AN ITEM STIMULUS APPROACH TO UNDERSTANDING TEST ITEM DIFFICULTY

by

Victoria Blanshteyn

A dissertation submitted to the Graduate Faculty in Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2012

This manuscript has been read and accepted for the Graduate Faculty in Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

December 15, 2011
Date

Dr. Charles Scherbaum
Chair of the Examining Committee

December 15, 2011
Date

Dr. Maureen O'Connor
Chair of the Examining Committee

Dr. Joel Lefkowitz

Dr. Lise Saari

Dr. Harold Goldstein

Dr. Kristen Shockley

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

AN ITEM STIMULUS APPROACH TO UNDERSTANDING TEST ITEM DIFFICULTY

by

Victoria Blanshteyn

Advisor: Professor Charles Scherbaum

Understanding what makes test items difficult is an important step in understanding how individuals solve items on a test and in mapping the cognitive processes that are involved. However, there remains a gap in understanding how general stimulus features of items (e.g., length of a test item) impact the difficulty of items for a range of item types. In an effort to reduce this gap, the current study tested the impact of item stimulus features on item difficulty. The proposed difficulty framework utilized the radical and incidental approach of item generation theory (e.g., Irvine, Dann, & Anderson, 1990), which allows items to be decomposed into the factors that are hypothesized to impact difficulty as well as examine the impact of different item stimulus features on difficulty. To test the proposed framework, the current paper incorporated linear latent trait modeling (Fischer, 1973), an IRT-based analytical approach that expresses item difficulty in terms of underlying factors of stimulus complexity rather than individual parameters. Results indicate that certain item stimulus features, including language ambiguity, negative wording, constructed-response items, and colloquial knowledge impact item difficulty. Implications for test development are discussed.

Acknowledgements

I extend my gratitude to my committee members, Drs. Charles Scherbaum, Joel Lefkowitz, and Lise Saari. Thank you for your encouragement, critical feedback and valuable insights. I also extend my gratitude to Drs. Harold Goldstein and Kristen Shockley, who served as outside readers.

I would especially like to thank my advisor and committee chair, Dr. Charles Scherbaum for his dedication, guidance and constant support through the many stages of the dissertation process. I am honored to have worked with you.

My deep gratitude goes to my parents whose profound influence has shaped me as a person and whose encouragement and support has motivated me in all phases of my life. You sacrificed much on my behalf and for my success and I am forever grateful.

I would like to thank my husband, Max, for his emotional support and encouragement during this process. You have made it possible for me to complete this degree and your outlook on life has kept me sane.

Finally, I would like to thank my daughter, Evelyn, who is the source of my inspiration and who has given me a new perspective on priorities.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Standardized tests and measurements are used primarily to distinguish between the ability or skill level of individuals taking those tests. These tests consist of a number of items that are written to assess particular aspects of the ability or skill domain. The items on these tests are critical components of standardized assessment (Lievens & Sackett, 2007) as item quality has a direct impact on the measurement properties and usefulness of the resulting test scores. The process of developing these test items is a daunting task (Lievens & Sackett, 2007), especially in mass testing, where there is a need for multiple parallel versions of a test. There are a number of challenges that test developers face in generating these tests including developing test construction approaches, writing large numbers of items, and pretesting items (Lievens & Sackett, 2007). An issue underlying all of these challenges is to understand how individuals of differing abilities approach and solve items on these tests and how it relates to the resulting performance on the items. At the core of this issue is the role that items and their properties play in this process. Essentially, the issue is to understand the sources of item difficulty.

It is well known that tests need to include items that vary in difficulty to assess different levels of ability accurately. However, at this point, we do not have a clear knowledge of the factors that influence item difficulty despite several attempts to understand them (e.g., Berk, Lohman, & Cassata, 2001; Carpenter, Just, & Shell, 1990; Embretson, 1995). Knowing what makes items difficult is an important step in understanding how individuals solve items on a test and in mapping the cognitive processes that are involved (e.g., Carpenter et al., 1990; Embretson, 1995). As previously mentioned, knowledge of what makes items difficult is also important in mass testing, where there is a need for mass generation of items of similar difficulty levels. Thus,

understanding what makes items difficult is important from the point of view of cognitive processing and mass testing.

