficulty drivers are:

1) To reduce the time spent on future item development by providing templates or instructions which account for difficulty drivers,
2) To reduce or eliminate the need for pilot testing by designating a predicted item difficulty for each item,
3) And to reduce the item discard and revision rate due to newly written items which poorly perform during live or pilot testing.

This study demonstrates how significant item difficulty drivers in items can be defined and a regression equation can be formed which allows for the accurate prediction of Rasch item difficulty. The methods section outlines how the items and their predicted Rasch item difficulties can be used as anchor items in the calibration of examinee ability levels (thetas) to determine the impact of using items with predicted Rasch item difficulties in an operational setting. Additionally, this research demonstrates a process for determining item difficulty drivers which can be implemented by test developers in different contexts to achieve the previously stated goals and benefits.

Continuing on, the literature review (Chapter II) provides a robust summary of the previous literature on: traditional item development and its shortcomings; accounting for

cognitive demands in item development; item difficulty modeling; and determining and accounting for linguistic and text-based variables using Coh-Metrix software (Grasser et al. 2004). Next, the methods section (Chapter III) outlines how the current research was conducted and how the proposed methods were implemented to achieve the desired outcomes. The results section (Chapter IV) provides the results from the implemented research methods. Finally, the discussion (Chapter V) outlines the study significance, limitations, and future considerations, which were determined using the results from the current research.

CHAPTER II

LITERATURE REVIEW

The following literature review contains four main sections. The first section examines the literature supporting traditional item development in certification and licensure exams which primarily focuses on item content and item writing best practices rather than the cognitive processing that an examinee would demonstrate in order to answer an item correctly. Section two will address the literature which is focused on cognition and the role of cognitive demands in item development. This section will include literature on evidence-centered design methods which encourage test developers to go beyond the content focused approach to item development in order to incorporate an evidence-based and cognitively focused approach. The third section of the literature review provides an in-depth look into previous research studies which have examined item difficulty and the different variables which factor into item performance. The fourth and final section will review the literature on linguistic features in test items and the impact that linguistic features may have on an item's difficulty.

Traditional Item Development

Traditional item writing practices have been in place for decades, focusing on best practices, item writer training, and a content centered approach, yet minimal changes have been made to the best practices put forth by Mosier et al. (1945) and Ebel (1951).

While changes to item writing best practices have been few and far between, content validity and content-related validity evidence, as described in the different editions of the Standards of Educational and Psychological Testing (here on referred to as the Standards), has transitioned from a major source of validity evidence to one of many supporting pieces of evidence in a validity argument.

Content validity first came about in the 1950's as one of three areas of test validity (the other two being criterion validity and construct validity). The original version of the Standards (1954, titled Technical Recommendations for Psychological Tests and Diagnostic Techniques), put forth recommendations for the dissemination of information which is encouraged to be distributed in a manual that accompanies the test. Content validity was originally defined in the Standards (1974) as "an aspect of validity that is required when the test user wishes to estimate how individual performances in the universe of situations that the test is intended to represent." Cronbach (1971) cautioned that content validity should be "restricted to the operational, externally observable side of testing, and not used for judgement on a subject's internal processes." A content model of validity was outlined by Guion (1977) with three main stipulations: observed performances are a representative sample from the content domain; observed performances are evaluated fairly; and the sample of observed performances is large enough to control for sampling error. Messick (1989) criticized content validity stating that while content validity can be used as support for an instrument representing the domain of relevance, it cannot be tied to the interpretation of test scores. As validity moved to a unified and construct centered approach, the term content validity was

removed from use and content-related validity evidence was substituted in its place. Beginning with the 1999 edition of the Standards, the term content validity was abandoned in favor of the term "content-related evidence." While this change may seem minor to a layman, it shows the shift from the view of test content being a primary argument for validity to a view of content-related evidence being only one piece of many in a validity argument. In the most recent version of the Standards (2014), evidence based on test content is only one of five major sources of evidence. The most recent Standards (2014) address content-related validity evidence in standard 1.1 which states "when the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with references to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent." Content-related validity evidence is further addressed (in regards to workplace testing and credentialing) by the Standards (2014) with standard 11.3 which states "When test content is a primary source of validity evidence in support of the interpretation for the use of a test for employment decisions or credentialing, a close link between test content and the job or professional/occupational requirements should be demonstrated."

The other source of validity evidence in the Standards (2014) are evidence based on response processes (most pertinent to the current research), evidence based on internal structure, evidence based on relations to other variables, and evidence for validity and consequences of testing. While an in-depth review of validity, and how views on validity have morphed throughout the years, is out of the scope of the current literature review, it

is important to note the shift from content being a primary factor of a validity argument to one of many pieces of evidence for validity. Likewise, it is important to note the inclusion of evidence based on response processes in the 1999 and 2014 version of the standards, as this evidence supports the need for examining response processes that may be responsible for item difficulty performance. The Standards (2014) addresses the importance of interpreting a construct based on assumptions tied to cognitive processes of examinees. The Standards (2014) states that "theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers." The Standards (2014) goes on to state that "questioning test takers from various groups making up the intended test-taking population about their performance strategies or responses to particular items can yield evidence that enriches the definition of the construct." Thus, it can be assumed that that the recommendations on response processes put forth by the Standards (2014) should not be employed only after the item development phase is complete, but throughout all phases of item development.

Mosier et al. (1945) was the first to put forth suggestions for constructing multiple-choice test items and defined the task of item writing as "phrasing a question in such terms that all prospective examinees understand the task set; those who have the requisite degree of knowledge and will give the intended answer; and all who do not, and will give another answer." While Mosier et al. (1945) and subsequent literature providing guidance for multiple choice items lacks in direction for directly accounting for the difficulty level of items, there is guidance provided on the linguistic features of items.

Two specific areas of guidance provided by Mosier et al. (1945) related to the linguistic features of multiple-choice items are:

1) Can modifying phrases, qualifications, etc. be removed from an item without impacting responses?
2) Are item ideas stated clearly, "with the answer an important part of the statement – not buried at the end of a preposition in a parenthetical clause?"

A final important take away from the Mosier et al. (1945) suggestions is the placement of responsibility of item quality on the item writer which is shown in the concluding statement which states that when an item writer submits an item, if the writer cannot "honestly predict" that an item will discriminate between qualified and unqualified examinees, then the item requires further revisions until that prediction can be made.

Ebel's chapter of *Educational Measurement* (1951) built on the original suggestions Mosier et al. (1945) provided and expanded on the Mosier et al. (1945) suggestions to item types other than multiple choice. Ebel begins the chapter with a view that is still agreed with by many in testing which is that item writing is an "art" and "essentially creative." The longevity of this view is demonstrated with Millman et al. (1989) citing the artistry of item writing nearly 40 years after Ebel's chapter was first published. The difference between the current research and PAD item development approaches with Ebel's approach is emphasized when Ebel (1951) states "just as there can be no set of rules for producing a good story or painting, so there can be no set of rules that will guarantee the production of good test items." The systematic and evidence-based approaches used by PAD methods such as AE directly contradict the notion that a

set of rules cannot be used to produce well performing test items. While Ebel's views differ greatly from the current PAD item development approaches, he does describe (perhaps unintentionally) an important weakness of using his prescribed item development which is the gap between a test content outline and the lack of guidance that a test content outline provides for the individual items that will be written by an item writer. Ebel (1951) goes onto state that the responsibility of developing an idea into an item that addresses a topic on the test content outline is placed on the item writer. Indeed, the responsibility, and freedom, for SMEs to generate ideas is presented as one of the most important concepts in item writing by Ebel (1951), so much so that Ebel dedicates an entire subsection of the chapter on item writing to "ideas for test items." Ebel (1951) continues on with the majority of the chapter dedicated to "suggestions for item writing." The Ebel (1951) general item writing suggestions and multiple-choice specific item writing suggestions are shown in Table 3.

Table 3. Ebel (1951) Item Writing Suggestions

Item writing suggestions - general
Express the items clearly
Use words with precise meaning
Avoid complex word arrangements
Include of all information needed to correctly respond to the item
Avoid unimportant ("nonfunctional") words
Avoid "unessential specificity" in item stems and response choices
Avoid "irrelevant inaccuracies"
Adapt the level of difficulty of the item to the test taker
Avoid clueing the correct response
Avoid "stereotyped phraseology"
Avoid "irrelevant sources of difficulty."
Item writing suggestions specific to multiple choice items
Use a direct question or incomplete statement

Include in an item stem any words that will be repeated in all option choices
Avoid negatively worded stem queries
Provide a response that is agreed upon as correct by SMEs
Make all response choices appropriate and plausible
Avoid “high technical” distractors
Avoid overlap in response options
Arrange responses in logical order
If an item addresses knowledge of a definition, include the term to be defined in the stem and the option choices for the correct definition in the response options
Avoid response options which present as true or false statements

A criticism that may be made of the Ebel (1951) suggestions and related to the current research is that the chapter’s earlier emphasis on the artistic and creative individual idea generation used by SMEs make his eighth general suggestion (adapt an item’s difficulty to the test taker) inherently difficult, since unique items with individual ideas are less systematic and thus will have wider ranging performance statistics that are more difficult to predict. This “unpredictability” is further highlighted in next chapter (Conrad, 1951) titled “The Experimental Tryout of Test Materials.” While an in-depth review of this chapter will not be provided in this section, it should be noted that Conrad (1951) provides seven purposes for pilot testing items with the first three addressing the need to determine problematic items, item difficulty, and item discrimination. These reasons for pilot testing contend that even if item writing suggestions such as those put forth by Ebel (1951) are closely followed, an item’s performance statistics still remain largely unpredictable until pilot testing is complete. As previously stated, and as the subsequent literature to be reviewed shows, different approaches for item difficulty modeling can be used by putting forth rules (which Ebel did not think were possible) in order to guide item development so that item performance may be predicted prior to pilot testing.

Haladyna et al. (1989) conducted a comprehensive review of previous item writing literature in order to produce a comprehensive set of rules for SMEs. Haladyna et al. (1989) created a taxonomy consisting of three areas for item writing rules to be classified under. Those three areas were (1) general item writing, which consists of procedural and content concerns, (2) stem construction, and (3) option development, which consists of the correct answer and the item's distractors. Haladyna et al. (1989) compiled a list of 43 item writing guidelines and used a group of guideline authors to rate each rule for importance and worthiness of inclusion in the list. The authors also designated if a guideline was "testable" or "value." Testable components could be tested to determine if the rule had impact on an item's statistical performance whereas values are those suggestions that cannot be tested but are deemed important by testing professionals. Haladyna et al. (1989) note that an important aspect and distinction of their research is the expert consensus rating given to each guideline. The authors do however note, that while their list is the most comprehensive source of item writing guidelines (at the time of publication), there remains an "urgent need" for additional research on item writing best practices.

Lane et al. (2015) outlines a comprehensive plan which addresses traditional item development in *The Handbook for Test Development* (2015). In the "twelve-steps for effective test development," step 2, content definition, step 3, test specifications, and step 4, item development, all address methods for item development which have been traditionally used for high stakes certification and licensure item development. Both Clauser et al. (2006) and Fein (2012) describe traditional methods specifically as the

methods which apply to certification and licensure testing in fourth edition of *Educational Measurement* (2006). The foundation for item development is formed by first creating a content definition which is used to create a test content outline or test specification. The test content outline, once created, is used by test developers and SMEs to guide the content and number of items which will make up the exam. In the first step, content definition, Lane et al. (2015) states "The validity of inferences for achievement test scores rests primarily and solidly on the adequacy and defensibility of the methods used to define the content domain operationally, delineate clearly the construct to be measured, and successfully implement procedures to systematically and adequately sample the content domain." In certification and licensure testing, defining the content domain is mainly achieved by the practice or job analysis process. Raymond et al. (2015) state that the purpose of a practice or job analysis is to "identify the job responsibilities of those employed in the profession." Raymond et al. go onto state that once the job responsibilities are known, a list of KSAs required for effective performance of the identified job responsibilities can be created, and those KSAs can serve as the basis for a test content outline or specification document. Both Raymond et al. (2015) and Clauser et al. (2006) state that most often, a group of SMEs who fill either the job being defined, or supervise the job being defined, are used in this process. Clauser et al. (2006) caution that "considerable work" remains after the list of KSAs for the job responsibilities is created. These considerations include using surveys to determine the frequency and criticality of each KSA to determine both the content that should be tested on and how many items should address each content area (content weightings). Fein (2012) describes how content

weightings are determined. Fein (2012) first states that using survey results, point values are assigned to KSAs based on difficulty, importance, and frequency. The values are then averaged and converted into a proportion which indicates "the proportion of content of the exam that should be covered by each element of content associated with a specific task." This process bridges the content definition phase to the test specification or test content outline phase. Indeed, the outcome of the content definition phase is the production of a document (test content outline/specification) which is used by test developers for item and test production.

Raymond et al. (2015) provide an alternative approach for creating a test content outline which includes six steps, and those steps are: (1) acquire relevant documentation; (2) obtain input from SMEs, (3) SMEs develop first draft of outline; (4) develop second draft; (5) review topics for job relevance; and (6) assign content weights. Raymond et al. (2015) caution however that "while this general approach works well in practice, one limitation of relying exclusively on SME panels is that the test plan may reflect little more than conventional wisdom and may include KSAs that really are not required for public protection but appear because of tradition."

Clauser et al. (2006) discuss content outlines and item classification stating that a common strategy is to use a matrix style approach to classify items to both a content area and to either (or both) a cognitive level or a task. Clauser et al. (2006) go on to describe different approaches to classifying items by cognitive level, with the most common being the use of Bloom's taxonomy (Bloom et al., 1956), which contains six cognitive levels ranging from knowledge (the most basic) to evaluation (the most complex). The authors

go onto to state that a "simpler" approach is to only use two cognitive levels, which are "recall of an isolated fact" or "application of knowledge." Regardless of the cognitive level rating system used, Clauser et al. (2006) encourage the use of "scenario-based" questions to accomplish several goals which include: avoiding exam items which only assess an examinee on "unimportant facts;" creating items with scenarios that a practitioner would be expected to solve in practice; and creating items which "have an appropriate level of factual knowledge while also requiring examinees to analyze factual situations and apply knowledge of facts to solve problems." Clauser et al. (2006) also discuss the classification of items to an examinee task. This approach ensures that items will not only assess factual knowledge, but also assess an examinee's ability to execute a task which would be expected once factual knowledge is demonstrated. Clauser et al. (2006) state that the challenge with task classification is to ensure the tasks are both relevant and those that would be expected of a new practitioner. Once a test content outline is developed, item development may begin. Using traditional methods, SMEs are provided content areas, and in a matrix style content outline, tasks or cognitive levels, and tasked with writing items which will assess examinees on the assigned areas. Fein (2012) and Baranowski (2015) cover the methods used in traditional item development (both item writing and item review). Baranowski (2015) begins with the important statement which grounds traditional item development (and was first stated by Cantor, 1987) which is that "item writing is frequently referred to as an art." Baranowski goes on to pose several questions regarding traditional item development such as "What

constitutes a high quality test item?" and "Do we know one when we see one or do we need performance statistics to determine one?"

Cognitive Demand in Item Development

In the book *Developing and Validating Test Items*, Haladyna et al. (2013) tie the cognitive demand of test items to content-related validity evidence. Haladyna et al. (2013) define content as "knowledge, skills, and abilities" and cognitive demand as "the expected mental complexity involved when a test item is administered to a typical test taker." The authors go on to note that if cognitive demand can be determined, test items will be able to focus on "exactly" what the construct represents. Haladyna et al. (2013) caution that defining cognitive demand may prove difficult due to variability in test taker process. This variability, which should be expected, makes pinpointing a single cognitive demand for test items difficult. Using "think aloud" methods with test takers would be advantageous to determine item cognitive demands, however this venture could be costly, and a more cost-effective approach that test developers employ is the classification of test items to a cognitive taxonomy. The most notable taxonomy which is used in practice comes from Bloom et al. (1956). Bloom's taxonomy consists of six levels which are (from lowest cognitive demand requirement to highest): (1) knowledge, (2) comprehension, (3) application, (4) analysis, (5) synthesis, and (6) evaluation. Haladyna et al. (2013) note that Bloom's cognitive taxonomy lacks validation and has proven "inadequate," citing several studies on classification consistency, test taker responses, and critical analysis of the taxonomy. The literature cited shows inherent weaknesses of classifying items to the taxonomy to make the argument that an item tests a single

cognitive demand for all test takers. Haladyna et al. (2013) propose a simplified approach for item cognitive demands, basing their proposed taxonomy on knowledge, skills, and abilities. Table 3.5 from Haladyna et al. (2013) is pictured below and provides a summary for the author's simplified cognitive rating proposal.

Figure 1. Haladyna et al. (2013) Summary of Cognitive Demands for Knowledge, Skills, and Abilities

Table 3.5 Summary of Cognitive Demands for Knowledge, Skills and Abilities

Cognition	Types	Demands
Knowledge	Fact, concept, principle, procedure	Recall/recognize Comprehend/understand Application
Skill	Mental, physical	Recall/recognition of procedure for performing skill Comprehension/understanding of procedure for performing the skill Performing the skill
Ability	Collection of structured and ill-structured tasks	Use knowledge and skills in the performance of each task

Gorin (2006) addresses how cognitive models can be used in the item writing process and encourages a shift to a more "scientific" approach. Gorin (2006) cites "increasing pressure to extract meaningful information about student skills and knowledge from item responses" as a reason for this shift. While the current research will focus on multiple choice items which are written by SMEs, Gorin (2006) posits that incorporating cognitive models into the item writing process has become easier to achieve with the use of innovative item types and automatic item generation. If cognitive models can be created, item generation can be employed to create items through the use of templates rather than the traditional approach of using SMEs for item writing. Gorin (2006) points out that item generation using cognitive models not only has psychometric

benefits, but also has economic benefits for testing organizations. Perhaps the most useful contribution of Gorin (2006) is the implications for practice section which encourages test developers "think outside the box" to "create construct definitions that are informative for item development" using techniques such as interviews. Gorin (2006) states that "test developers must consider even more rigorous methods of item examination before operational use that provides explicit evidence regarding the skills, knowledge, and processes measured by the items." Gorin (2006) addresses several aspects for incorporating cognition in the design of both tests and items. An important contribution of Gorin (2006) are the recommendations provided by the author for the development of tests which address content that has not previously been studied using cognitive approaches and item difficulty modeling. Gorin (2006) emphasizes the benefits of better defining constructs and using cognition as a driver for item development by citing several previous authors who state that "construct definitions including descriptions of individual cognitive processes and hypothesized relationships among processes can provide a stronger foundation for test development and score interpretation (Embretson, 1994; Mislevy, 1994; Messick, 1995)." While Gorin (2006) addresses several important benefits for defining a construct, including construct mapping (Wilson, 2004), these methods fall outside the scope of the current proposed research in that they require the development of a test, and the items which make up the test, from the ground up. Gorin (2006) does however provide several useful tools for researchers to examine the cognitive aspects found in operational test items. Gorin (2006) encourages the use of the qualitative collection method of "think alouds" which require examinees to verbalize

their thought processes when responding to an item. In the initial step for item difficulty modeling, Gorin (2006) proposes hypothesizing "skills, knowledge, and processes" which are required to answer the item correctly. Indeed, Gorin (2006) cites (Bejar, 1991, and Bennett, 1999) when stating "the key to item difficulty modeling is to identify the relevant features that drive item processing and to estimate their impact." Gorin (2006) does caution that verbal protocols such as "think alouds" may be subject to the examinee having a different interaction with the item while "thinking aloud" that they would while simply taking the test item during an examination setting. Despite this shortcoming, Gorin (2006) notes that the information gained from verbal protocols can lead to the discovery of processing components which were not previously known or hypothesized. Gorin (2006) also notes that the use of verbal protocols may allow test developers to determine examinee response strategies which could then be accounted for in accounting for variance in item difficulty. A final, and important note by Gorin (2006), is that "test developers must consider even more rigorous methods of item examination before operational use that provides explicit evidence regarding the skills, knowledge, and processes measured by the items" and that "item design should proceed from sources of cognitive complexity related to the construct of interest, rather than unrelated surface features." The methods proposed in chapter 3, such as "think alouds," can be used to determine those unrelated surface features which may be impacting difficulty without providing information on the knowledge or skill that the item is intended to assess. This aspect also speaks to one of the shortcomings of items written by SMEs using the "artistic" approach which lends itself to introducing irrelevant content or surface features

within items which may impact performance but do not provide information on the construct of interest. Gorin (2006) concludes by emphatically stating that "the future success of cognitive-based test development depends heavily on the ability of test developers and practitioners to learn and adopt the methods based on cognitive psychology."