A number of frameworks about the sources of item difficulty have been offered (e.g., Carpenter et al., 1990; Freedle & Kostin, 1997; Gorin & Embretson, 2006; Irvine, Dann, & Anderson, 1990; Roccas & Moshinsky, 2003). These frameworks can be categorized into those that focus on cognitive processes and those that focus on item stimulus features. Most of the existing frameworks have focused primarily on item difficulty from a cognitive processing point of view. For example, in a difficulty framework designed for the Raven's Progressive Matrices, Carpenter et al. (1990) illustrated that more difficult problems require abstract reasoning and goal management. These frameworks define difficulty in terms of cognition of the individual completing test items. From the point of view of these theories, the amount of cognitive processing necessary to respond correctly to a test item is what makes item difficult. The insights from this research are beneficial for developing theories of cognitive information processing and training examinees to approach test items appropriately. These frameworks, however, do not define difficulty based on item properties.

From the point of view of item writing and test development, there is also a need to understand how stimulus features of items impact difficulty. Furthermore, understanding what item stimulus features impact item difficulty is an important step in mapping the cognitive processes that are involved (Carpenter et al., 1990; Embretson, 1995). Nonetheless, among the theories on item difficulty, there are only a few frameworks that have stimulus-focused features (e.g., Gorin & Embretson, 2006). Most other difficulty frameworks are entirely concerned with cognitive operations and properties of item response behaviors, not item stimulus features.

However, as discussed above, from the standpoint of test development, there is a need for an item difficulty taxonomy that is focused on item stimulus features.

Furthermore, within the frameworks that focus on item stimulus features, there are those frameworks that focus on specific test items types (e.g., sentence completion items) and those with difficulty factors that can be applied across different item types. Most frameworks are focused on a particular item type in their discussion of item difficulty. For example, in analyzing sentence completion items, Sheehan and Mislevy (2001) found that adding a second blank to an item can lessen item difficulty. Thus, in analyzing how item stimulus features impact item difficulty, most research has focused on specific item types. At the same time, there remains a gap in understanding item difficulty from general item stimulus features that could be applied across different item types.

Thus, there is a need for a general framework of item difficulty stimulus features that can be applied across different tests and item types. In an effort to reduce this gap, the current paper will propose and test a general item difficulty framework that will focus on stimulus features of items. The proposed framework will utilize item generation theory (e.g., Irvine et al., 1990), which allows items to be decomposed into the factors that are hypothesized to impact difficulty as well as examine the impact of various item stimulus features on difficulty. Item generation theory (e.g., Irvine et al., 1990) was developed for mass testing and allows test developers to produce comparable forms of tests without extensive item pre-testing (Lievens & Sackett, 2007). Item generation theory has not been utilized in this capacity in prior research. The proposed framework will also incorporate features of existing item difficulty frameworks (e.g., Freedle & Kostin, 1997) and analytical approaches that are well suited to test the impact of these stimulus features on item difficulty. Specifically, a modern psychometric approach to modeling item

properties called linear latent trait modeling (Fischer, 1973) will be used. This analysis is based on item response theory and expresses item difficulty in terms of underlying stimulus features that create item difficulty. Item response theory (IRT) predicts the probability that a respondent will select a particular response option based upon the level of the construct possessed by the individual and the properties of the item. As will be reviewed in subsequent sections, IRT offers many benefits over classical psychometric techniques for examining research questions about item difficulty. In the next chapters the notion of item difficulty, IRT, linear latent trait modeling, and item generation theory will be addressed. A detailed discussion of item difficulty and theories/frameworks on item difficulty will follow.

The purpose of the current paper is to examine general stimulus features of item difficulty. The goal involves gaining a deeper understanding about the factors that impact item difficulty. The hope is that this knowledge will improve the quality and fairness of standardized tests as well as enable more efficient generation of parallel test items. The current paper makes no recommendations about using these hypothesized factors to manipulate levels of difficulty on a test to achieve a desired difficulty level. However, it is a potential caveat of learning and gaining a deeper understanding about item difficulty features. Such knowledge could be used for nefarious purposes. Item difficulty could potentially be manipulated to create a disparate testing experience for different groups. However, despite this potential caveat, the current paper aims at providing test developers with a better way of generating parallel tests and improving the quality of items on standardized tests. In no way does the current paper provide recommendations about how to manipulate test difficulty to attain some purpose or a certain difficulty level.

Chapter 2: Item Difficulty

Item difficulty has been traditionally defined in terms of examinee performance on an item at the group level (Scheuneman & Gerritz, 1990). In other words, item difficulty indicates how difficult it is for respondents to endorse a particular item (namely, the proportion of respondents answering the item correctly). Thus, when most test-takers endorse an item it is considered easy whereas when only few endorse it, the item is considered difficult. Overall, item difficulty is defined from a measurement perspective. Currently, there are no definitions or perspectives on item difficulty that are non-measurement focused. From this measurement perspective, there are two widely used approaches to defining item difficulty, the classical and modern measurement theory (e.g., item response theory). Both approaches are an attempt to assess unobservable psychological constructs and item stimulus features. However, they differ dramatically in the way that they accomplish this task.