Graf et al. (2005) set out to create a set of item models which would generate multiple-choice items that perform psychometrically equivalent to each other. The study used a retroactive approach, with previously created items with ideal statistics, as "sources" to base item models on. Graft et al. (2005) state that "an important aspect of item model development is to capture students' approaches to solving problems and to represent common misconceptions among the options." After the item models, and items were generated, the authors used the item statistics to analyze the cognitive aspects of the items. The authors note that it is difficult in practice to identify which item features impact difficulty and the extent of influence on difficulty that these features have. Graft et al. (2005) report that determining a cognitive framework is "necessarily complicated" and not "feasible" in an operational setting. In the study, the authors presented a "key model, distractor model, and option model." These features are important in that the authors found that the distractor models accounted for additional variability among the generated items. The authors do however caution that the analysis of distractor model impact was conducted retroactively using item performance statistics, and without those statistics, the differences may not have been recognizable due to the similarity of the distractor models. Heeding that caution, the use of similar item models may not produce

the previously cited benefit of cost savings from reduced or eliminated pilot testing. Graft et al. (2005) also caution that the findings with the distractor models concerned a specific item model, and that distractor models could not explain variance in item performance when applied to other developed item models in the study. Graft et al. (2005) conclude their study by citing the need for more effective and efficient methods for creating item models.

Fulkerson et al. (2011) conducted a study on the cognitive processes which SMEs experience when creating items. The authors cited the numerous studies on item writing as it relates to test taker response processes and the lack of research on the SMEs' cognitive processes when writing test items. The research was unique in that it not only reported on the cognitive aspect of items from the writer's development process, but also reported several useful strategies to determine those cognitive processes such as "think aloud" methods using story boards. The authors defined a storyboard as "a written description of the narrative, images, animation, and/or video that will be developed for a test scenario." Fulkerson et al. (2011) cite the three important "phases" which take place during item writing and which were reported by Fulkerson et al. (2009). The three phases include the (1) initial representation phase, (2) the exploration phase, and (3) the solution phase. While the research was conducted from the item writer perspective, the approaches for determining the cognitive processes can be utilized for studies such as the proposed research which approach items from the reviewer perspective. Fulkerson et al. (2011) used a training session with study participants prior to the first "think-aloud" session to provide information on the research, the scientific evidence that the research

was based on, and the upcoming "think aloud" tasks. Following the training session, "think aloud" sessions were held with single individuals and lasted one hour each. Fulkerson et al. (2011) reported that during the "think aloud" sessions, writers were asked to respond to a writing assignment and verbalize their cognitive information while writing. To capture data during the "think aloud" sessions, transcripts were converted into statements which were coded as "categories of revised problem-solving cognitive processes" and "categories of requisite knowledge structures." The researchers also reported an interrater agreement to demonstrate the reliability of the coders. Fulkerson et al. (2011) provided two tables, one on item writer cognitive process categories, and one on knowledge structure categories, which are useful for future researchers to consider when examining the inner workings of test items. The notable cognitive process and knowledge structure categories reported on by Fulkerson et al. (2011) that can guide the analysis of previously written items along with the definitions provided by Fulkerson et al. (2011) are:

1) Schema activation – application of mental structures drawing on experience
2) Operator – active searching for content and solutions
3) Extraneous – information that is irrelevant to the item
4) General and pedagogical content knowledge – domain specific knowledge and instructional practices.

Another unique aspect of this research is the comparison of experienced (more than one year of item writing experience) and novice SMEs. These differences may be applicable when determining the cognitive process an entry level examinee undergoes when sitting

for a licensure or certification exam. Fulkerson et al. (2011) conclude that novices "spend more of their writing time defining the task and evaluating ways to select and sequence assessment content." This conclusion may be applicable to how entry level examinees approach a test item, by first determining a task and then selecting and sequencing the content provided in the item to solve that task, but further research is needed to support this possibility.

Recent research conducted by Lenzer et al. (2016) focused on cognitive pretesting of items, and while the research was based on survey items, their methods can be explored for viability in multiple-choice test items. Lenzer et al. (2016) list the goals of cognitive pretesting as determining item comprehensibility, problems within the items and the causes of those problems, and identifying possible improvements for the items. The authors list four questions which cognitive pretesting can aim to answer about items. The four questions posed by Lenzer et al. (2016) are:

1) How do participants interpret the items and/or terms within the items?
2) How do participants retrieve information and/or events from memory?
3) How do participants arrive at a response?
4) How do participants assign their internally determined responses to an actual item response?

Additionally, Lenzer et al. (2016) provide five methods for obtaining cognitive pretesting data which include (1) "think aloud" techniques, (2) probing techniques, (3) paraphrasing, (4) confidence ratings, and (5) sorting. The purpose of the "think aloud" technique is to have respondents talk out their thought processes as they respond to an

item. The authors suggested asking respondents “While you are answering the following question, can you tell me what you are thinking?” The probing technique is described by the authors as asking follow-up questions based on a response. Four types of probing can be used which include comprehension probing (understanding of the question), category (or response selection) probing (reasoning behind the respondents answer choice), information retrieval probing (what information was retrieved from a respondent’s memory before selecting their response), and general/elaborative probing (explanation of answers or thought processes). Lenzer et al. (2016) describe paraphrasing as asking respondents to restate questions in their own words in order to gain a better understanding on the understanding of the question. Confidence ratings are used by the authors to determine if respondents are confident in their responses, and more importantly if they are not confident, the reasons for the lack of confidence. Finally, the authors describe the sorting technique as having respondents sort terms of items in either their own determined categories or a list of categories provided by the authors. Lenzer et al. (2016) recommend cognitive interviews take place in a quiet space, and if possible, the interviews be voice recorded (or even video recorded to determine any visual cues that the respondents put forth). Further recommendations for cognitive interviews provided by Lenzer et al. (2016) include conducting between five and 30 interviews with durations between 60 and 90 minutes. Analyzing data from cognitive interviews and methods for conducting the analysis are also addressed by Lenzer et al. (2016). Quantitative analysis can be conducted using a coding scheme for the responses with three steps of coding: open coding, which is the coding of responses by topic or categories; axial coding, which

is the process of determining group differences from the open coding; and selective coding, which is identifying "subordinate" topics which connect the open coding categories. In the selecting coding process, Lenzer et al. (2016) state that the researcher must "formulate a hypothesis that describes the phenomena that an item captures." Lenzer et al. (2016) caution readers that there are no universal rules to how cognitive pretesting should be completed, and their work only provides guidance and suggestions for conducting cognitive pretesting. The authors go on to state that the techniques chosen by a researcher when conducting cognitive pretesting should be determined by both the interest of the researcher and the behavior patterns of the respondents.

The work by Mislevy & Haertel (2006) on ECD has served as a basis for many of the modern principled and cognitive focused assessment design approaches being used and explored today. The ECD approach to test design is rooted in creating assessments which are based on evidence which may be used to create a validity argument. ECD, as described by Mislevy & Haertel (2006), is a layered approach, which allows for assessments to be created using layers which build upon one another to create the final assessment product. The layers are all dependent on each other and each subsequent layer must use the foundational elements from the previous layers in order for the assessment to achieve a robust validity argument. The five layers of ECD are (1) domain analysis, (2) domain modeling, (3) conceptual assessment framework, (4) assessment implementation, and (5) assessment delivery. Table 1 from Mislevy & Haertel (2006) is pictured below, and outlines each assessment layer's role, key entities, and selected knowledge representations.

Figure 2. Mislevy & Haertel (2006) Layers of Evidence-Centered Design for Education Assessments

Table 1. Layers of Evidence-Centered Design for Educational Assessments

Layer	Role	Key Entities	Selected Knowledge Representations
Domain Analysis	Gather substantive information about the domain of interest that has direct implications for assessment; how knowledge is constructed, acquired, used, and communicated.	Domain concepts, terminology, tools, knowledge representations, analyses, situations of use, patterns of interaction.	Representational forms and symbol systems used in domain (e.g., algebraic notation, Punnett squares, maps, computer program interfaces, content standards, concept maps).
Domain Modeling	Express assessment argument in narrative form based on information from Domain Analysis.	Knowledge, skills, and abilities; characteristic and variable task features, potential work products, potential observations.	Toulmin and Wigmore diagrams, PADI design patterns, assessment argument diagrams, "big ideas" of science.
Conceptual Assessment Framework	Express assessment argument in structures and specifications for tasks and tests, evaluation procedures, measurement models.	Student, evidence, and task models; student, observable, and task variables; rubrics; measurement models; test assembly specifications; PADI templates and task specifications.	Algebraic and graphical representations of measurement models; PADI task template; item generation models; generic rubrics; algorithms for automated scoring.
Assessment Implementation	Implement assessment, including presentation-ready tasks and calibrated measurement models.	Task materials (including all materials, tools, affordances); pilot test data to hone evaluation procedures and fit measurement models.	Coded algorithms for rendering tasks, interacting with examinees and evaluating work products; tasks as displayed; IMS/QTI representation of materials; ASCII files of item parameters.
Assessment Delivery	Coordinate interactions of students and tasks: task-and test-level scoring; reporting.	Tasks as presented; work products as created; scores as evaluated.	Renderings of materials; numerical and graphical summaries for individual and groups; IMS/QTI results files.

A notable contribution of ECD and the link between ECD and cognitive processing and item difficulty modeling, is the accounting for both characteristic and variable features of test items. Characteristic variables are the parts of test items which directly address the knowledge or skill in which the item aims to assess. Variable features are defined by Mislevy & Haertel (2006) as those item features which the test developer may manipulate to affect an item's difficulty. Mislevy & Haertel (2006) directly link ECD to the cognitive demand literature when they state that the "domain modeling layer is important for

improving the practice of assessment, especially for the higher-level reasoning and capabilities for situated actions that cognitive psychology call to our attention." At the domain modeling stage of ECD, the KSAs must be explicitly stated, and characteristic and variable features defined, so that item's may be formed which account for these layers during the item development or conceptual assessment framework layer.

Item Difficulty Modeling

Going as far back as 1947, educational measurement experts have attempted to predict and explain item difficulty in an attempt to create valid and reliable assessments. Tinkelman & Sherman (1947) and Lorge & Kruglov (1954) were the first experts who attempted to predict item difficulty using SMEs as judges of items. Both studies found that while the judges were able to predict relative item difficulty (ordering of items), the judges were not able to accurately predict absolute item difficulty. While these studies and their methods of prediction were simplistic in design, they provided the basis for an important (and complex) concept that measurement experts are still attempting to explain today. If item difficulty can be accurately predicted, the need for costly pilot testing may be reduced or eliminated entirely. Another benefit from the ability to predict item difficulty is increased measurement accuracy by focusing items on important points of a score scale. Ferrara, Steedle, & Frantz (2018) provided a robust summary of the different studies and models that have been used to predict item difficulty. A noticeable gap in the current literature is item difficulty studies conducted for complex high-stakes certification exams in scientific fields such as medicine. Qunbar (2019) conducted the first study for modeling item difficulty in a medical certification context with results indicating the need

for further studies to be conducted in this context. The proceeding literature reviewed analyzes some of the various methods and concepts which have been used to predict item difficulty. The studies chosen implemented a range of methods and modeling approaches across several different assessment contexts for predicting and explaining item difficulty. Additionally, some studies were chosen based on their success in predicting item difficulty, with both successful and unsuccessful studies being reviewed so that lessons may be learned from both. The different types of explanatory variables used by the studies include content related variables, combinations of skills, cognitive features, and linguistic patterns. For each study, the statistical or theoretical model used, variables explaining item difficulty, and assessment context and generalizability is discussed. As research continues to advance in this area, understanding and building upon prior studies will be critical for measurement professionals to determine methods which are accurate, generalizable, and provide answers to the questions first posed by Tinkelman & Sherman (1947).

Embretson & Kingston (2018) modeled item difficulty using a straightforward approach and their results supported the notion that the items performed in a predictable fashion. Embretson & Kingston (2018) used two forms of item modeling, family variant (similar in difficulty and cognitive complexity) and structurally variant (reduced in difficulty and cognitive complexity), to generate items that were tested on a 7th grade mathematics achievement test. Additionally, items were generated using these item modeling approaches for both a 6th and an 8th grade mathematics achievement test. The process began with a panel of SMEs selecting previously used items and creating a

family variant item model, substitution variables, and constraints. Embretson & Kingston (2018) state that "adding context directly increases translation difficulty and may also increase integration difficulty, due to the added working memory burden." This was used as the study's rationale for creating the structural variant item models by "removing irrelevant sources of cognitive complexity." By removing these "irrelevant sources," the items produced from the structural variant item models were predicted to perform at an easier difficulty than both the family variant item models and operational items which were used to create them. The results showed that the predicted difficulty levels were accurate for both item model variants. Three statistical methods were used to evaluate the items: classical test theory (CTT), the 2PL item response theory (IRT) model, and the generalized linear model (GLM). The operational items and family variant model items did not have significant differences in difficulty showing that controlling for item difficulty through substitution variables was successful. The structural variant model items were found to have a significant difference in item difficulty for both the 6th and 7th grade items. The GLM, holding item identity and family membership as fixed factors, revealed that the family variant and structural variant item models were a significant predictor of item difficulty. While these findings were encouraging, Embretson & Kingston (2018) highlight in the discussion that the items created had similar content levels, appearance, and syntax to the operational items.

Sheehan (1997) put forth a non-parametric tree-based approach (TBA) to item difficulty modeling which aims to "model the complex non-linear ways in which skills interact with different item features to produce changes in item difficulty." In the TBA

(also referred to as a Classification and Regression Tree Analysis or CART) once a difficulty model is created, it is translated into a proficiency model which aligns with different sets of skills that are demonstrated when answering an item correctly. Sheehan (1997) does not provide a method for generating test items but instead intends to explain item difficulty to provide diagnostic information to test takers and users. The TBA in this study utilized two inputs: a vector of IRT item difficulty estimates, and a matrix of hypothesized skill classifications. The matrix of skill classifications was created using studies of factors which affect item difficulty, SMEs, and an analysis of domain tasks. In the TBA, the value of a response is regressed onto different sets of predictor variables (or sets of skills). A loss function is then used to create an interval band which explains the sets of skills that can be assumed to be mastered when a response vector is located within the interval band. Items which require the same or similar sets of skills to answer them correctly are grouped into "schemas." Items that require the same skills may be grouped into different schemas if the skills need to be applied in a different way. Once the schemas are developed, a computer algorithm further splits the items by skill classifications into clusters, and then into even smaller subsets called "nodes." The TBA used also allowed for a manual step called "pruning" where one may manually collapse nodes if it makes practical sense or if the differences in skills are minor and thus would provide minimal information if split. Sheehan (1997) used the SAT I verbal reasoning test to validate the TBA. The "schemas" used in this study were vocabulary in context; main idea and explicit statement; inference about an author's underlying purpose, assumptions, attitude, or rhetorical strategy; and application or extrapolation. The choice

of using the SAT I verbal reasoning test allowed the TBA to be used for items of several different complexities and which require different sets of skills. This aspect of the TBA is an advantage when compared to the previous study where the items were very similar to each other. The model tested against a best- and worst-case scenario. In the worst-case scenario, each item only assesses a single skill, and there are no clusters or nodes. In the best-case scenario, each item assesses a unique combination of skills. In this study, the SAT I verbal reasoning items were clustered, and the cluster model sum of square residuals were used to evaluate the amount of variance in student response vectors that was explained by the clusters. The item difficulty modeling using the TBA was successful as the model using eight clusters explained 90% of the variance in response vectors and the model using nine clusters explained 91%. Sheehan (1997) noted that the TBA allows for combinations of skills to be analyzed, which in turn allows for the approach to be taken with higher order questions. Additionally, whereas the linear model requires each skill or variable to have the same impact on item difficulty, the TBA models how the combination of skills impact difficulty. Finally, the discussion outlines what is perhaps the most influential advantage of the TBA which is "allowing for skill mastery to be transformed from an unobservable trait to an observable trait."

Two research studies conducted in 1990 used data from the 1985 National Assessment of Educational Progress (NAEP) to analyze how underlying variables explain variance in scores. Kirsch & Mosenthal (1990) used the variables and types of variables outlined in Table 4.

Table 4. Variables Underlying the Performance of Young Adults (Kirsh & Mosenthal, 1990)

Variable Type	Variable
Document (structural complexity and length)	Number of organizing categories*
	Number of embedded organizing categories
	Deepest level of embedded organizing categories
	Number of specifics**
	Number of embedded specifics
	Deepest level of embedded specifics
Task (relationship of information in question and information in document)	Number of organizing categories required by task**
	Deepest level of embedded organizing categories required by task
	Number of specifics required by task**
	Deepest level of embedded specifics required by task
Process (strategies for using a document to answer items)	Degrees of correspondence**
	Type of information**
	Plausibility of distractors

**Kirsch & Mosenthal defined organizing categories as "the highest unit of analysis, consisting of a generalized term or category that serves to summarize or synthesize specific information."*

Kirsch & Mosenthal (1990) used both correlations and a regression analysis to identify which variables were significant in explaining a respondent's total score. The study found that the total amount of variance accounted for by the five significant variables was 89%. While the results of this study provide several contributions related to document literacy, the notable contribution for the context of this paper is how the significant variables for task and process may help measurement professionals in explaining item difficulty. For task variables, both the number of organizing categories required by task and the number of specifics required by task were found to be significant. The regression coefficients for

both variables reveal what one may expect, in that a participant score decreases as either organizing categories required by task or number of specifics increases (β = -3.93 and -3.75). Other significant variables found in this study were degrees of correspondence and types of information. Kirsch and Mosenthal (1990) define degrees of correspondence as a variable which "deals with relation between given information in a question and corresponding information in the document" and define types of information as "primarily focusing on type of information and how it refers to the processes necessary to generate requested information based on one or more nodes from a document's information hierarchy." The regression coefficients for these variables are also informative, as they show a participant score increases as these variables increase (β = -3.93 and -3.75). While this study does not directly apply to typical certification exam formats and items, the findings and documentation based on the outlined variables can be used to inform future studies aimed at predicting item difficulty.

Sheehan & Mislevy (1990) used the NAEP Document Literacy scale data to "describe a cognitive processing model for solving the exercises and a structure relating item parameters in the psychometric model to salient item features in the cognitive model." The authors first define distinct skills demonstrated by the survey responses which were document literacy, prose literacy, and quantitative literacy. The authors note the report completed by Kirsch & Jungeblut (1986) for its importance for recognizing the different "types and levels of skills adults use in their everyday interactions with printed materials." Sheehan & Mislevy (1990) note the shortcomings of IRT models which are their failure to account both the cognitive processes that examines demonstrate when

answering items correctly or incorrectly and the features of items which drive difficulty. Sheehan & Mislevy (1990) describe the cognitive model put forth by Kirsch & Mosenthal (1988) which hypothesized a cognitive model for document literacy which included a four-step cognitive processing model consisting of three levels of organization. This model, while hypothesized about the NAEP data and survey responses, is informative to how examinees may process item content regarding different areas of expertise. The four steps and three levels of organization hypothesized by Kirsch & Mosenthal (1988) are provided in Table 5.

Table 5. Kirsh & Mosenthal (1988) Cognitive Model for Document Literacy

Step	Description
1	Identify the information given and requested in the task directive
2	Search document until requested information has been located
3	Match the information provided in the document to the information requested in the directive
4	Determine whether the identified match adequately meets the criterion of the task
Level of Organization	Category title
1	Organizing
2	Specific (SPE)
3	Semantic feature

Kirsch & Mosenthal (1988) also identified three variables which were present in items and responses which were material variables, directive variables, and process variables. Both the material and directive variables in the research addressed content specific information regarding both information on medications and information on directives for

taking the medication. The process variables outlined by Kirsch & Mosenthal (1988) are important because they address the examinee process for responding to an item. The three process variables put forth were degree of correspondence, or "how explicitly the information requested in a question matches corresponding information in the text," type of information, or "the type and number of restrictive conditions that must be held in mind in identifying and matching features," and the plausibility of distractors, or distracting information in an item or document which may lead an examinee away from a correct response.

Sheehan & Mislevy (1988) proposed a method for accounting for examinee cognitive processing into a psychometric model which built upon previous work using the LLTM. This method was labeled as "a two-stage empirical Bayes regression model (EB)." Using the full EB model and accounting for variables set forth by the cognitive model for document literacy, Sheehan & Mislevy (1988) found an R-squared of .81 when applying the model to the NAEP survey items. Sheehan & Mislevy (1988) conclude their work by presenting a compelling argument for future research that ties cognitive features of items to item difficulty and psychometric models. Sheehan & Mislevy (1988) state that "it is increasingly recognized that mere high reliability coefficients do not guarantee a good test, nor do high predictive relationships guarantee a valid test." Sheehan & Mislevy (1988) go on to state that "the onus has been placed (appropriately!) upon the tester to demonstrate that the skills tapped in an educational test are in fact those deemed important to measure."

Embretson (1998) further contributed to the research on linking cognitive processing to test items, and further, test validation. Embretson (1998) provided two reasons why construct validation should be expanded which were advances in psychology as it relates to construct validity and 2) the concept of integrating "test design into the construct being measured." Embretson (1998) outlined the stages of cognitive design systems in the figure pictured below.

Figure 3. Embretson (1998) Cognitive Design Systems

Table 1
Cognitive Design Systems

Stage
Specify general goals of measurement
Construct representation (meaning)
Nomothetic span (significant)
Identify design features in task domain
Task-general features (mode, format, conditions)
Task-specific features
Develop a cognitive model
Review theories
Select or develop model for psychometric domain
Revise model
Test model
Evaluate cognitive model for psychometric potential
Evaluate cognitive model plausibility on current test
Evaluate impact of complexity factors on psychometric properties
Anticipate properties of new test
Specify item distributions on cognitive complexity
Distribution of item complexity parameters
Distribution of item features
Generate items to fit specifications
Artificial intelligence?
Evaluate cognitive and psychometric properties for revised test domain
Estimate component latent trait model parameters
Evaluate plausibility of cognitive model
Evaluate impact of complexity factors on psychometric properties
Evaluate plausibility of the psychometric model
Calibrate final item parameters and ability distributions
Psychometric evaluation
Measuring processing abilities
Banks items by cognitive processing demands
Assemble test forms to represent specifications
Fixed content test
Adaptive test
Validation: Strong program of hypothesis testing

Embretson (1998) aimed to control item stimulus properties (determined using the cognitive model) so that the properties could be manipulated to control the level of item difficulty while also eliminating irrelevant properties which may impact an item's difficulty. Embretson (1998) cited the theory proposed by Carpenter et al. (1990) in which examinees apply lower and higher-level relationships when responding to an item.

The greater the numbers and levels of relationships, the more memory capacity is required for an examinee to correctly answer an item. Embretson (1998) accounted for these relationships by developing "item structures" which control for the numbers and types of relationships found in a test item. The item structures were then used to generate individual test items using 22 objects and seven attributes. The distractors for each item were developed so that one or more objects or attributes shared with the stem information were incorrect. Embretson (1989) found that a proportion of .773 of difficulty in item variance could be attributed to the item structural model. Regarding the cognitive model, Embretson (1998) states that the "best prediction of item difficulty and response time was obtained by a model with a single variable to represent working memory load." To account for this variable, Embretson (1998) used a variable titled "relational level" consisting of a five-point scale. The 5-points used in the scale along with their corresponding variable code were: 1) identify; 2) pairwise; 3) figure addition/subtraction; 4) distribution of three; and 5) distribution of two. Embretson (1998) found that a proportion of .71 of item difficulty variance could be attributed to the relational level (or working memory load) variable. Thus, Embretson (1998) demonstrated how using a combination of item structures, objects, attributes, and a "relational level" variable, an item bank can be generated with items which address higher level abilities and perform psychometrically sound. Further, Embretson (1998) tied the cognitive design system approach for generating test items to construct validity by demonstrating that memory load can be used as a variable to link an item's demand on examinee cognitive processing to the item's difficulty level.

Gorin & Embretson (2006) state that while there have been numerous approaches to using cognitive modeling to predict item difficulty, "the methodological approaches are similar in that once the relevant strategies and knowledges are integrated into a cohesive cognitive model, related features of existing items can be quantified." Gorin & Embretson (2006) used a coding approach to create variables for GRE-V test items in an attempt to account for processing of reading comprehension test questions. The study aimed to build upon several previously studied difficulty and cognitive modeling approaches, including the incorporation of Flesch's (1948) reading grade level, Anderson's (1982) 4-point scale for distractor reasoning, Sheehan & Ginther's (2001) and Embretson and Wetzel's (1987) correspondence and item format variables. This study, unlike the prior studies, attempted to model items that feature longer passages by using two additional variables, one that accounted for passage-length and interactions with the variables from the other models, and one that accounted for decision-processing requirements for special formats (such as the use of the word "except" in an item query). Both an investigator and a natural language processor were used to code variables within the test items. While the Sheehan and Ginther (2001) model accounted for 25% of variance in item difficulty, and the Embretson & Wetzel (1987) model accounted for 28% of variance in item difficulty, the model in this study was able to improve upon both models and account for 34% of the variance in item difficulty. Gorin & Embretson (2006) conclude their paper by pointing out that the methods used were retrofitted to previously existing items and that an optimal approach is to first develop a model which

experiments with variables which may explain difficulty variance and then develop items around the experimental model.

Ferrara et al. (2018) conducted three studies using the CART model to analyze items for three different assessment contexts: (1) high school language arts, mathematics, science, and social studies; (2) grade 6 through 9 science and social studies; and (3) a national achievement test program. The studies used a combination of both hypothesized variables (item design, content, cognitive, and linguistic) and response demand variables (content, cognitive, and linguistic) which were previously used by Ferrara et al. (2011). Each variable type had a subset of specific variables which were defined for coding purposes. A group of professional SMEs coded items to the variables after receiving training on the definitions and application of the variables. An additional activity reviewing rater agreement and consensus was also conducted to ensure accurate coding of items. The CART model was chosen because it includes an importance statistic (how well variables act as predictors) and a non-parametric approach. Ferrara et al. (2018) also chose to use a bootstrap technique (random forest approach) and conditional R-squares to minimize bias and provide cross validation. The third study conducted used only importance statistics and was unique to the assessment context. For the first two studies, the initial results showed that the variable for item type and maximum points were the most important predictors of difficulty, as one would expect. In many assessments, the item type and maximum points per item are fixed, so these variables would not be used. Ferrara et al. (2018) chose to also provide the results when excluding item type and maximum points in an effort to show the impact of the other variables on item difficulty.

The variance in difficulty explained by content, cognitive, and linguistic demands was found to be significant (defined by the authors as >.10) for language arts R-squared = .44) and social studies (R-squared = .18) in study 1, and grade 4 social studies (R-squared = .19) and grade 5 science (R-squared = .13) in study 2. These results show that even when using a robust model (CART with bootstrapping) and quality variable coding process, explanation of a majority of variance is difficult to achieve. The discussion section is especially helpful to those considering similar studies. The authors point out the impact that item type and maximum points can have, and caution those who may use these as variables in future studies. Also highlighted is the lack in current literature on variables that can be generalized to other assessment contexts. The importance of modeling item difficulty is touched on and how results (when successful) can be used for operational activities such as training SMEs to target difficulty levels with their items. Finally, Ferrara et al. (2018) note that response demands which are broader (such as the cognitive level or the "depth of knowledge" demand), and therefore more generalizable to other assessment contexts, may be problematic in that they provide "general and confusing" information about specific items. Ultimately, Ferrara et al. (2018) acknowledge the difficulties experienced in achieving desirable and generalizable results for studies such as their own and encourage future attempts to model item difficulty.

Two recent dissertations completed by graduate students at the University of North Carolina, Greensboro, addressed item modeling and item difficulty. Masters (2010) set out to predict item difficulty using assessment engineering methods, and specifically task models and item templates in the context of an insurance licensure exam. This study,

like Embretson & Kingston (2018) used operational items and SMEs to determine a variable for knowledge objects in the items which explain item difficulty. In this study, an additional step of SMEs reviewing and rating the difficulty levels of the operational item options was also completed. Next, task models and item templates were developed using the knowledge object and distractor difficulty variables. The results showed that the items developed using the item templates fit the Rasch model better than the operational items when using the variation of infit and outfit statistics (.18 infit/.25 outfit for operational items and .05 infit/.06 outfit for templated items). The study showed that 11 of the 14 item templates produced items which met similar acceptance rates as the operational items. A final analysis of the items used SMEs to rate the templated item's based on frequency and importance, as well as to rate the items on a separate scale of distractor complexity. These difficulty ratings were correlated to the actual item difficulties (after testing) and it was determined that neither method was successful (both accounted for the explanation of less than 1 percent of variance in item difficulty. Qunbar (2019) used a machine learning approach to produce item predicting variables and model the relationship of those variables with the known item difficulties from the ABP GP certification exam. This study is unique in that it did not use SMEs when creating the variables and classifications which predict item difficulty. The study deployed word counts as the variable for predicting difficulty (including the word counts of item stems, keys, and distractors) using the linear least squares regression (LLS), principle components regression (PC), partial least squares regression (PLS), CART, and artificial neural network regression models. The results of the study showed that using item

representations based on word counts was not effective in predicting item difficulties, regardless of the statistical model used. It should be noted that a similar study by Rupp, Garcia, & Jamieson (2001) was able to successfully use item word counts (aggregated and not split out by stem, key, and options) to account for 31 percent of the variance in item difficulty, and the context of pediatric items in this study may have produced the less than desirable results.

As Ferrara et al. (2018) summarized, there has been no shortage in the pursuit for measurement professionals to explain and predict item difficulty. Various methods (automatic item generation, assessment engineering [and other principled assessment design methods], retrofitting of theoretical item difficulty variables, machine learning), assessment contexts (SAT, GRE, licensure and certification, grade-level subjects), and statistical models (Rasch/IRT, GLM, PLS, PC, CART, ANN) have been used in this pursuit. Regarding item difficulty studies, Ferrara et al. (2018) state that "the empirical literature is promising" and that "there is plenty of opportunity for improvements in theoretical development and empirical results." Building upon previous research to further the prediction and explanation of item difficulty was demonstrated in the work by Gorin & Embretson (2006). Research should continue to build upon prior studies, by theorizing and implementing new methods and models, and in certain instances combining previous methods and models with new experimental models.

Coh-Metrix Linguistic Variables

Grasser et al. (2004) define both cohesion and coherence as they apply to computational linguistics. The authors define the distinction between the two terms with

cohesion being “a characteristic of text” and coherence being “a characteristic of the reader’s mental representation of the text content.” While cohesion applies to actual parts of the text, such as words, or phrases, coherence applies to how a reader may apply knowledge and skills to a text in their understanding of it. Using the concepts of both cohesion and coherence, Grasser et al. (2004) created the web-based tool named Coh-Metrix. Grasser et al. (2004) define cohesion gaps as those areas of text which require readers to apply information which was previously learned or text which has been read. Perhaps the most intriguing ability of the Coh-Metrix software when applied in a context such as pediatric clinical vignettes is that it identifies and accounts (within the variables coded) for cohesion gaps. Indeed, Grasser et al. (2004) state that cohesion gaps “can be beneficial for high-knowledge readers because their knowledge affords successful inference making.” Grasser et al. (2004) go on to state that “these results highlight the importance of pinning down linguistic and discourse features of cohesion and of better understanding the properties of world knowledge.” Grasser et al. (2004) also point out the shortcomings of previously used readability formulas and note that such formulas fail to account for language features such as discourse components which can impact the readability of text and cause text to be more difficult to read and comprehend. The Coh-Metrix web-based software is a copy and paste application which analyzes and codes the text with 108 variables. The variables account for various textual features, ranging from simple counts of words and sentences, to more complex features, such as syntactic pattern density such as noun and verb phrase density. Grasser et al. (2004) conclude their work by stating that the Coh-Metrix tool should lead to new understanding of language

processing and that future research should allow to determine if the language processing components are appropriate for the those who are interacting with the text.

Conclusions from a Review of the Literature

The literature review conducted aimed to address several areas where the current research will build upon previously conducted research. Literature on traditional item development methods, and the shortcomings of those methods, was reviewed. This important first step demonstrated how the current research aims to build upon the traditional item development model, including best practices which were first put forth by Mosier (1945), and which are currently used by numerous high-stakes licensure and certification organizations including the ABP. Next, the past literature on cognitive demands was reviewed to understand how these demands can be accounted for in assessment and item design. This section reviews literature which serves as an important foundation to the current proposed research by addressing the current researches goal of further understanding the cognitive demands of pediatric certification items and how those cognitive demands impact an item's difficulty. Further, the review of cognitive demand literature shows the notable gap of past research in examining cognitive demand of advanced items, which contain clinical vignettes and assess high levels of problem solving, such as the pediatric certification items that will be studied in the current research. This section also highlights the methods employed by ECD, which demonstrate how test developers may use such studies as the current one to make a more robust validity argument. The third section in the literature review examined previous item difficulty modeling studies in order to demonstrate the different approaches that

researchers have taken to better understand item difficulty. Like the previous section, this review demonstrated the gap in this area of the literature as it pertains to higher level items like those that will be included in the proposed study and the lack of variables which may be generalized to constructs other than those which are included in the studies themselves. Finally, a brief review of the literature provided by the authors of the software Coh-Metrix demonstrated how the Coh-Metrix tool and variables examine linguistic processing. Research on the use of a software such as Coh-Metrix to examine linguistic variables in the context of high-stakes test items has yet to be completed. The reoccurring theme found throughout this literature review is that while immense work has been completed around item development, item difficulty modeling, understanding of how examinee's interact with items, and linking item features (including content and cognitive processing), with the ultimate decisions that are made from performance on the items, a gap still remains in examining these concepts in higher level items used for high-stakes decisions, such as a the ABP GP certification exam (and items).

CHAPTER III

METHODS

The ability to predict item difficulty without the need for field testing in high-stakes certification testing has various benefits which include a reduction in costly pilot testing, item discard rates, and SME time spent on item revisions due to poor performance. Currently, test developers are faced with the predicament of either (1) pilot testing every item in order to determine item difficulty or (2) decomposing items to determine objective indicators (measures) to predict item difficulty. A review of the literature has revealed two previously used methods to predict item difficulty which are (a) coding items to account for features in the items such as cognitive variables and (b) using statistical models such as the LLTM put forth by Fischer (1973) and the GLTM put forth by Embretson (1984). A notable gap in current research is the demonstration of a workable and repeatable process for item difficulty modeling in high-stakes certification and licensure items which can inform future item production. Creating a workable and repeatable process for predicting item difficulty would have several impactful benefits for certification and licensure organizations. If a successful process is discovered, test developers will be able to work with SMEs to code previously written but unused test items with difficulty predicting variables to estimate the difficulty of the items prior to testing. Test developers will also be able to train SMEs to account for difficulty predicting variables when they are writing and reviewing items. As previously cited, the

2019 general pediatrics exam had 33% of new items perform with a difficulty level which was either too high or too low, causing the items to either be discarded or revised by SMEs. With each ABP item costing approximately $3,500, the current poor performance rate, and subsequent item discard rate, is less than ideal. Accurately predicting Rasch item difficulty prior to field testing will reduce the current item discard rate at the ABP and produce a cost and time savings stemming from poor performing items. These benefits, which would be realized with a workable and repeatable process for item difficulty modeling, and the gap in literature defining such a process, motivated the current research and research questions.

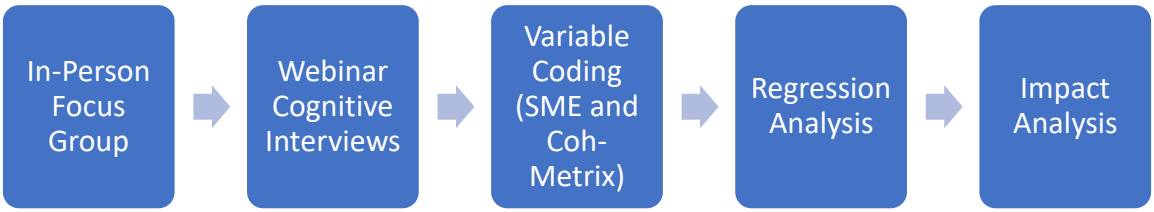
Research Questions

1) What operationally feasible process can be implemented that allows SMEs and test developers to code content- and cognitive-related variables, from previously used test items, in order to predict item difficulty modeling and produce cost and time savings due to poor item performance?
2) Can replicable cognitive complexity-oriented variables (metrics) be developed by two methods: (a) SMEs; and (b) computed linguistic/text-based-features variables derived from Coh-Metrix?
3) What proportion of Rasch item difficulty variance in pediatric certification items do variables defined by SMEs (content knowledge, cognitive process, reading comprehension, distractor relationship to the key, etc.) and variables provided by Coh-Metrix linguistic software account for both individually and jointly?

4) Using Pearson's correlation, what is the strength of the relationship between an examinee's ability level (theta) on a pediatric certification exam determined using 96 items with predicted Rasch item difficulties and an examinee's ability level determined by freely calibrating all items with live data?

To address the current research questions, a mixed methods exploratory sequential design was used. This research design was chosen as it uses two phases: (1) an exploratory qualitative phase and (2) subsequent quantitative phase which is informed and built upon by the exploratory qualitative phase (Watkins & Gioia, 2015). A flow chart outlining the research design and milestones is presented in figure 3.

Figure 4. Flow Chart of Research Milestones



During the exploratory qualitative phase, the researcher utilized SMEs to conduct a focus group and subsequent cognitive interviews to define content and cognitive variables which drive a test item's difficulty. The quantitative phase then used the results from the qualitative phase in combination with Coh-Metrix linguistic variables (108 variables) to create a regression equation which allows for the prediction of item difficulty prior to collecting live testing data. The item difficulty which will be used in the quantitative phase of the study (as the dependent variable) is the Rasch item difficulty, defined by the equation: $P = 1/ [1 + \exp (b-t)]$. The Rasch item difficulty model was chosen due to the

common use of the model in high-stakes certification testing (the model supports smaller sample sizes than the 2- and 3PL item response theory models) and the ABP's use of the Rasch model in current operational test items. The baseline used to determine the success of the regression equation was a proportion of .80 variance explained, which was recommended by Bejar (1983). A final quantitative component of the study demonstrated how a test developer can examine the impact of using predicted Rasch item difficulties by comparing examinee ability levels (thetas) calibrated using live testing data with examinee ability levels calibrated using predicted Rasch item difficulties.

Study Items

Three sets of ABP GP test items were used in the research study. First, a set of 12 test items was used to conduct a SME focus group. The set of 12 test items (6 pairs of items sharing classifications and universal tasks), were chosen as part of a larger set of items which shared classifications, but which had large differences in Rasch item difficulty. Next, a second set of 101 test items were used to conduct cognitive interviews with SMEs and to conduct a principle component analysis on 108 Coh-Metrix variables which were coded to the items. Finally, a third set of 96 test items were coded with both SME content and cognitive based variables and with Coh-Metrix principle component scores (derived using the analysis from the set of 101 test items). The third set of 96 test items were also used to create a regression equation for predicting Rasch item difficulty and the results from that equation were used for the impact portion of the study. Both the second and third set of test items were chosen due to (1) their inclusion on the 2017 General Pediatrics certification exam, (2) their close mapping to the published weightings

for the exam, and (3) their placement on a 2018 non-proctored exam which rendered the items exposed (and not confidential). While the items closely map to the published content outline weightings (domain and universal task), they do not match exactly due to the restrictions of items which were eligible for use in the study.

The ABP GP certification exam is a 335-item exam which is administered annually in October. The exam is created based on the GP test content outline which is published on the ABP website. The content outline consists of 25 content domains, with each domain containing up to four levels of more specific content areas within the domain. The content outline also outlines four universal tasks which define the way in which knowledge of a content area may be demonstrated in pediatric clinical practice. Both content domains and universal tasks have published weights, which were determined using a survey of practicing general pediatricians and a group of SMEs. Each item on the exam is classified to both a content area and universal task and each exam is built so that the 335 items on the exam represent the published content and universal task weightings.

Each test item used in the study was written, reviewed, and approved for use on the ABP GP certification exam by a committee of SMEs. Additionally, each item was professionally edited by ABP staff, and conforms to internal and AMA editorially guidelines. Each item was written to address a blueprint domain and universal task. Prior to the start of this study, a one-way analysis of variance (ANOVA) was conducted on all currently active general pediatric items (n = 2,419) to determine if either the coded content domain or universal task indicators have a significant effect on an item's Rasch

item difficulty. The results of the ANOVA were then used during the SME focus group to inform the SMEs of the effects of these previously coded variables during the variable creation process. Table 6 outlines the general pediatric blueprint (domains, universal tasks and definitions, and published weightings) and the corresponding number of items in each of the three sets of items used in the research study.

Table 6. Classification of Research Items

Domain	Content Area	Published Weighting	Number of Items: Set 1 (*n* = 12)	Number of Items: Set 2 (*n* = 101)	Number of Items: Set 3 (*n* = 96)
1	Preventative Pediatrics/Well-Child Care	8%	5	7	7
2	Fetal and Neonatal Care	5%	0	6	5
3	Adolescent Care	5%	0	4	4
4	Genetics, Dysmorphology, and Metabolic Disorders	3%	0	4	4
5	Mental and Behavioral Health	5%	0	6	6
6	Child Abuse and Neglect	4%	0	4	4
7	Emergency and Critical Care	4%	0	4	4
8	Infectious Diseases	7%	0	7	7
9	Oncology	2%	0	1	2
10	Hematology	4%	0	4	4
11	Allergy and Immunology	4%	0	4	4
12	Endocrinology	4%	0	4	3
13	Orthopedics and Sports Medicine	4%	0	4	4
14	Rheumatology	2%	0	2	1

15	Neurology	5%	0	6	6
16	Eye, Ear, Nose, and Throat	4%	0	4	4
17	Cardiology	4%	0	2	2
18	Pulmonology	5%	0	6	6
19	Gastroenterology	4%	2	4	4
20	Nephrology, Fluids, and Electrolytes	4%	0	4	4
21	Urology and Genital Disorders	3%	0	2	2
22	Skin/Dermatology	4%	0	4	4
23	Psychosocial Issues	2%	0	2	1
24	Ethics	2%	0	4	2
25	Research Methods, Patient Safety, and Quality Improvement	2%	0	2	2
Total		100%	12	101	96
Universal Task	Definition	Published Weighting	Published Weighting		Published Weighting
Basic Science and Pathophysiology	Understanding best practices, clinical guidelines, and foundational pediatric knowledge, including normal and abnormal function of the body and mind in an age specific development context	20%	0	6	6
Epidemiology and Risk Assessment	Recognizing patterns of health and disease and understanding the variables that influence those patterns	10%	0	3	2

Diagnosis	Using available information (e.g., patient history, physical exam) to formulate differential diagnoses, choose appropriate tests, and interpret test results to reach a likely diagnosis	35%	8	49	50
Management and Treatment	Formulating a comprehensive management and/or treatment plan, including reevaluation and long-term follow-up, taking into account multiple options for care	35%	4	43	38
Total		100%	12	101	96

There was an acceptable amount of variance in the Rasch item difficulties for both the training and study sets of items for the Coh-Metrix analysis and the regression equation to be successful. The Rasch item difficulties for both sets of items were determined by calibrating the items (along with the other items found on the 2017 General Pediatrics exam) using the live data from approximately 2,200 examinees. The variance of the items in the training set (101 items) was 2.07 (SD = 1.44). The variance of the items in the study set (96 items) was 2.21 (SD = 1.49). The histograms in figure 4 and figure 5 show the distribution of Rasch item difficulties for each set of items.

Figure 5. Distribution of Training Set Rasch Difficulties (Set 2)

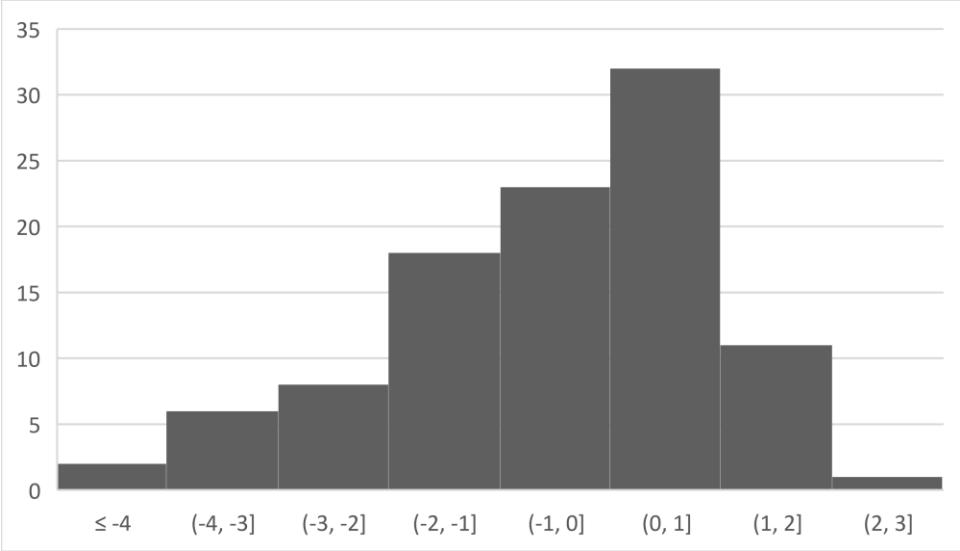
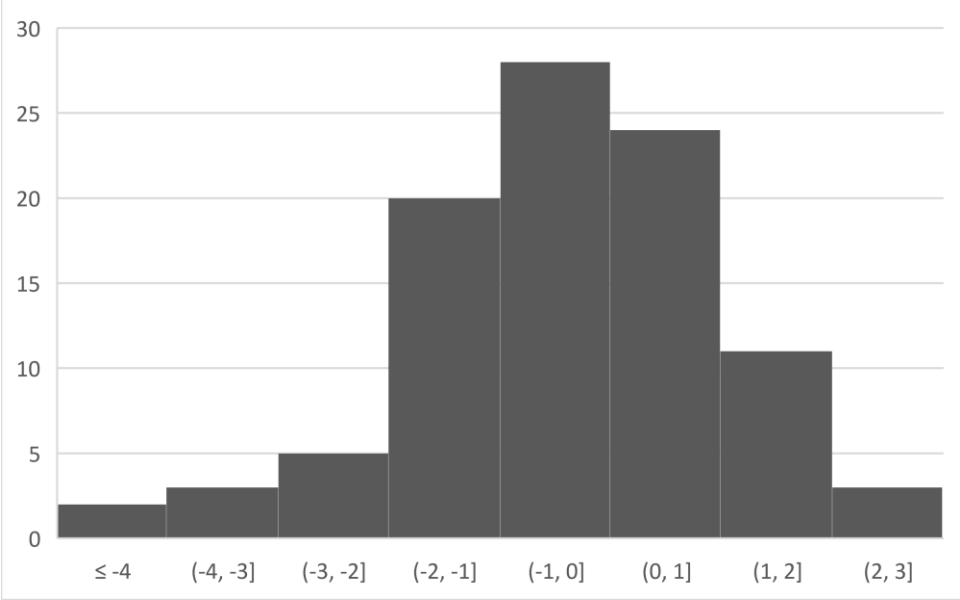


Figure 6. Distribution of Study Set Rasch Difficulties (Set 3)



SME Content and Cognitive Processing Variables (Research Question 1 and 2a)

The decision to use SMEs for the exploratory qualitative portion of the study, which included participation in a focus group and cognitive interviews, was based on the insight that SMEs can provide in understanding both general pediatrics test item content and the knowledge and skills possessed by examinees sitting for the general pediatrics exam. The group of five SMEs were selected to participate in the study based on both their membership on the ABP General Pediatrics examination committee and their availability to attend a focus group which was held during the 2019 annual General Pediatrics committee meeting. Each of the five SMEs had extensive experience in item writing and item review of general pediatrics test items as evidenced by their length of service on the committee which ranged from four to six years. Additionally, the five SMEs came from varying practice settings with three currently in private practice, one currently hospital based, and one currently serving as the Vice Dean of Medical Education at a large university.

The first step in the exploratory qualitative phase of the study was the use of a focus group (attended by the five SMEs previously discussed). The decision to use a focus group was based on the ability of this qualitative approach to allow for the articulation of ideas and insight into a topic (Peters, 2019). Peters (2019) also states that focus groups "allow evaluators to understand how people think or feel about something." The need of the current research to first understand a pediatrician's thought process when answering test items, including what may be causing an item to be easy or difficult, made the use of a focus group a logical first step in this process. The focus group was

moderated using the recommendations for moderating a focus group provided by Beverly (2019). The agenda used for the focus group, and corresponding focus group recommendation (Beverly, 2019) is outlined in Table 7.

Table 7. Agenda for Initial In-Person Subject Matter Expert Meeting

Session	Purpose	Time (minutes)	Recommendation Addressed (Beverly, 2019)
Introduction	Introduction to the goals and purpose of the study and explanation of the benefits of predicting item difficulty	45	- Give an overview - Set ground rules - Request permission to record
Item Review and Think Aloud	Review of previously used general pediatrics certification items and discussion on cognitive processes and variables which may be driving item difficulty	90	- Moderate tweaking questions and question order based on the conversation
Conclusion	Summary of item review and cognitive interviews and outline of next steps	15	- Summarize what was said - Verify session recorded successfully

The introduction to the focus group included a presentation on the research goals and questions. Also included in the introduction were questions to keep in mind during the item review portion. The questions were based on previous related literature (Ferrara et al., 2018) and the researcher's experience with general pediatric item development. The questions were:

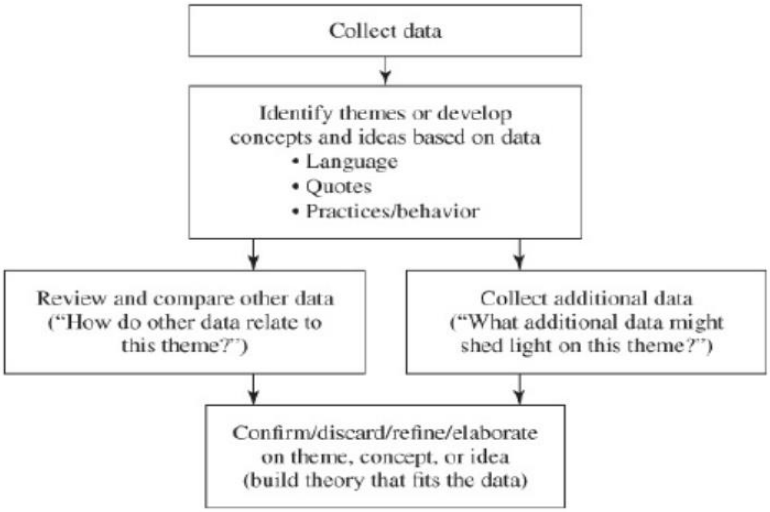
1. How many mental steps are required, or concepts understood, to arrive at the correct answer?
    a. An example of this is the relational complexity variable defined by Ferrara et al. (2018) as "the number of (a) concepts that examinees must hold in mind, (b) facts that examinees must hold in mind, or (c) cognitive processes that examinees must undertake in order to process and respond to an item and their relationships to one another."
2. How common is the patient presentation?
3. What is the quality of each distractor?
4. Are there any different approaches that would be taken to answer the item correctly (based on the Keehner at al. (2017) notation that different approaches may be taken to answer, and asked to the SMEs who did not initially walk through the item response process)?

Following the introductory presentation, the SMEs were advised of the ANOVA results on content domain and universal task effects on Rasch item difficulty. The ANOVA was conducted on 2,973 previously tested General Pediatric certification items to determine if an item's content domain or universal task has a significant effect on the Rasch item difficulty. There was not significant effect found on the Rasch item difficulty for the 25 content domains [$F(24, 2394) = 1.497$, $p = 0.057$]. Additionally, there was not significant effect found on the Rasch item difficulty for the four universal tasks [$F(3, 2415) = 2.506$, $p = 0.057$]. These results were shared so that the SMEs could consider these factors when discussing the difficulty drivers of items. During the focus group discussion, the SMEs

walked through the cognitive process they would expect an entry level examinee taking the general pediatrics exam to experience when answering the items.

After the initial discussions, the SMEs were provided with the item's actual percent correct and prompted to discuss any discrepancies with their perceived difficulty level of an item and the actual difficulty. The focus group was recorded and transcribed following the meeting as recommended by Bloor et al. (2001) using Trint transcription software (Kofman, 2019). The transcription, along with the researchers notes, were then used to conduct a qualitative analysis for both the in-person and SME webinars following the grounded theory approach and analysis outlined by Taylor et al. (2015). The steps for following the grounded theory approach, as displayed in Taylor et al. (2015), are shown below.

Figure 7. Taylor et al. (2015) Grounded Theory Approach



The first step of this approach, identifying themes, concepts, and ideas (from the transcribed audio and researcher notes), was conducted for the focus group. During the

qualitative analysis of the focus group, coding was performed on the data to understand commonly stated aspects of item response processes. Coding "involves bringing together and analyzing all the data bearing on major themes, ideas, concepts, interpretations, and prepositions (Taylor et al., 2015)." This analysis included capturing "concepts" (Rubin, 2005) which are words or terms that are used in interviews and which address the research problem. Each concept of what may be driving an item's difficulty was given a unique code and recorded in a Microsoft Excel spreadsheet. This coding process is also based on the open coding process for the cognitive pre-testing of items put forth by Lenzer et al. (2016). Taylor et al. (2015) recommend providing quotes to illustrate concepts derived from the coding process. During the qualitative analysis of the focus group, specific quotes from the focus group, which enforced the coded data, were recorded to reinforce the analysis and results. Using the commonly cited concepts from the focus group, a set of questions was developed for prompting the SMEs during the next portion of the study which consisted of five cognitive interview webinars.

The next step in the research addressed the grounded theory approach step on collecting additional data. Five, 1.5-hour cognitive interview webinars were conducted with the SMEs to discuss general pediatrics items. Cognitive interviewing as a method to guide the SME discussion was chosen due to its effectiveness in "exploring specific components of a question and affordances for targeting issues of interest (in this case, item difficulty drivers)" (Keehner et al., 2017). Using the Keehner et al. (2017) small sample size recommendation for cognitive interviews, four of the webinars were conducted with one SME, and one webinar was conducted with two SMEs. The use of

cognitive interviews to explore item difficulty was also recommended in the research conducted by Gorin (2006). The agenda used during the individual webinars is outlined in Table 8.

Table 8. Agenda for Individual SME Webinars

Session	Purpose	Time (minutes)
Introduction	Presentation on research study goals, recap of in-person meeting findings, and outstanding questions answered	15
Item Review and Cognitive Interviews	Review of previously used general pediatrics certification items and discussion on cognitive processes and variables which may be driving item difficulty	65
Conclusion	Summary of item review and cognitive interviews and outline of next steps	10

At the beginning of each webinar, a presentation was made to refresh the SMEs on the goals of the research, present the in-person meeting findings, and to answer any questions which remain from the introductory in-person meeting. During the webinars, the SMEs were asked to walk through their process in answering a test item and asked questions about each item which were formed by the focus group qualitative analysis. The items reviewed during the webinars came from the 101-item training set and were reviewed by alternating difficulty level to ensure that items of all difficulties were reviewed. The questions posed to the SME for each item, stemming from the focus group analysis, were:

1. Using a 1 to 4 scale (1 being rare and 4 being a “classical” presentation) how commonly is the patient presentation seen with the subsequent diagnosis or management being asked for?
2. Using a 1 to 4 scale (1 being a few times in your career or a few times in the course of training and 4 being several times per month) how commonly is the item’s diagnosis or patient presentation seen in practice or touched on in training?
3. For each distractor, using a 1 to 4 scale (1 being not plausible and easily eliminated and 4 being highly plausible and difficult to differentiate from the correct answer) how plausible is each distractor?
4. How many pieces of information are required to answer the item correctly?
5. How many pieces of extraneous information are present in the item?
6. Are keywords or patterns that are specifically addressed in training found within the item?

Additional notes were also recorded during the webinars on any concepts which were not covered by the researcher questions. Answers to each of the questions, and additional concepts, were recorded in a Microsoft Excel spreadsheet. Each webinar concluded with a summary of the item level variables addressed by the questions and the SMEs were asked if there were any additional factors driving item difficulty that they noticed after their review of the items. Following the final webinar, using the questions and recorded answers for each item, the final step in the Taylor et al. (2015) grounded theory approach was conducted to confirm, discard, and elaborate on the difficulty variables. Pearson’s correlation was used to determine if the responses to the interview

questions correlated with the Rasch item difficulty. The results from this analysis were then used to confirm the use of the difficulty concepts for the research study. The researcher notes were also analyzed using the previously cited qualitative coding process (Taylor et al., 2015) to determine if additional difficulty driving concepts would be included in the variable coding process.

Following the webinar analysis, metadata were created within the ABP's item banking system which allowed for the SMEs to code the difficulty variables for each study item. A coding assignment was created, evenly dividing the items among the SMEs both in number of items and in item difficulty (each SME was assigned an equal amount of easy, medium, and difficult items). The statistics for each item were hidden (using the hide statistics item banking setting) so that the SME coding responses were not biased and to demonstrate how the process would work if implemented with previously unused test items. A detailed email was sent to each SME outlining how to conduct the coding assignment within the item banking software. An attachment was included with the email which detailed each variable to be coded and the scale to code each variable with. Finally, it was emphasized in the instructions that SMEs should complete the assignment upon logging into the software and promptly log out when finished so that the time spent on the assignment could be recorded and reported on. Each SME was assigned 20 items and given 10 days to complete the assignment. Following the initial coding assignment, the assignments were rotated so that each SME was assigned a second set of 20 items. The second round of coding was used to produce a rater agreement rate for the coding process. During the second coding assignment, SMEs were unable to see the round 1

coding values to ensure that coding responses would not be biased. Each SME was given 10 days to complete the second coding assignment. After completion of the second round of coding, any item with a different (yes/no) pattern recognition code, was assigned to a third SME so that a final determination could be made on which variable value to use. Once all of the variable coding was complete, the items with variable values were exported into a Microsoft Excel spreadsheet and a rater agreement percentage and Fleiss' kappa value was calculated to determine the level of consistency between SMEs during the coding process.

Coh-Metrix Linguistic Variables (Research Question 1 and 2b)

In addition to the SME defined content and cognitive processing variables, each item was analyzed using Coh-Metrix linguistic software (Grasser et al., 2004) to determine if any linguistic features in the items predict item difficulty. Coh-Metrix is "a system for computing computational cohesion and coherence metrics for written and spoken texts." (Grasser et al., 2004) Further, Coh-Metrix allows readers, writers, educators, and researchers to instantly gauge the difficulty of written text for the target audience." (Grasser et al., 2018). The Coh-Metrix tool codes 108 linguistic variables that fall within 11 linguistic indices. The 11 Coh-Metrix linguistic indices along with number of variables within each and examples are provided in Table 9.

Table 9. Coh-Metrix Variable Indices

Indices label	Number of Variables	Example Variables
Descriptive	11	Paragraph & sentence count, word length

Text easability principle component scores	16	Narrativity, verb cohesion, word concreteness
Referential cohesion	10	Stem, noun, and content overlap
Latent semantic analysis	8	Latent semantic overlap in all and adjacent sentences
Lexical diversity	4	"Type-token" ration
Connectives	9	Additive, positive, and negative connectives
Situation model	8	Causal and intentional verb incidences
Syntactic complexity	7	Words before main verb and number of modifiers per noun phrase
Word information	22	Noun, verb, adjective, adverb, and pronoun incidence
Readability	3	Flesch reading easy and grad level

The complete list of Coh-Metrix variables and definitions is provided in Appendix A. For each item analyzed by the Coh-Metrix software, the text of the item was copied and pasted into the online Coh-Metrix platform, placing a hard return between the item stem and each item answer choice. The Coh-Metrix software was then executed and the value for each variable (provided in the Coh-Metrix output) was recorded in an Excel spreadsheet. Due to the large number of Coh-Metrix variables and the limited sample size of 96 study items, a principle component analysis (PCA) was performed using the 101-item training set. The PCA allowed for a reduction in dimensionality and variables by reducing the 108 Coh-Metrix variables into components. Using the retained components from the PCA, component scores were calculated for each of the 96 study items. To determine the number of components to retain, the scree plot and "elbow" rule was used. Once the components were determined, the coefficient for each variable within each retained component was recorded in an Excel spreadsheet. After the PCA was completed on the training set of items, the 96 items that were coded with the SME variables were

analyzed using the Coh-Metrix application. Since the PCA was conducted using the correlation matrix, the Coh-Metrix variables for the study items were standardized to match the standardization that took place on the training items during the PCA. The standardized variables for each of the 96 study items were then used to create a component score, by multiplying each variable by the variable's coefficient from the training set. Finally, the products of each variable and coefficient within an identified component were summed to create a component score for each variable and item.

Multiple Regression Analysis (Research Question 3)

Once the 96 items were coded with both the SME defined variables and the Coh-Metrix component scores, a backward ordinary least squares regression analysis was conducted using SPSS with the Rasch item difficulty value as the dependent variable and the SME variables and Coh-Metrix component scores as the independent variables. The assumptions of normality, homoscedasticity, and multicollinearity were checked after running the analysis to ensure all assumptions were met. Normality of the data was checked using the predicted probability (P-P) plot. Homoscedasticity was checked by plotting the predicted and residual values on a scatterplot. Multicollinearity of the data was be checked by ensuring all VIF values were less than 10. Finally, a regression equation was created using the backward regression model with the most significant R-squared and significant beta coefficients retained.

Impact Study (Research Question 4)

A final analysis was planned, but not carried out, to demonstrate the impact of using predicted Rasch item difficulties on the calibration of examinee ability levels. Due

to a lower than expected R-squared, this portion of the study was not completed, but the methods for completing this described below to illustrate to future researchers how an impact study on item difficulty modeling research can be carried out.

Using the regression equation and coded variable values, a predicted Rasch item difficulty for each item can be created. The test items and examinee ability levels can then freely calibrated, with Winsteps software, using previously collected live data. Next, the test items can be fixed with the predicted Rasch item difficulties (using the Winsteps anchor set function) and the examinee ability levels can be calibrated using the fixed item difficulty values. A Pearson correlation can then calculated to show the relationship between the two examinee ability levels (freely calibrated and anchored using predicted Rasch item difficulties).

Conclusion

This chapter described the methods used in the current research to create a workable and repeatable process for test developers to accurately predict item difficulty without the need for pilot testing. The methods outlined how test developers can define content and cognitive processing variables for test items to answer research question one, 2a, and 2b. These research questions aimed to define a cost-effective and operationally feasible process for SMEs and test developers to code content- and cognitive-related variables to previously used test items. The success of these methods was measured by both the replicability of the process and the amount of variance which the variables were able to explain in the test items. The methods described to use a multiple regression analysis to create an item difficulty regression equation were implemented to answer

research question three which is “can cognitive variables defined by SMEs, and linguistic variables coded using Coh-Metrix software, account for a significant amount of variance in Rasch item difficulty?” Finally, the methods on conducting an impact demonstrated how future researchers can demonstrate the impact of using predicted item difficulties on examinee scoring decisions. Ultimately, the most consequential benefit of these activities was the ability to define a process which test developers can implement in the future to determine significant difficulty driving variables and account for those variables during the item writing and review stages of item development.

CHAPTER IV

RESULTS

In this chapter, the process and results from the SME focus group and cognitive interviews, Coh-Metrix analysis, SME variable coding process, multiple regression analysis, and impact analysis, will be presented along with an explanation how each result addresses the research questions stated in the previous methods chapter. Due to the several different analyses required for completion of this study, including both qualitative and quantitative analyses, a mixed methods exploratory sequential design (Watkins & Gioia, 2015) was chosen for use. The results section will first report on the process and results from the qualitative analysis conducted following both the in-person SME focus group and the five SME cognitive interview webinars. These results demonstrate how the previously outlined process was used to define content and cognitive variables within test items. Second, the results from the SME variable coding process will presented including the average time this activity required, which speaks to the operational feasibility of the process, and the interrater reliability determined using two rounds of coding, which demonstrates consistency in the variable coding process. The results from the Coh-Metrix analysis, including the PCA results on the training set of items, will be then be presented. Next, the results of the multiple regression analysis will be presented. Finally, the impact study will be discussed as this portion of the research was not completed due the .80 variance threshold not being met.

Focus Group and Cognitive Interview Webinar Results

As outlined in the previous chapter, both an in-person focus group with five SMEs and five subsequent cognitive interview webinars were held in order to address research questions 1 and 2a. A qualitative and quantitative analysis was then conducted using the grounded theory approach (Taylor et al., 2015) to define the item difficulty variables to use when answering research question 3. A total of 12 GP test items with varying difficulties were discussed during the in-person focus group. Following the focus group, Trint transcription software (Kofman, 2019) was used to transcribe the audio recording of the session. The transcription was then used to code concepts and themes which were discussed in the focus group along with memorable quotes from the participants which reinforced the coded concepts. Table 10 provides the concepts which were coded using the focus group transcription along with the frequency (in number of items) that each concept was cited.

Table 10. Focus Group Coded Concepts and Frequencies

Difficulty Driver	Frequency
Processing of multiple pieces of information required	5
Keywords/pattern recognition used when solving	4
Uncommon patient presentation	4
Not commonly seen or trained on	3
Extraneous information in stem	3
Options not plausible	3
Commonly seen or trained on	3
Missing information which would be seen in practice	3
Common (classic) patient presentation	3
Plausible distractors	3
Diagnose and management required	2
Patient age not commonly seen with diagnosis	2
Diagnosis provided in stem	1

No extraneous information in stem	1
Important topic/content	1
Table in stem	1
Compound correct answer (two parts)	1

The coded concepts which addressed the same topic or theme were then combined to create a final list of concepts. Table 11 provides the coded concepts and frequency which the concepts were cited after the initial concepts addressing the same theme were combined.

Table 11. Reduced Focus Group Concepts and Frequencies

Concept	Frequency
Common/uncommon patient presentation	7
Frequency that a condition is seen in practice or trained on in training	6
Distractor plausibility	6
Processing of multiple pieces of information required	5
Amount of extraneous information in the stem	4
Keywords/pattern recognition used when solving	4
Missing information which would be seen in practice	3
Diagnose and management required	2
Diagnosis provided in stem	1
Important topic/content	1
Table in stem	1
Compound correct answer (two parts)	1

The most frequently cited concept discussed was how common or uncommon a patient presentation described in an item corresponded to the diagnosis or management option which the examinee was asked to identify. In one instance, the group agreed that a classical patient description made an item easy (where the diagnosis was asked for). One SME stated:

> This I would get because this is a classic picture that has always been described to me this way.

Another SME, speaking about an item with an uncommon patient presentation stated:

> The age of this child with this condition is not common and a bit of a 'red herring.' And a continuous murmur is a very rare presentation for this condition. The question writer made this one an unusual presentation.

The frequency which a practitioner sees a patient presentation and condition in practice or training was another commonly cited concept. When citing this concept, and how an item assessing knowledge of a commonly seen condition is easier to answer correctly, one SME stated:

> This to me would be very easy because it is a great description of something that you see and do every single day. If a trainee ever went to a clinic, then they saw this.

On a less commonly seen patient, one SME commented:

> I've seen one case of rickets in 30 years. If you haven't seen it, who the heck remembers?

Distractor plausibility was another concept which was brought up during the discussion. The SMEs vocalized that being able to easily eliminate less plausible distractor options makes an item easier since even those with less knowledge can narrow the options and use a process of an elimination approach. On one item, while admitting to not being greatly familiar with the item content area, a SME said:

> Not being a cardiologist, I did reason this one out. I took C and D right off the list because C applies to babies and D doesn't exist. I was down to A, B, and E, and if the area doesn't have clear external borders, it's either A or E, and then I was down to 50-50.

A separate SME commenting on the same item stated:

> A and B seemed like the same thing to me. I'm not sure what D is. D is not plausible since it doesn't exist. That's how we are eliminating these options. ASD I would expect that the s^2 would be split, so cross that out, and VSD would have a whole systolic murmur, which takes that out, which leads me down to D and E, and D does not exist, so I know it is E.

Distractor plausibility was addressed on another item, with the options being plausible and thus making the item more difficult. During this discussion, a SME stated:

> When I was looking at the options and trying to figure out 'well why is this choice even a choice,' maybe they're trying to say erythema and crusting. So, this is a superinfection and then I need to worry about that. And if you were learning that way, you might choose B. And on top of that, I got to D and E and I was like, alright, I've never used that for skin, and cephalosporin and noraxon, maybe I eliminate both of those since they are both a cephalosporin. So, I think the only option I can quickly eliminate is C. Then, you have to decide using the description what you are treating, impetigo or cellulitis. So, the question is, what's the answer?

Kirsch & Mosenthal (1988) noted distractor plausibility in their work on item difficulty modeling in the NAEP survey as well, which added credence to the use of this as an item difficulty variable.

The amount of information an examinee is required to process in order to answer an item correctly, with larger amounts making items more difficult, was discussed by the

group. When discussing an item which required multiple pieces of information, and the processing of that information to answer the item correctly, a SME stated:

> This is a four-step question. You have to understand this, and then you have to go back to this. Your figuring out is this physical exam normal or abnormal and you're also figuring out is the patient's development normal or abnormal. There are multiple pieces of information. And if either is not normal, what category does this go in, and then based on category, what test do I do? So, it takes more information and steps in the thinking process.

The amount of extraneous information in an item, requiring an examinee to differentiate extraneous information from the information needed to correctly respond, was discussed by the group as impacting an item's difficulty. While cited in several different items, the concept was elaborated on in one specific item. When discussing this item, one SME stated:

> What makes this item harder is there are multiple factors in this scenario. Some of them are distractors. So, it is irrelevant that the child is doing well in kindergarten to the rest of the problem. So, they're giving you a lot of extraneous stuff with key pieces that you need to pay attention to. You have to figure out what is important to answer the question and what is irrelevant to answer the question. They're telling you he is atopic below the eyes which is not necessarily helpful, but it might be. You have to think about that, is it helpful or is it not helpful to answer the question. You have to say helpful or not helpful, relevant or not relevant, before you can even answer the question. So, to the question of difficulty, there are multiple distractors in the stem.

Extraneous information was a previously cited item difficulty component in the work by Fulkerson et al. (2011) and Kirsch & Mosenthal (1988) (distracting information) which further promoted the use of this variable in the current research.

Pattern and keyword recognition were also cited during the focus group and the ability to identify patterns to correctly answer items impacting how easily an item can be answered. The SMEs agreed that in pediatric training, trainees are taught to recognize specific patient patterns in order to identify underlying conditions. When speaking on pattern and keyword recognition, one SME stated:

> And so much for me in these items is pattern recognition and word choice. It's like whenever you see the word 'sandpaper rich' you know its Scarlett fever. Those buzzwords lead you to the right answer.

In response, another SME commented on both pattern recognition and keywords, and how question writers sometimes avoid the use of common terms to make items more difficult. The SME stated:

> Sometimes item writers will avoid those keywords and come up with some other terms so it's not so much clueing and to make it more challenging.

Pattern and keyword recognition relate to a previously cited item difficulty component which Fulkerson et al. (2011) defined as 'schema activation' or the "application of mental structures drawing on experience."

Using the results from the individual and grouped difficulty driver concepts, along with the review of the cited and additional notable quotes, targeted questions were created for use during the SME webinar sessions with difficulty driving concepts which were cited four or more times having specific questions based on them. The following targeted questions were developed and asked for each item reviewed during the SME webinars.

1. Using a 1 to 4 scale (1 being rare and 4 being a “classical” presentation) how commonly is the patient presentation seen with the subsequent diagnosis or management being asked for?
2. Using a 1 to 4 scale (1 being a few times in your career or a few times in the course of training and 4 being several times per month) how commonly is the item’s diagnosis or patient presentation seen in practice or touched on in training?
3. For each distractor, using a 1 to 4 scale (1 being not plausible and easily eliminated and 4 being highly plausible and difficult to differentiate from the correct answer) how plausible is each distractor?
4. How many pieces of information are required to answer the item correctly?
5. How many pieces of extraneous information are present in the item?
6. Are keywords or patterns that are specifically addressed in training in the item?

In some instances, where an item did not lend itself to a topic or difficulty driver which the question targeted, an N/A was recorded in the spreadsheet. For example, an item assessing the General Pediatrics scholarly activities domain would not lend itself to the question regarding patient presentation commonality and thus would have an N/A coded for that question and item. A total of 53 of the training set items were discussed over the course of the five webinars with an item difficulty breakdown of 19 easy, 16 medium, and 18 difficult items. Following the final webinar, the values for each recorded question response (minus an N/A responses) were correlated using Pearson’s Product Momentum correlation with the corresponding item difficulties. The plausibility rating for each item’s distractors was summed and the total for each was used to create a new

difficulty driver labeled “Total Distractor Plausibility,” which was then correlated with the item difficulties. The webinars and resulting question answers and correlations with item difficulty were used to address the final two steps of the Taylor et al. (2015) grounded theory approach which are “collecting additional data” and “confirm, discard, refine, and elaborate concepts.” The resulting item difficulty drivers and their correlations with item difficulties are presented in table 12.

Table 12. Correlation Between Item Difficulty Drivers and Item Difficulty

Difficulty Driver	Pearson’s Correlation
Common/uncommon patient presentation	-.360
Frequency of condition/management being seen in practice or trained on	-.209
Distractor plausibility	.645
Number of pieces in stem required to answer	.405
Number of extraneous pieces of information in stem	.500
Keywords/patterns (Yes/No)	-.380

Each difficulty driver correlated above the .20 level and in the expected direction (positive or negative) with an item’s Rasch difficulty. These results confirmed the focus group qualitative results. Using a combination of the focus group qualitative analysis results, and targeted question correlation results, the six difficulty driving concepts which emerged from the focus group and which were further examined during the webinars, were retained for use in the study.

Following the finalization of SME variables, drop down metadata were created using the ABP’s item banking software and the SMEs were each given 19 to 20 items to code with the six difficulty driving variables. The SMEs were provided a detailed

instruction set on how to complete the assignment within the item banking system and a document which outlined each variable and coding scale. The SMEs were also given the chance to communicate any concerns with the variable definitions or scales prior to the start of variable coding. The SMEs reported having no concerns or disagreements with the finalized set of difficulty drivers. The communication to the SMEs and variable definition and scale attachment can be found in Appendix D.

The amount and average time spent (rounded to the nearest minute) by SMEs during the item variable coding process is presented in table 13.

Table 13. SME Time Spent (Minutes) on Item Variable Coding

SME	Round 1	R1 Average Per Item	Round 2	R2 Average Per Item
1	46	2.4	43	2.3
2	63	3.1	62	3.3
3	65	3.4	64	3.4
4	54	2.8	134	7.1
5	49	2.6	56	2.9
Total	277	-	359	-
Average	55	2.9	56*	3.8

*Rater 4 was not included in the round 2 average as they most likely remained logged into the system while not coding the items therefore inflating their total time spent (134)

Following the second round of coding, the consistency between raters was examined to determine the amount of agreement between the first and second rater on each variable. For each variable, the percent of agreement was calculated. The rater consistency by percent of agreement between raters is summarized in table 14.

Table 14. Rater Agreement Rates

Variable	Rater 1 and 5 (n=38)	Rater 2 and 3 (n=19)	Rater 2 and 4 (n =20)	Rater 3 and 4 (n =19)	Average
Frequency Diagnosis is Seen/Trained On (1 to 4)	44.7%	26.3%	40.0%	47.4%	39.6%
Common/Uncommon Patient Presentation (1 to 4)	47.4%	47.4%	55.0%	42.1%	48.0%
Pieces of Information Required to Answer (continuous)	10.5%	15.8%	20.0%	5.3%	12.9%
Pieces of Extraneous Information (continuous)	13.2%	15.8%	5.0%	21.1%	13.8%
Pattern/Keyword Recognition Required (Yes/No)	63.2%	84.2%	40.0%	57.9%	61.3%
Option A Plausibility (1 to 4)	42.9%	23.5%	37.5%	38.5%	35.6%
Option B Plausibility (1 to 4)	39.3%	25.0%	31.3%	46.7%	35.6%
Option C Plausibility (1 to 4)	41.4%	46.7%	38.5%	53.3%	45.0%
Option D Plausibility (1 to 4)	41.9%	50.0%	20%	28.6%	35.1%
Total Average	38.3%	37.2%	31.9%	37.9%	36.3%
Average Excluding Variables 3 and 4 (continuous variables)	45.8%	43.3%	37.5%	46.3%	43.2%

*The pieces of information required to answer correct and pieces of extraneous information were continuous variables, and therefore a lower percentage of agreement rate was expected

The most consistently rated variable was the pattern/keyword recognition variable, which was expected as the variable only had two options for coding. Likewise, both the number of pieces of information required to answer an item correctly and the number of pieces of erroneous information present were the most inconstantly rated variables which was also expected as those variables were both continuous. While the overall percentages of agreement were not as high as expected, it is worth noting that 61 percent of the

disagreements on the six variables using a 1 to 4 scale were rated within one rating of each other. The percentages of rater agreement do however indicate a need for additional training in the coding process in future iterations of this process.

In addition to rater agreement, the Fleiss' kappa statistic was calculated for the variables which were coded as yes/no or using the 1 to 4 scale. The Fleiss' kappa statistics are reported in table 15.

Table 15. Fleiss' Kappa Rater Agreement

Variable	Rater 1 and 5 (n=38)	Rater 2 and 3 (n=19)	Rater 2 and 4 (n =20)	Rater 3 and 4 (n =19)
Frequency Diagnosis is Seen/Trained On (1 to 4)	.245	.039	.129	.207
Common/Uncommon Patient Presentation (1 to 4)	.235	.223	.118	-.066
Pattern/Keyword Recognition Required (Yes/No)	.244	.481	-.250	-.267
Option A Plausibility (1 to 4)	.098	-.144	.140	.100
Option B Plausibility (1 to 4)	.109	.432	-.032	.121
Option C Plausibility (1 to 4)	.145	.300	.005	.195
Option D Plausibility (1 to 4)	.198	.221	-.169	-.157

Using the guidelines from Landis & Koch (1977) on assessing the strength of Cohen's kappa, only one set of raters reached a rate of moderate agreement (>.40). The majority of the Fleiss' kappa statistics were classified as either a poor (<.20) or fair (.20 - .39) strength of agreement.

To determine the final variable values, the average of each variable code was taken from the round one and round two ratings. When a difference between the round one and round two coding was found for the variable addressing pattern and keyword

recognition (Yes/No), a third SME was asked to provide a coding to determine the final code to use. As stated in the methods chapter, the distractor plausibility ratings were first averaged between raters and then summed across the distractors for each item to create a single variable titled "total distractor score."

Coh-Metrix Linguistic and Text-Based Variables Analysis

Coh-Metrix linguistic software was used to analyze both a training set of 101 items and the 96 items used in the current study. Due to the large number of variables (108), a PCA was conducted on the training set of items using SPSS statistical software to reduce the variables into components and apply the component variable coefficients to create component scores for the study items.

Training Items (101 Items)

One hundred and one training items were analyzed by Coh-Metrix linguistic software and the 108 Coh-Metrix variables for each item were recorded in a Microsoft Excel spreadsheet. Three resulting Coh-Metrix variables were not analyzed. These variables were WRDPRP1s, WRDPRP1p, and WRDPRP2. Each of these three variables addresses an incidence score for pronouns in the first and second person. The three variables were excluded due a score of zero being reported for each of the training items (leading to zero variance within each of the variables). This result is due to the ABP's internal editorial style which does not allow for the use of pronouns (in the first and second person) in test items. The remaining 105 Coh-Metrix variables were analyzed for the 101 training items by conducting a PCA analyzing the correlation matrix within SPSS statistical software. The resulting PCA analysis revealed nine components to retain and

apply to the study items. The decision to retain nine components was made by analyzing the PCA scree plot and applying the “elbow rule.” The cumulative variance explained by the nine components was 67.162%. The resulting scree plot from the PCA conducted on the training set of items is displayed in figure 8 and the eigenvalues and total percent of variance explained for the nine components is displayed in table 16.

Figure 8. PCA Scree Plot (101 Training Items)

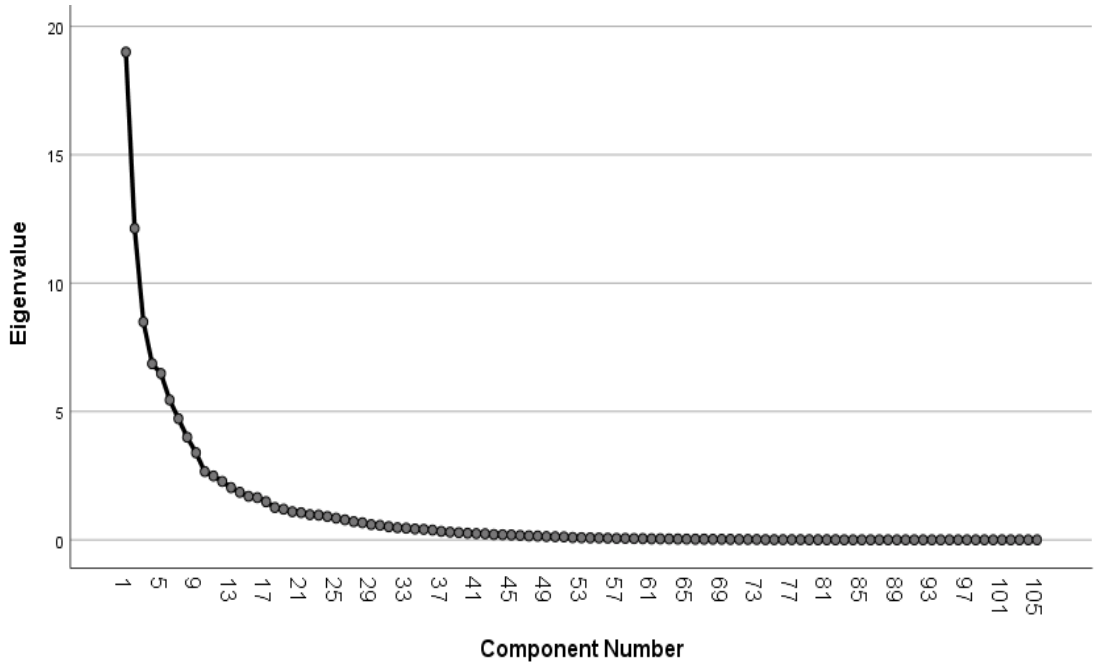


Table 16. PCA Retained Component Eigenvalues and Variance Explained

Component	Eigenvalue	% of Variance	Cumulative % of Variance
1	19.000	18.095	18.095
2	12.132	11.554	29.650
3	8.488	8.084	37.734
4	6.860	6.534	44.268
5	6.481	6.172	50.440
6	5.447	5.188	55.628
7	4.721	4.496	60.124
8	3.996	3.806	63.930
9	3.394	3.232	67.162

The ninety-six items used in the current research to address the research questions were analyzed by Coh-Metrix linguistic software and the 108 Coh-Metrix variables for each item were recorded in a Microsoft Excel spreadsheet. The three Coh-Metrix

variables which were excluded from the training set (WRDPRP1s, WRDPRP1p, and WRDPRP2) were also excluded from the 96 study items due to values of zero being recorded for each variable and item. The resulting 105 Coh-Metrix variables for each study item were recorded in a Microsoft Excel spreadsheet. The PCA conducted on the training items used the correlation matrix (which standardizes the variables), which required the values of the study items to be standardized prior creating the component score variables (by subtracting from each variable value the average of the variable and dividing by the variable's standard deviation). A component score for each of the nine components was calculated for each of item by multiplying the component coefficient from the training set by the study item variable value and summing the products over the 105 variables within the component. The nine component scores for each of the 96 study items were then saved in a Microsoft Excel spreadsheet.

Multiple Regression Results

A backward multiple regression, using an *F* probability removal criterion of .10 at each step, was calculated to predict Rach item difficulty on the six SME variables and nine Coh-Metrix component variables. After running the initial regression, two items were shown to be outliers in the analysis. The two item's SPSS Casewise diagnostics values are shown in table 17.

Table 17. Casewise Diagnostics for Excluded Items

Case Number	Std. Residual	IRTb	Predicted Value	Residual
53	-3.593	-4.18	.0909	-4.27468
75	-4.281	-6.78	-.8981	-5.87891

The decision was made to remove the two outlier items from the data and to calculate the multiple regression using the remaining 94 study items. Using the backward method, it was found that the unconditional model with all the variables entered explained a significant amount of the variance in Rasch item difficulty, $F(15, 77) = 3.128$, $p < .01$, $R^2 = .379$, R^2 Adjusted = .258. The total distractor score variable and Coh-Metrix Factor 3, 5, 6, and 7 variables were all found to be significant in the unconditional model, $p < .05$. The final model created in the backward regression retained the five significant variables found in the unconditional model and explained a significant amount of the variance in Rasch item difficulty, $F(1, 86) = 2.696$, $p < .01$, $R^2 = .324$, R^2 Adjusted = .285. Equation 1 displays the equation created by the final model for predicting Rasch item difficulty.

Equation 1. Regression Equation for Predicting Rasch Item Difficulty

Rasch Item Difficulty = -1.868 - .277 (Coh-Metrix Factor 7) +
.25 (Total Distractor Score) + .392 (Coh-Metrix Factor 3) + .281 (Coh-Metrix Factor 5) +
.485 (Coh-Metrix Factor 6)

The unconditional and final model ANOVA summaries are presented in table 18 and 19.

Table 18. ANOVA Results for Unconditional Regression Model

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	58.104	15	3.874	3.128	.001
Residual	95.346	77	1.238		
Total	153.450	92			

Table 19. ANOVA Results for Final Regression Model

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	49.768	5	9.954	8.352	.000
Residual	103.682	87	1.192		
Total	153.450	92			

In the final regression model, total distractor score and Coh-Metrix Factors 3 and 6 were significant predictors of Rasch item difficulty at the $p < .01$ level, and Coh-Metrix factors 5 and 7 were significant predictors of Rasch item difficulty at the $p < .05$ level. The assumptions of normality, homoscedasticity, and multicollinearity were checked after running the analysis to ensure all assumptions were met. No VIF values were found to be greater than 10. The assumptions of normality and homoscedasticity were also met, and the respective plots showing the assumptions were met are displayed in figures 9, 10, and 11. The regression coefficients and coefficient correlations for both the unconditional model and the final regression model can be found in Appendix C.

Figure 9. Distribution of Dependent Variable (IRTb)

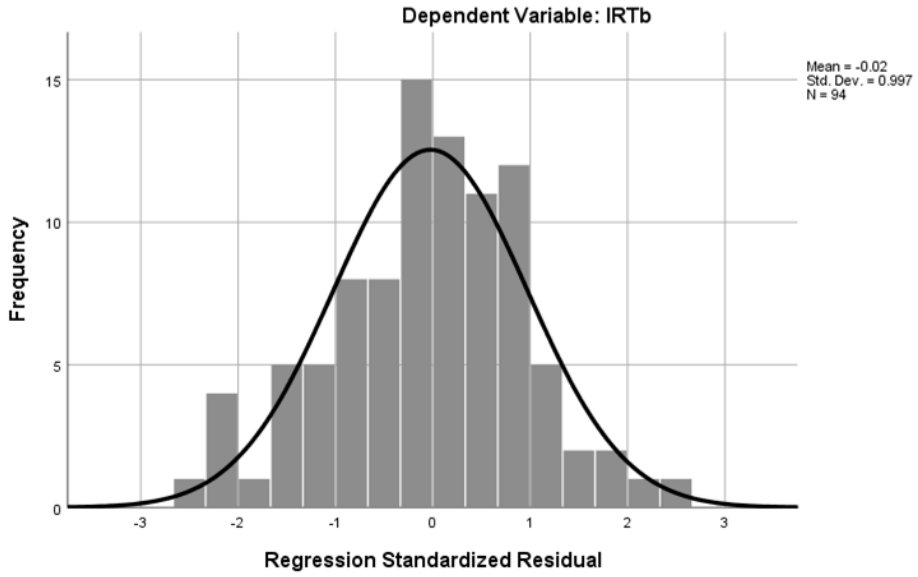


Figure 10. Normal P-P Plot of Regression Standardized Residuals

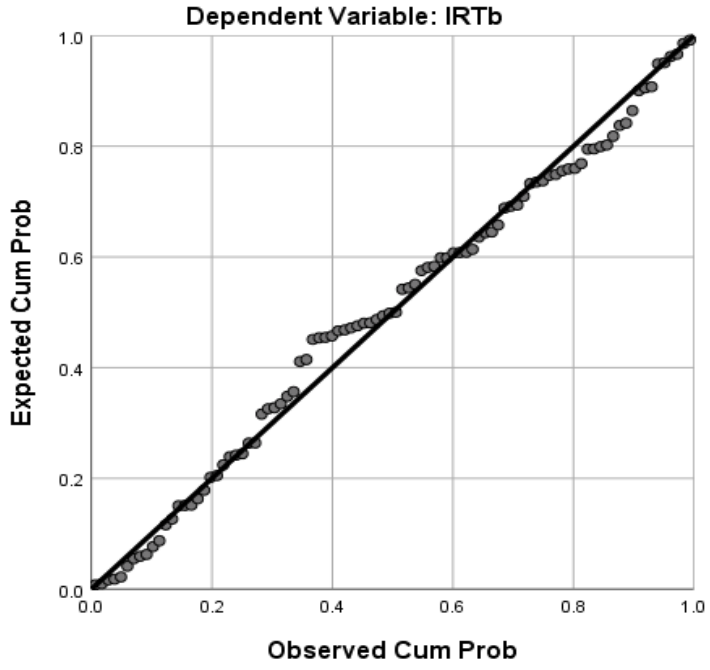
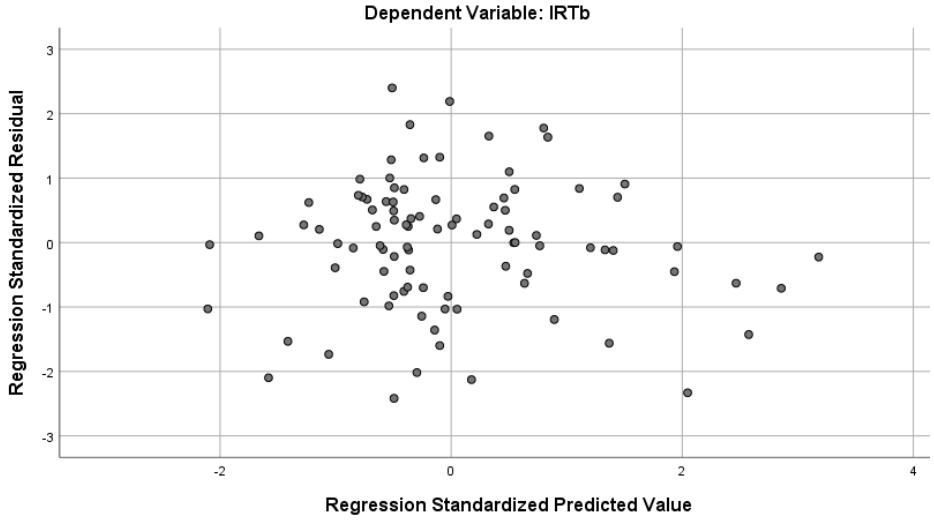


Figure 11. Scatter Plot of Standardized Residuals and Standardized Predicted Residuals



Impact Study Results

While the regression equation reported on in the previous section was found to be significant in predicting Rasch item difficulty, the final R-squared (.324) and adjusted R-squared (.285) did not explain enough variance in Rasch item difficulty to justify completing the impact study which was outlined in chapter 3. The planned impact study would be beneficial for similar studies where higher R-squared values (perhaps those meeting the Bejar, 1983, .80 criteria) are achieved. The impact study is further discussed in chapter 5.

CHAPTER V

DISCUSSION

The following chapter will present a discussion which addresses the research results and the implications of those results on each research question put forth in chapter 3. The limitations, next steps, and recommendations as they relate to each research question will be addressed. Finally, the significance of the current research and future steps for related research will be addressed including areas where future researchers may build upon the current research to achieve desirable outcomes.

Research Question 1

To address research question 1, an operationally feasible process for certification and licensure organizations to predict item difficulty using content- and cognitive related variables defined by SMEs was created and implemented. The current research successfully demonstrated how using a focus group and cognitive interview webinars; SMEs were able to cite different variables which they believed drove the difficulty of items. Testing organizations may be concerned with the associated costs, including SME time and meeting resources, with item difficulty modeling research. The current research only required one 2.5-hour in-person meeting which was conducted as part of an already scheduled item review meeting. If future research can align item difficulty research with already scheduled item review meetings, the costs related to meeting space, SME lodging, and SME travel, can be avoided (as was the case with the current research).

The initial focus group was used as an exploratory first step based on the ability of focus groups to allow for the articulation of ideas and insight into a topic (Peters, 2019). This step was needed in the current research as there is no prior literature examining content- and cognitive variables within general pediatric (or related medical specialty) items. As item difficulty modeling research in the medical certification progresses, exploratory meetings, such as the focus group conducted in this research, may be avoided in favor of targeted cognitive interviews based on prior literature.

SME time commitment is another valid concern with item difficulty modeling research as many SMEs are unpaid volunteers, with full-time jobs, who are already committing time to activities such as item writing and item review. The current research required SME time commitment of (1) a 2.5-hour in-person meeting (which as previously mentioned was conducted during a meeting they had already committed to attend), (2) a 2-hour webinar, and (3) two item coding assignments which took approximately 1 hour each. The total time commitment for each SME in the process put forth was 6.5 hours. None of the SMEs in the current research expressed concerns with the time commitment. Future studies may aim to quantify the amount of SME time spent reviewing and revising poorly performing items as a justification for time savings that can be realized from studies which successfully model item difficulty.

Another aspect of an operationally feasible process is the availability of resources for the researcher to carry out item difficulty modeling research. The current research was conducted using a webinar/conference calling service, SPSS statistical software, and Microsoft Excel. These three resources, or similar software packages which can achieve

the same outcomes, are readily available in most testing organizations. Additionally, the focus group discussion was transcribed using a free trial of Trint transcription software (Kofman, 2019). Future studies with more extensive transcription needs will need to account for a transcription software license. While the current research content- and cognitive variables were not as predictive (of Rasch item difficulty) as desired, the process put forth to define the variables was an operationally feasible and cost effective approach that other organizations may use moving forward to conduct similar research.

A limitation to this portion of the study, which will likely be a limitation in future studies, is the amount of SME time that can be spent on such a process. Increasing the number of focus groups and cognitive interviews will undoubtedly provide more in-depth insight into what drives an item's difficulty. Unfortunately, these processes require access to SMEs, and SME ability to commit time to such a project. This consideration, and maximizing access to SMEs, should remain a priority when planning future item difficulty modeling studies.

Research Question 2

To answer research question 2, both SMEs and Coh-Metrix software were used to create content, cognitive, linguistic, and text-based variables which were then coded to general pediatrics certification items. The background of the SMEs used to define the content- and cognitive variables in the current research was a limitation of the study. The difficulty of the study items was based on the performance of entry-level practitioners; however, the SMEs used in the current research came from a pool of practicing general pediatricians who have experience in item writing for the GP exam. While the SMEs

were instructed to base their ‘walk throughs’ of items on the problem solving processes that an entry-level practitioner would experience, there may have been a gap between what they viewed as driving an item’s difficulty and what actually drove the item’s difficulty for entry-level practitioners taking the exam. A similar limitation was cited by Fulkerson et al. (2011) when examining the cognitive differences between experienced and novice item writers. Future consideration to this limitation should be considered, however using SMEs closer to that of the test taking population may prove difficult as many testing organizations (such as the ABP) only use experienced practitioners for exam development activities.

Following the creation of the six variables, the SMEs coded 96 items with each variable. A second round of coding was then conducted to determine how consistently the SMEs coded each item and variable. Three of the variables were coded using a 1 to 4 scale, two of the variables were coded on a continuous scale, and one variable was coded yes or no. The average rate of agreement for variables using a 1 to 4 scaled was 39.8 percent. The percent of exact and adjacent agreement was 76.4. The average rate of agreement for the two continuous variables was 13.4 percent. While the rate of agreement for the continuous variables was expected to be lower, the 1 to 4 scale variables and yes/no variable had less than desirable rates of agreement. Chaturvedi & Shweta (2015) cite 75 to 90 percent as an acceptable level of agreement. The variables using a 1 to 4 scale and the yes/no variable both had lower rates of agreement than this benchmark. Chaturvedi & Shweta (2015) also recommend using related literature and studies as a benchmark to determine successful rates of agreement. Ferrara et al. (2018) found an

average rater agreement of 76.5 percent (on similar variables with smaller coding scales) which provides a useful benchmark for the current research. These results show that the coding process most likely would have benefited from a group training, which allowed the SMEs to code example items and discuss with each other any coding disagreements. This step should be included in future item difficulty modeling studies which require SME coding of variables so that a higher rate of agreement may be achieved. Additionally, reconvening the SMEs in the current research, to discuss the areas of coding disagreement, is a logical next step before beginning future item difficulty modeling research at the ABP.

The Coh-Metrix variable creation was a straightforward process which required copying and pasting each item (with a hard return placed between the item stem and each option) in the software and recording the output in a Microsoft Excel spreadsheet. Coh-Metrix variables can be seen as an initial step to use in determining if the linguistic or text-based variables coded are significant predictors of an item's difficulty. It should also be noted that Coh-Metrix software is currently free to use, which is appealing considering the other resources required during item difficulty modeling studies.

One limitation to the Coh-Metrix software is the large number of variables which are coded (108). This many variables requires either a large number of items to be analyzed (to avoid overfitting a regression model) or use of a method such as a PCA to reduce the number of variables into components. Neither of these solutions is ideal. Coding a large number of items is time consuming and the availability of that many items for coding may not be plausible. Conducting a PCA, such as was done in the current

research, is a useful workaround, however the components are not easily interpretable (if at all). Nevertheless, significant PCA components, while not interpretable, may provide test developers with a 'screening tool' for item difficulty to determine if an item needs further revisions before being administered.

Another limitation when using the Coh-Metrix software is the online nature of the platform and organizational concerns which may arise from inputting secure test items into such a platform. While Coh-Metrix states that it does not share data that is input into the software, it cannot guarantee the security of data that is input. This limitation impacted the current research, as it restricted the items which could be used in the research to only those that had previously been administered on a non-proctored exam.

Research Question 3

The backward multiple regression conducted to predict Rasch item difficulty based on the SME and Coh-Metrix variables provided mixed results. While both the unconditional model and final reduced model were significant, neither approached the level of .80 variance explained recommended (Bejar, 1983) to avoid pilot testing items. Further, only one of the SME variables, total distractor plausibility, was found to be significant. To better understand the significance of the total distractor plausibility variable, a regression was performed post-hoc, first with the four significant Coh-Metrix components, and then including the distractor plausibility variable. The R-squared was increased by .06 when including the distractor plausibility variable showing that while significant, the variable did not account for a large amount of variance in Rasch item difficulty. The regression results, along with the results obtained from the Qunbar (2019)

study on text complexity as a source of item difficulty in GP test items, indicate the need for further investigation into what variables are driving difficulty in GP test items.

Next steps in this process will include sharing these results with the SMEs that worked on the project to determine if they believe there was anything missed from the discussions and if the coding discrepancies can be reconciled. If the SMEs believe that the variables they defined are still valid, coding additional items (after additional coding training) may be warranted to ensure the current findings were not due to the small sample size or coding discrepancies.

The four Coh-Metrix components will be noted as significant and used in future research on GP item difficulty modeling. An additional next step (with security measures in mind), given the four significant Coh-Metrix component variables, may be to code all ~8,000 items in the GP item bank to the Coh-Metrix variables, followed by a regression analysis (which would not require a PCA). This analysis may provide more interpretable information on the individual linguistic and text-based Coh-Metrix variables which explain significant variance in GP Rasch item difficulties.

One limitation experienced during the process was the selection of items used in the current research. Due to security concerns with Coh-Metrix software, only 197 items were made available for the study (of which 101 were needed to train the Coh-Metrix PCA). This left only 96 items to regress with 15 item difficulty variables. To account for this, both the R-squared and adjusted R-squared (which adjusts for numbers of predictors in the model) were reported in the results. Additionally, the PCA was conducted on 101 items for 108 Coh-Metrix variables, and a more powerful result may have been achieved

if the training set contained more items or did not require a PCA. Alternate regression models, such as CART, may also be considered for future studies if a larger sample size of items is available.

Research Question 4

As noted in the results chapter, the impact study portion of the current research was not carried out due to the low R-squared (.324) and adjusted R-squared (.285) achieved. While these results were significant, there was still a large proportion of Rasch item difficulty which remained unexplained and thus rendering an impact study on using predicted Rasch item difficulties to determine examinee ability levels unwarranted. While the current research did not achieve results, which warranted the impact study, future studies should benefit from such a study if the Bejar (1983) .80 recommendation is achieved. A major goal of item difficulty modeling research is to reduce or eliminate the need for the costly pilot testing of items, and impact studies such as the one outlined in chapter 3, provide researchers with the opportunity to demonstrate to stakeholders the implications on examinees if predicted difficulties are used in place of those based on live testing data.

Significance and Future Directions

Item difficulty modeling remains an important, yet relatively unexplored research topic in certification and licensure exams. Understanding the content, cognitive, linguistic, and text-based variables within an item, which drive the item's difficulty level, will allow testing organizations to reduce costly pilot testing, reduce item discard rates from poor performance, and ensure that items are measuring the intended construct

without artificially creating difficulty by means such as including erroneous information in an item's stem to trick the test taker into an incorrect answer.

As previously stated, the ABP 2019 General Pediatrics certification exam had 159 newly tested items return with poor performance. The current research was significant in that it provided an operational and repeatable process for addressing the issue of poor item performance by defining item difficulty variables in order to predict item difficulty before live testing occurs. In addition to addressing poor performance, the current research methods demonstrated how test developers can take steps to ensure that items are performing at difficulty levels equivalent to the abilities they are intended to measure. Sheehan & Mislevy (1988) first cited this benefit of item difficulty modeling when stating "the onus has been placed (appropriately!) upon the tester to demonstrate that the skills tapped in an educational test are in fact those deemed important to measure." During the focus group discussion when discussing one item which had good performance and a difficulty level near the current cut score, one SME stated:

> They didn't need to write the question this way. I suspect this question writer was a cardiologist and trying to make a simple thing a little more complicated so it wouldn't be so easy for everybody to get it.

Another SME referred to that same item as a 'zebra,' a term used by SMEs to describe items which address either a rare concept or a common concept described in an uncommon way. The significance of discussions such as these is that some items which are performing at an acceptable difficulty may not be performing at a level which matches examinee ability with item difficulty. While exam validity was outside the scope

of the current research, the item difficulty modeling process demonstrated showed how test developers can use the methods to further bolster the validity argument. This concept roots back to the arguments made by advocates of PAD, which were cited in chapter 1 and 2, and future research will benefit from using the principles set forth by PAD methods when conducting item difficulty modeling studies.

While the results of the regression portion of the study were not overwhelmingly positive, the process put forth which addressed research question 1 and 2, is one that may be used by the ABP and other testing organizations moving forward to model item difficulty. Future item difficulty modeling studies can build on the current research by learning from the results and discussions presented in chapters 4 and 5. Specifically, future item difficulty modeling studies should dedicate as much time as possible to the focus group and cognitive interview process. While SME access for item difficulty modeling will be a limitation for many organizations, test developers should be encouraged to plan for the maximum amount of time allowable for SME focus groups and cognitive interviews. Additionally, the coding portion of the current research demonstrates the need for SME training and a consensus among SMEs of how to code different variables. While this will require additional SME time commitment, achieving a high-level of rater agreement is paramount to the success of an item difficulty modeling process such as the one put forth in the current research. Another recommendation for future studies is the use of Coh-Metrix software to code variables (or to code principle components as the current research did) as a first step, as the software is a free resource which can help to explain variables which would otherwise go unidentified by SMEs. It

would benefit test developers to understand the recurring limitation of SME access and time commitment, and to adapt future item review sessions (which mostly occur as in-person meetings and part of regularly scheduled exam development activities) to include targeted questions which aim to explore item difficulty variables. Currently, item review sessions focus mainly on approving items based on content, but strategic planning on facilitating the sessions to focus on both content and cognitive variables which are driving an item's difficulty will allow test developer's to kill two birds with one stone (using the SME time to approve items, and provide information that is gained through focus groups and cognitive interviews, concurrently).

Finally, future studies should examine if reaching the .80 variance explained threshold set forth by Bejar (1983) is necessary, or if a time and cost savings can be realized by using predictive modeling to screen for outlier items which fall on the extreme ends of the difficulty spectrum. With acceptable item difficulties ranging from 35 percent correct to 95 percent correct, this would allow test developers to revise or eliminate those falling outside of that range without the need of live data. This method, while not requiring the Bejar (1983) threshold to be achieved, would create the same advantages of avoiding pilot testing for those items which will ultimately perform either too easy or too difficult.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME]. (1974). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association. American Psychological Association.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association. American Psychological Association.

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

American Psychological Association, & American Educational Research Association. National Council on Measurements Used in Education (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Psychological Bulletin, 51(2), 1-38.

Anderson, R. C. (1982). How to construct achievement tests to assess comprehension. *Review of Educational Research*, 42, 145-170.

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

Bejar, I. I. (1991). A generative approach to psychological and educational measurement. *ETS Research Report Series, 1991*(1), i-54.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). A feasibility study of on-the-fly item generation in adaptive testing. *ETS Research Report Series, 2002*(2), i-44.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational measurement: Issues and practice, 18*(3), 5-12.

Blooms, B. S. (1956). *Taxonomy of Educational Objectives, Handbook 1: The Cognitive Domain*. New York, David McKay Co Inc.Bloor, M., Frankland, J., Thomas, M. & Robson, K. (2001). Analysis. In Bloor, M., Frankland, J., Thomas, M., & Robson, K. *Introducing Qualitative Methods: Focus groups in social research* (pp. 58-73). London: SAGE Publications Ltd doi: 10.4135/9781849209175

Brennan, R. L., & National Council on Measurement in Education. (2006). *Educational measurement*. Praeger Publishers.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review*, 97, 404–431.

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.

Chaturvedi, S. R. B. H., & Shweta, R. C. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods. *Indian Academy of Applied Psychology, 41*(3), 20-27.

Conrad, H. S. (1951). The experimental tryout of test materials. Educational Measurement. Washington, DC: *American Council on Education*, 123.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement (2nd ed.)* (pp. 443-507). Washington, DC: American Council on Education.

Ebel, R. L. (1951). Writing the test item. *Educational measurement*, 185-249.

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

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika*, 49, 175-186.

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.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied psychological measurement*, *11*(2), 175-193.

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

Fein, M. (2012). *Test development: Fundamentals for certification and evaluation*. American Society for Training and Development.

Ferrara, Lai, Reilly, & Nichols. Principled approaches to assessment design, development, and implementation." *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*. First Edition. NJ: John Wiley & Sons, Inc., 2017 Published 2017 by John Wiley & Sons, Inc. 41 – 74.

Ferrara, S., Steedle, J. T., Frantz, R. S. (2018). Item response demands, predicting item difficulty, and validity of inferences from test scores. Measured progress.org. Found at: https://www.measuredprogress.org/wp-content/uploads/2018/05/Item-Response-Demands-Predicting-Item-Difficulty-and-Validity-of-Inferences-from-Test-Scores.pdf

Ferrara, S., Svetina, D., Skucha, S., & Davidson, A. H. (2011). Test development with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practic*e, *30*(4), 3-15.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32, 221-233.

Fulkerson, D., Mittelholtz, D.J., & Nichols, P. D. (April, 2009). *The psychology of writing items: Improving figural response item writing*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Fulkerson, D., Nichols, P., & Mittelholtz, D. (2010, May). What item writers think when writing items: Towards a theory of item writing expertise. In *annual meeting of the American Educational Research Association*, Denver, CO.

Furter, R. (2015). *Principled assessment as a foundation for standard setting*. Unpublished doctoral dissertation. University of North Carolina at Greensboro, NC.

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.

Graf, E. A., Peterson, S., Steffen, M., & Lawless, R. (2005). Psychometric and cognitive analysis as a basis for the design and revision of quantitative item models. *ETS Research Report Series, 2005*(2), i-33.

Grasser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193-202.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix Software. Retrieved from: www.cohmetrix.com.

Guion, R. M. (1977). Content validity—The source of my discontent. *Applied Psychological Measurement, 1*(1), 1-10.

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., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Keehner, M., Gorin, J. S., Feng, G., & Katz, I. R. (2017). Developing and validating cognitive models in assessment. *The handbook of cognition and assessment: Frameworks, methodologies, and applications*, 75-101.

Kirsch, I., & Jungeblut, A. (1986). Literacy profiles of young adults. *Educational Testing Service, Princeton, NJ*.

Kirsch, I. S., & Mosenthal, P. B. (1988). Understanding document literacy: Variables underlying the performance of young adults. *ETS Research Report Series, 1988*(2), i-67.

Kirsch, I. S.; & Mosenthal, P. B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly*, 25, pp. 5-30.

Kofman, J. (2019). Trint Transcription Software. Retrieved from: https://trint.com/about-us/

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

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

Lane, S., Raymond, M. R., Haladyna, T. M., & Downing, S. M. (2015). Test development process. In *Handbook of test development* (pp. 19-34). Routledge.

Lenzner, T., Neuert, C., & Otto, W. (2016). Cognitive Pretesting (Version 2.0).

Lenzner, T., Neuert, C., & Otto, W. (2016). *GESIS Survey Guideline*. Mannheim, Germany: GESIS-Leibnitz Institute for the Social Sciences.

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. (2013). Automatic item generation for computerized adaptive testing. In *Automatic item generation: Theory and practice* (pp. 196-216). Routledge, New York, NY.

Luecht, R. M., Burke, M. (2019). *Reconceptualizing items: from automatic item generation to task model families*. Paper presented October 2017 at MARCES 2017, College Park MD. Manuscript submitted January 2019.

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.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5-11.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In RL Linn (Ed.), *The American Council on Education/Macmillan series on higher education*. Educational measurement (pp. 335-366). New York, NY, England: Macmillan Publishing Co, Inc; American Council on Education.

Mislevy, Robert J. Evidence and inference in educational assessment. *Psychometrika 59.4* (1994): 439-483.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence centered design. *ETS Research Report Series*, *2003*(1), i-29.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6-20.

Mosenthal, P. B., & Kirsch, I. S. (1991). Toward an explanatory model of document literacy. *Discourse Processes, 14*(2), 147-180.

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.

Nichols, Kobrin, Lai, & Koepfler. The role of theories of learning and cognition in assessment design and development." *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*. First Edition. NJ: John Wiley & Sons, Inc., 2017 Published 2017 by John Wiley & Sons, Inc. 41 – 74.

Pellegrino, J. W., Mumaw, R., and Shute, V. 1985. Analyses of spatial aptitude and expertise. In S. Embretson (Ed.), *Test Design: Developments in Psychology and Psychometrics* (pp. 45-76). New York: Academic Press.

Qunbar, S. (2019). *Automatic item difficulty modeling with test item representations*. Unpublished doctoral dissertation. University of North Carolina at Greensboro, NC.

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.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing, 1*(3-4), 185-216.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105-126.

Sheehan, K. M., & Ginther, A. (2001, April). What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section. Paper presented *at the annual meeting of the National Council on Measurement in Education*, New Orleans, LA.

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

Sheehan, K., & Mislevy, R. J. (1994). A Tree-Based Analysis of Items from an Assessment of Basic Mathematics Skills.

Taylor, S. J., Bogdan, R., & DeVault, M. (2015). Introduction to qualitative research methods: A guidebook and resource. John Wiley & Sons.

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

Watkins, Daphne, and Deborah Gioia. Mixed methods research. OUP Us, 2015.

Whitely, S. E. 1980. Multicomponent latent trait models for ability tests. *Psychometrika*, 45, 479-494.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge.

APPENDIX A

COH-METRIX VARIABLE DEFINITIONS (GRASSER ET AL., 2004)

Number of paragraphs (DESPC)
This is the total number of paragraphs in the text. Paragraphs are simply delimited by a hard return.

Number of sentences (DESSC)
This is the total number of sentences in the text. Sentences are identified by the OpenNLP sentence splitter (http://opennlp.sourceforge.net/projects.html).

Number of words (DESWC)
This is the total number of words in the text. Words are calculated using the output from the Charniak parser. For each sentence, the Charniak parser generates a parse tree with part of speech (POS) tags for clauses, phrases, words and punctuations. The elements on the leaves of a parse tree are tagged words or punctuations. In Coh-Metrix, words are taken from the leaves of the sentence parse trees.

Mean length of paragraphs (DESPL)
This is the average number of sentences in each paragraph within the text. Longer paragraphs may be more difficult to process.

Standard deviation of the mean length of paragraphs (DESPLd)
This is the standard deviation of the measure for the mean length of paragraphs within the text. In the output, d is used at the end of the name of the indices to designate that it is a standard deviation. A large standard deviation indicates that the text has large variation in terms of the lengths of its paragraphs, such that it may have some very short and some very long paragraphs. The presence of headers in a short text can increase values on this measure.

Mean number of words (length) of sentences in (DESSL)
This is the average number of words in each sentence within the text, where a word is anything that is tagged as a part-of-speech by the Charniak parser. Sentences with more words may have more complex syntax and may be more difficult to process. While this is a descriptive measure, this also provides one commonly used proxy for syntactic complexity. However, Coh-Metrix provides additional more precise measures of syntactic complexity discussed later in this chapter.

Standard deviation of the mean length of sentences (DESSLd)
This is the standard deviation of the measure for the mean length of sentences within the text. A large standard deviation indicates that the text has large variation in terms of the lengths of its sentences, such that it may have some very short and some very long

sentences. The presence of headers in a short text may impact this measure. Narrative text may also have variations in sentence length as authors move from short character utterances to long descriptions of scenes.

Mean number of syllables (length) in words (DESWLsy)
Coh-Metrix calculates the average number of syllables in all of the words in the text. Shorter words are easier to read and the estimate of word length serves as a common proxy for word frequency.

Standard deviation of the mean number of syllables in words (DESWLsyd)
This is the standard deviation of the measure for the mean number of syllables in the words within the text. A large standard deviation indicates that the text has large variation in terms of the lengths of its words, such that it may have both short and long words.

Mean number of letters (length) in words (DESWLlt)
This is the average number of letters for all of the words in the text. Longer words tend to be lower in frequency or familiarity to a reader.

Standard deviation of the mean number of letter in words (DESWLltd)
This is the standard deviation of the measure for the mean number of letters in the words within the text. A large standard deviation indicates that the text has large variation in terms of the lengths of its words, such that it may have both short and long words.

Narrativity: PCNARz, PCNARp
Narrative text tells a story, with characters, events, places, and things that are familiar to the reader. Narrative is closely affiliated with everyday, oral conversation. This robust component is highly affiliated with word familiarity, world knowledge, and oral language. Non-narrative texts on less familiar topics lie at the opposite end of the continuum.

Syntactic Simplicity: PCSYNz, PCSYNp
This component reflects the degree to which the sentences in the text contain fewer words and use simpler, familiar syntactic structures, which are less challenging to process. At the opposite end of the continuum are texts that contain sentences with more words and use complex, unfamiliar syntactic structures.

Word Concreteness: PCCNCz, PCCNCp
Texts that contain content words that are concrete, meaningful, and evoke mental images are easier to process and understand. Abstract words represent concepts that are difficult to represent visually. Texts that contain more abstract words are more challenging to understand.

Referential Cohesion: PCREFz, PCREFp
A text with high referential cohesion contains words and ideas that overlap across sentences and the entire text, forming explicit threads that connect the text for the reader. Low cohesion text is typically more difficult to process because there are fewer connections that tie the ideas together for the reader.

Deep Cohesion: PCDCz, PCDCp
This dimension reflects the degree to which the text contains causal and intentional connectives when there are causal and logical relationships within the text. These connectives help the reader to form a more coherent and deeper understanding of the causal events, processes, and actions in the text. When a text contains many relationships but does not contain those connectives, then the reader must infer the relationships between the ideas in the text. If the text is high in deep cohesion, then those relationships and global cohesion are more explicit.

Verb Cohesion: PCVERBz, PCVERBp
This component reflects the degree to which there are overlapping verbs in the text. When there are repeated verbs, the text likely includes a more coherent event structure that will facilitate and enhance situation model understanding. This component score is likely to be more relevant for texts intended for younger readers and for narrative texts (McNamara, Graessar, &Louwerse, 2012).

Connectivity: PCCONNz, PCCONNp
This component reflects the degree to which the text contains explicit adversative, additive, and comparative connectives to express relations in the text. This component reflects the number of logical relations in the text that are explicitly conveyed. This score is likely to be related to the reader's deeper understanding of the relations in the text.

Temporality: PCTEMPz, PCTEMPp
Texts that contain more cues about temporality and that have more consistent temporality (i.e., tense, aspect) are easier to process and understand. In addition, temporal cohesion contributes to the reader's situation model level understanding of the events in the text.

Noun overlap (CRFNO1 and CRFNOa)
These are measures of local and global overlap between sentences in terms of nouns. Adjacent noun overlap (CRFNO1) represents the average number of sentences in the text that have noun overlap from one sentence back to the previous sentence. Among the co-reference measures, it is the most strict, in the sense that the noun must match exactly, in form and plurality. Whereas local overlap considers only adjacent sentences, global overlap (CRFNOa) considers the overlap of each sentence with every other sentence. As shown in Table 4.1, just over 50 percent of the adjacent sentences contained an overlapping noun, and 43 percent of the sentence pairs in the text contained an overlapping noun when comparing all of the sentences (global overlap).

Argument overlap (CRFAO1 and CRFAOa)
These local and global overlap measures are similar to noun overlap measures, but include overlap between sentences in terms of nouns and pronouns. Argument overlap occurs when there is overlap between a noun in one sentence and the same noun (in singular or plural form) in another sentence; it also occurs when there are matching personal pronouns between two sentences (e.g., he/he). The term argument is used in a linguistic sense, where noun/pronoun arguments are contrasted with verb/adjective predicates (Kintsch & Van Dijk, 1978). Consider argument overlap for the science passage in Table 4.1 in the second column. Note that in comparison to noun overlap, it is less strict because it considers the overlap for example between cells and cell. Argument and stem overlap would also include overlap between pronouns, such as it to it, or he to he, which noun overlap does not include.

Stem overlap (CRFSO1, CRFSOa)
These two local and global overlap measures relax the noun constraint held by the noun and argument overlap measures. A noun in one sentence is matched with a content word (i.e., nouns, verbs, adjectives, adverbs) in a previous sentence that shares a common lemma (e.g., tree/treed; mouse/mousey; price/priced). Notably, the outcome for stem and argument overlap in Table 4.1 were identical; however, this will not always be the case.

Content word overlap (CRFCWO1, CRFCWO1d, CRFCWOa, CRFCWOad)
This measure considers the proportion of explicit content words that overlap between pairs of sentences. For example, if a sentence pair has fewer words and two words overlap, the proportion is greater than if a pair has many words and two words overlap. This measure includes both local (CRFCWO1) and global (CRFCWOa) indices, and also includes their standard deviations (CRFCWO1d, CRFCWOad). In the example provided in Table 4.1, the content word overlap both locally and globally was lower than that estimated by the binary overlap scores. This measure may be particularly useful when the lengths of the sentences in the text are a principal concern.

Anaphor overlap (CRFANP1, CRFANPa)
This measure considers the anphor overlap between pairs of sentences. A pair of sentences has an anphor overlap if the later sentence contains a pronoun that refers to a pronoun or noun in the earlier sentence. The score for each pair of sentences is binary, i.e., 0 or 1. The measure of the text is the average of the pair scores. This measure includes both local (CRFANP1) and global (CRFANPa) indice.

LSA sentence adjacent: LSASS1
This index computes mean LSA cosines for adjacent, sentence-to-sentence (abbreviated as "ass") units. This measures how conceptually similar each sentence is to the next sentence.

LSASS1d: This index computes standard deviation of LSA cosines for adjacent, sentence-to-sentence (abbreviated as "ass") units. This measures how consistent adjacent sentences are overlapped semantically.

LSA sentence all: LSASSp
Like LSA sentence adjacent (LSAassa), this index computes mean LSA cosines. However, for this index all sentence combinations are considered, not just adjacent sentences. LSApssa computes how conceptually similar each sentence is to every other sentence in the text.

LSASSpd
This index computes the standard deviation of LSA cosine of all sentence pairs within paragraphs.

LSAPP1
This index computes the mean of the LSA cosines between adjacent paragraphs.

LSAPP1d
This index is the standard deviation of LSA cosinces between adjacent paragraphs.

LSAGN
This is the average givenness of each sentence.

LSAGNd
This is the standard deviation of giveness of each sentence.

Type-token ratio: LDTTRc
Type-token ratio (TTR) (Templin, 1957) is the number of unique words (called types) divided by the number of tokens of these words. Each unique word in a text is considered a word type. Each instance of a particular word is a token. For example, if the word dog appears in the text 7 times, its type value is 1, whereas its token value is 7. When the type-token ratio approaches 1, each word occurs only once in the text; comprehension should be comparatively difficult because many unique words need to be decoded and integrated with the discourse context. As the type-token ratio decreases, words are repeated many times in the text, which should increase the ease and speed of text processing. Type-token ratios are computed for content words, but not function words. TTR scores are most valuable when texts of similar lengths are compared.

LDTTRa
Type token ratio for all words.

LDMTLDa
MTLD lexcical diversity measure for all words.

LDVOCDa
VOC lexical diversity measure for all words.

All connectives, CNCAll: This is the incidence of all connectives.

Causal Connectives: CNCCaus
This is the incidence score of causal connectives. Among the various types of connectives, only causal connectives (CNCCaus) discriminated between the high and low cohesion texts, presumably because the researchers who created the texts primarily manipulated causal cohesion and not additive, temporal, or clarification connectives.

CNCLogic
This is the incidence score of logic connectives.

CNCADC
This is the incidence score of adversative/contrastive connectives.

CNCTemp
This is the incidence score of temporal connectives.

CNCTempx
This is the incidence score of extended temporal connectives.

CNCAdd
This is the incidence score of additive connectives.

CNCPos
This is the incidence score of positive connectives.

CNCNeg
This is the incidence score of negative connectives.

SMCAUSv
This is the incidence score of causal verbs.

Causal content: SMCAUSvp
This is the incidence of causal verbs and causal particles in text.

Intentional content: SMINTEp
This is the incidence of intentional actions, events, and particles (per thousand words).

Causal cohesion: SMCAUSr
This is a ratio of causal particles (P) to causal verbs (V). The denominator is incremented by the value of 1 to handle the rare case when there are 0 causal verbs in the text.

Cohesion suffers when the text has many causal verbs (signifying events and actions) but few causal particles that signal how the events and actions are connected.

Intentional cohesion: SMINTEr
This is the ratio of intentional particles to intentional actions/events.
SMCAUSlsa
This is the LSA overlap between verbs.

SMCAUSwn
This is the WordNet overlap between verbs.

Temporal cohesion: SMTEMP
This is the repetition score for tense and aspect. The repetition score for tense is averaged with the repetition score for aspect.

Words before main verb: SYNLE
This is the mean number of words before the main verb of the main clause in sentences. This is a good index of working memory load.

Modifiers per NP: SYNNP
This is the mean number of modifiers per noun-phrase.

SYNMEDpos
This is the mean minimum editorial distance score between adjacent sentences computed from part of speech tags. Notice that the editing actions were performed on POS tags in two sentences instead of letters in two words. See Coh-Metrix book for details.

SYNMEDwrd
This is the minimum editorial distance score between adjacent sentences computed from words. Notice that the editing actions were performed on words in two sentences instead of letters in two words. See Coh-Metrix book for details.

SYNMEDlem
This is the minimum editorial distance score between adjacent sentences from lemmas. Notice that the editing actions were performed on lemmas in two sentences instead of letters in two words. See Coh-Metrix book for details.

Syntactic structure similarity adjacent: SYNSTRUTa
This is the proportion of intersection tree nodes between all adjacent sentences.

Syntactic structure similarity all 01: SYNSTRUTt
This is the proportion of intersection tree nodes between all sentences and across paragraphs.

DRNP
This is the incidence score of noun phrases.

DRVP
This is the incidence score of verb phrases.

DRAP
This is the incidence score of adverbial phrases.

DRPP
This is the incidence score of preposition phrases.

DRPVAL
This is the incidence score of agentless passive voice forms.

Negations: DRNEG
This is the incidence score for negation expressions.

DRGERUND
This is the incidence score of gerunds.

DRINF
This is the incidence score of infinitives.

WRDNOUN
This is the incidence score of nouns.

WRDVERB
This is the incidence score of verbs.

WRDADJ
This is the incidence score of adjectives.

WRDADV
This is the incidence score of adverbs.

Personal pronoun: WRDPRO
This is the number of personal pronouns per 1000 words. A high density of pronouns can create referential cohesion problems if the reader does not know what the pronouns refer to.

WRDPRP1s
This is the incidence score of pronouns, first person, single form.

WRDPRP1p
This is the incidence score of pronouns, first person, plural form.

WRDPRP2
This is the incidence score of pronouns, second person.

WRDPRP3s
This is the incidence score of pronouns, third person, single form.

WRDPRP3p
This is the incidence score of pronouns, third person, plural form.

WRDFRQc
This is the average word frequency for content words.

WRDFRQa
This is the average word frequency for all words.

WRDFRQmc
This is the average minimum word frequency in sentences.

Age of acquisition (WRDAOAc)
Coh-Metrix includes the age of acquisition norms from MRC which were compiled by Gilhooly and Logie (1980) for 1903 unique words. The c at the end of the index name indicates that it is calculated for the average ratings for content words in a text. Age of acquisition reflects the notion that some words appear in children's language earlier than others. Words such as cortex, dogma, and matrix (AOA= 700) have higher age-of-acquisition scores than words such as milk, smile, and pony (AOA =202). Words with higher age-of-acquisition scores denote spoken words that are learned later by children.

Familiarity (WRDFAMc)
This is a rating of how familiar a word seems to an adult. Sentences with more familiar words are words that are processed more quickly. MRC provides ratings for 3488 unique words. Coh-Metrix provides the average ratings for content words in a text. Raters for familiarity provided ratings using a 7-point scale, with 1 being assigned to words that they never had seen and 7 to words that they had seen very often (nearly every day). The ratings were multiplied by 100 and rounded to integers.

Concreteness (WRDCNCc)
This is an index of how concrete or non-abstract a word is. Words that are more concrete are those things you can hear, taste, or touch. MRC provides ratings for 4293 unique words. Coh-Metrix provides the average ratings for content words in a text. Words that score low on the concreteness scale include protocol (264) and difference (270) compared to box (597) and ball (615).

Imagability (WRDIMGc)
An index of how easy it is to construct a mental image of the word is also provided in the merged ratings of the MRC, which provides ratings for 4825 words. Coh-Metrix provides the average ratings for content words in a text. Examples of low imagery words are reason (285), dogma (327), and overtone (268) compared to words with high imagery such as bracelet (606) and hammer (618).

Meaningfulness (WRDMEAc)
These are the meaningfulness ratings from a corpus developed in Colorado by Toglia and Battig (1978). MRC provides ratings for 2627 words. Coh-Metrix provides the average ratings for content words in a text. An example of meaningful word is people (612) as compared to abbess (218). Words with higher meaningfulness scores are highly associated with other words (e.g., people), whereas a low meaningfulness score indicates that the word is weakly associated with other words.

Polysemy (WRDPOLc)
Polysemy refers to the number of senses (core meanings) of a word. For example, the word bank has at least two senses, one referring to a building or institution for depositing money and the other referring to the side of a river. Coh-Metrix provides average polysemy for content words in a text. Polysemy relations in WordNet are based on synsets (i.e., groups of related lexical items), which are used to represent similar concepts but distinguish between synonyms and word senses (Miller et al., 1990). These synsets allow for the differentiation of senses and provide a basis for examining the number of senses associated with a word. Coh-Metrix reports the mean WordNet polysemy values for all content words in a text. Word polysemy is considered to be indicative of text ambiguity because the more senses a word contains relates to the potential for a greater number of lexical interpretations. However, more frequent words also tend to have more meanings, and so higher values of polysemy in a text may be reflective of the presence of higher frequency words.

Hypernymy (WRDHYPn, WRDHYPv, WRDHYPnv)
Coh-Metrix also uses WordNet to report word hypernymy (i.e., word specificity). In WordNet, each word is located on a hierarchical scale allowing for the measurement of the number of subordinate words below and superordinate words above the target word. Thus, entity, as a possible hypernym for the noun chair, would be assigned the nu1mber 1. All other possible hyponyms of entity as it relates to the concept of a chair (e.g., object, furniture, seat, chair, camp chair, folding chair) would receive higher values (see also Chapter 2). Similar values are assigned for verbs (e.g., hightail, run, travel). As a result, a lower value reflects an overall use of less specific words, while a higher value reflects an overall use of more specific words. Coh-Metrix provides estimates of hypernymy for nouns (WRDHYPn), verbs (WRDHYPv), and a combination of both nouns and verbs (WRDHYPnv).

Flesch Reading Ease: RDFRE
The output of the Flesch Reading Ease formula is a number from 0 to 100, with a higher score indicating easier reading. The average document has a Flesch Reading Ease score between 6 and 70.

Flesch_Kincaid Grade Level: RDFKGL
This more common Flesch-Kincaid Grade Level formula converts the Reading Ease Score to a U.S. grade-school level. The higher the number, the harder it is to read the text. The grade levels range from 0 to 12.

RDL2
This is the second language readability score.

APPENDIX B

SME VARIABLE DEFINITIONS AND CODING INSTRUCTIONS

Variables

1. **Diagnosis/Treatment Seen/Trained On (1 – 4)**
    a. Using a 1 to 4 scale (1 being a few times in your career or a few times in the course of training and 4 being several times per month in practice or training) how commonly is the item's diagnosis or patient presentation **seen** in practice or touched on in training?
    b. N/A is provided as an option and may be used if an item addresses an area such as ethics or scholarly activities.
2. **Patient Presentation (1 – 4)**
    a. Using a 1 to 4 scale (1 being rare presentation and 4 being a "classical" presentation) how commonly is the patient presentation seen with the subsequent diagnosis or management being asked for?
    b. N/A is provided as an option and may be used if an item addresses an area such as ethics or scholarly activities.
3. **Number of pieces of information needed**
    a. How many pieces of information in the stem are required to correctly answer the item?
        i. Each individual piece of information should be counted when coding for this variable. If five lab values are provided, and all five are required to answer the item correctly, then those would count as 5 pieces of information (rather than 1). If five symptoms or patient history pieces of information are provided, each individual one that is needed to answer the item should be counted.
4. **Number of extraneous info NOT needed**
    a. How many pieces of extraneous information are present in the stem which do not help answer the item correctly?
    b. Each individual piece of information should be counted when coding for this variable. If five lab values are provided, and three are required to answer correctly but two of the values are not needed, then two should be counted.
5. **Pattern Recognition or Keywords (Yes or No)**
    a. Are there keywords or patterns that are specifically addressed in training that are used to correctly answer the item?

Variables 6 through 9
Option A through D Plausibility (1 to 4)

    b. For each option, using a 1 to 4 scale (1 being not plausible and easily eliminated and 4 being highly plausible and difficult to differentiate from the correct answer) how plausible is each distractor?
    c. This variable should be left blank for the correct answer choice.

APPENDIX C

REGRESSION COEFFICIENTS AND COEFFICIENT CORRELATIONS

Regression Coefficients, Unconditional Model

Model	Unstandardized Coefficients		Standardized Coefficients			Correlations		
	B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part
(Constant)	-2.064	.995		-2.075	.041			
Diagnosis & Treatment	.038	.171	.028	.221	.825	.047	.025	.020
Seen/Trained On Patient Presentation	.109	.233	.056	.469	.640	.008	.053	.042
Information Needed to Answer	.011	.071	.017	.151	.880	.029	.017	.014
Extraneous Information	-.018	.086	-.026	-.209	.835	-.081	-.024	-.019
Pattern/Keyword Recognition	-.361	.268	-.131	-1.348	.182	-.143	-.152	-.121
Total Distractor Score	.225	.102	.222	2.214	.030	.325	.245	.199
Coh-Metrix Factor 1	.082	.141	.060	.582	.562	.044	.066	.052
Coh-Metrix Factor 2	-.254	.148	-.181	-1.716	.090	-.127	-.192	-.154
Coh-Metrix Factor 3	.446	.188	.282	2.366	.020	.218	.260	.213
Coh-Metrix Factor 4	-.082	.147	-.059	-.556	.580	-.015	-.063	-.050
Coh-Metrix Factor 5	.363	.154	.274	2.348	.021	.216	.259	.211
Coh-Metrix Factor 6	.441	.138	.320	3.205	.002	.313	.343	.288
Coh-Metrix Factor 7	-.314	.145	-.213	-2.168	.033	-.182	-.240	-.195
Coh-Metrix Factor 8	-.014	.138	-.009	-.100	.920	.028	-.011	-.009
Coh-Metrix Factor 9	-.049	.134	-.035	-.362	.719	.075	-.041	-.032

Regression Coefficients, Final Model

Model	Unstandardized Coefficients		Standardized Coefficients			Correlations		
	B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part
(Constant)	-1.868	.579		-3.226	.002			
Total Distractor Score	.250	.091	.247	2.754	.007	.325	.283	.243
Coh-Metrix Factor 3	.392	.146	.248	2.687	.009	.218	.277	.237
Coh-Metrix Factor 5	.281	.119	.212	2.365	.020	.216	.246	.208
Coh-Metrix Factor 6	.485	.126	.351	3.832	.000	.313	.380	.338
Coh-Metrix Factor 7	-.277	.133	-.188	-2.084	.040	-.182	-.218	-.184

Coefficient Correlations, Unconditional Model

Model		Factor 9	Factor 2	Pattern Recognition	Number of pieces of information	Patient Presentation	Factor 8	Total Distractor Score	Factor 7	Factor 1	Factor 6	Factor 4	Factor 5	Factor 3	Extraneous info	Diagnosis/Treatment Seen/Trained On
Correlations	Factor 9	1.000	.060	-.079	-.078	-.048	.122	-.042	.196	-.105	-.088	.203	-.021	-.099	-.028	.128
	Factor 2	.060	1.000	-.101	-.120	.038	.019	.105	.042	-.256	.251	-.080	-.107	-.286	.081	-.088
	Pattern Recognition	-.079	-.101	1.000	.052	-.006	.041	-.048	-.044	.187	.008	.057	-.117	.132	-.169	-.067
	Number of pieces of information	-.078	-.120	.052	1.000	-.019	.037	-.068	.024	.109	.042	-.171	-.409	.452	.355	.060
	Patient Presentation	-.048	.038	-.006	-.019	1.000	.010	-.123	-.082	.072	.199	-.090	.060	.075	.031	-.624
	Factor 8	.122	.019	.041	.037	.010	1.000	.031	.009	.107	-.028	-.085	-.169	.000	-.011	-.016
	Total Distractor Score	-.042	.105	-.048	-.068	-.123	.031	1.000	.064	-.089	-.090	-.211	-.148	-.114	.297	.038
	Factor 7	.196	.042	-.044	.024	-.082	.009	.064	1.000	.079	-.024	.213	-.187	.111	.102	.070
	Factor 1	-.105	-.256	.187	.109	.072	.107	-.089	.079	1.000	.049	-.066	-.064	.203	.003	-.258
	Factor 6	-.088	.251	.008	.042	.199	-.028	-.090	-.024	.049	1.000	-.103	-.091	.175	.054	-.191

	Factor 4	.203	-.080	.057	-.171	-.090	-.085	-.211	.213	-.066	-.103	1.000	.218	.023	-.326	.184
	Factor 5	-.021	-.107	-.117	-.409	.060	-.169	-.148	-.187	-.064	-.091	.218	1.000	-.284	-.494	-.020
	Factor 3	-.099	-.286	.132	.452	.075	.000	-.114	.111	.203	.175	.023	-.284	1.000	.337	.036
	Extraneous info	-.028	.081	-.169	.355	.031	-.011	.297	.102	.003	.054	-.326	-.494	.337	1.000	-.037
	Diagnosis/Treatment Seen/Trained On	.128	-.088	-.067	.060	-.624	-.016	.038	.070	-.258	-.191	.184	-.020	.036	-.037	1.000
	Factor 9	.018	.001	-.003	-.001	-.001	.002	-.001	.004	-.002	-.002	.004	.000	-.003	.000	.003
	Factor 2	.001	.022	-.004	-.001	.001	.000	.002	.001	-.005	.005	-.002	-.002	-.008	.001	-.002
	Pattern Recognition	-.003	-.004	.072	.001	.000	.002	-.001	-.002	.007	.000	.002	-.005	.007	-.004	-.003
	Number of pieces of information	-.001	-.001	.001	.005	.000	.000	.000	.000	.001	.000	-.002	-.004	.006	.002	.001
Covariances	Patient Presentation	-.001	.001	.000	.000	.054	.000	-.003	-.003	.002	.006	-.003	.002	.003	.001	-.025
	Factor 8	.002	.000	.002	.000	.000	.019	.000	.000	.002	-.001	-.002	-.004	1.340E-6	.000	.000
	Total Distractor Score	-.001	.002	-.001	.000	-.003	.000	.010	.001	-.001	-.001	-.003	-.002	-.002	.003	.001
	Factor 7	.004	.001	-.002	.000	-.003	.000	.001	.021	.002	.000	.005	-.004	.003	.001	.002

Factor 1	-.002	-.005	.007	.001	.002	.002	-.001	.002	.020	.001	-.001	-.001	.005	3.827E-5	-.006
Factor 6	-.002	.005	.000	.000	.006	-.001	-.001	.000	.001	.019	-.002	-.002	.005	.001	-.004
Factor 4	.004	-.002	.002	-.002	-.003	-.002	-.003	.005	-.001	-.002	.022	.005	.001	-.004	.005
Factor 5	.000	-.002	-.005	-.004	.002	-.004	-.002	-.004	-.001	-.002	.005	.024	-.008	-.007	-.001
Factor 3	-.003	-.008	.007	.006	.003	1.340E-6	-.002	.003	.005	.005	.001	-.008	.035	.005	.001
Extraneous info	.000	.001	-.004	.002	.001	.000	.003	.001	3.827E-5	.001	-.004	-.007	.005	.007	-.001
Diagnosis/Treatment Seen/Trained On	.003	-.002	-.003	.001	-.025	.000	.001	.002	-.006	-.004	.005	-.001	.001	-.001	.029

Coefficient Correlations, Final Model

Model		Total Distractor Score	Factor 7	Factor 6	Factor 5	Factor 3
Correlations	Total Distractor Score	1.000	.079	-.108	-.026	-.140
	Factor 7 Variable	.079	1.000	-.007	-.178	.079
	Factor 6 Variable	-.108	-.007	1.000	-.059	.262
	Factor 5 Variable	-.026	-.178	-.059	1.000	-.074
	Factor 3 Variable	-.140	.079	.262	-.074	1.000
Covariances	Total Distractor Score	.008	.001	-.001	.000	-.002
	Factor 7 Variable	.001	.018	.000	-.003	.002
	Factor 6 Variable	-.001	.000	.016	-.001	.005
	Factor 5 Variable	.000	-.003	-.001	.014	-.001
	Factor 3 Variable	-.002	.002	.005	.001	.021