Classical Measurement Theory

Classical measurement theory (commonly referred to as classical test theory, CTT) involves the estimation of the unobservable "true score" as a linear combination (i.e., sum or average) of responses to test items (Embretson & Reise, 2000). The true score is defined as an examinee's expected score on a test over repeated administrations of parallel tests. However, given that the true score can never be known, it is accepted that the observed test score is the product of a true score and random measurement error:

$$O = T + E \quad (1)$$

The observed score is impacted by two item stimulus features in CTT, the true score and the error. CTT assumes that each person has a true score, *T*, that could be obtained if the test were free from measurement error. A person's true score is the number of items answered

correctly over an infinite number of parallel test administrations. However, a true score cannot be observed because of inherent errors in the examinee, (e.g., mood), in response behaviors (e.g., bubbled in the wrong box), in the test (e.g., printing error), and in the administration process (e.g., hot or cold room). Therefore, it is assumed that observed score, O, plus measurement error, E, is the true score. It is also assumed that error is random, that over repeated testing the mean of error is zero, that the true score and error are uncorrelated, and that error on one item is uncorrelated with the error on other items. Thus, one of the weaknesses of CTT is the inability to directly determine the latent true score. Instead of the true score, an observed score, which is flawed due to measurement error, is used.

In CTT, item difficulty is typically defined as the proportion of test-takers who correctly respond to the test item. That is, item difficulty is a group level phenomenon. Specifically, it is operationalized as a group-level item mean. In CTT, *p*-values, defined as proportion of persons responding to the item correctly, define difficulty. Difficulty indices range inversely from 0.0 to 1.0 with 0.0 reflecting very difficult items and 1.0 reflecting very easy items. However, items with difficulties at the extremes provide little useful information about the differences between examinees. Information about differences among examinees is maximal at $p = 0.50$.

As noted above, CTT is not without limitations. One of them is that validity, reliability, and difficulty indices obtained for test items are sample-dependent and may not be useful for different samples or populations. Thus, test and examinee characteristics are not separate in CTT. One cannot interpret the difficulty of the test items without knowing the ability level of examinees completing the items. Rather, difficulty is sample-dependent and may vary for different samples and populations. For example, a verbal ability test given to a sample of highly-educated examinees will have difficulty parameters that will be different when the same ability

test is given to sample of less-educated examinees. Thus, test and examinee characteristics are confounded in CTT

Thus, because of sample-dependency there is no way to estimate item difficulty independently. Therefore, if one is interested in developing or testing an item difficulty framework, one would not be able to do so with the sample-dependent nature of CTT. Additionally, defining test or item difficulty in terms of a sample of respondents does not provide information about the factors that influence test difficulty regardless of who completes the items. It only indicates what factors make items more or less difficult for that particular sample.

The current paper will propose and test a general item difficulty framework that will focus on stimulus features of items. However, given that CTT difficulty indices are defined by sample and item characteristics, CTT techniques will not yield difficulty factors that could be generalized to other tests or situations. To understand item difficulty independent of tests and respondents one needs to turn to the modern psychometric approach of item response theory, which is discussed in the next section.

Item Response Theory

In Item Response theory (IRT), item difficulty is defined in terms of the level of the latent characteristic needed to successfully respond to the test item. IRT is a model-based approach to understanding non-linear relationships between latent characteristics, item characteristics, and response characteristics (Embretson & Reise, 2000). The purpose of IRT is to provide a framework for evaluating how well individual items on assessments or tests work. Most commonly, IRT is utilized by psychometricians for developing and refining item banks for tests and equating different versions of these tests (Hambleton, Swaminathan, & Rogers, 1991).

IRT models are referred to as latent trait models. In IRT, item responses are viewed as observable manifestations of hypothesized constructs. These attributes or constructs, such as chemistry knowledge, cannot be directly observed. However, in IRT, these constructs are inferred from the responses.

Depending on how the data are scored, IRT predicts the probability that a respondent will select a particular response option based upon the level of the latent characteristic assessed by a measure that is possessed by the individual and the properties of the item. Like CTT, IRT focuses on the level of the characteristic possessed by the individual and properties of test items (i.e., item discrimination and item difficulty). Unlike CTT, IRT estimates the relationship of latent person and item stimulus features with the same model and scale. In other words, IRT models the response of an examinee with a certain level of the latent characteristic to each item on the test.

IRT rests on two principles. The first principle is that the performance of an examinee on a test item can be predicted from their level of the latent characteristic assessed by the measure. The second principle is that the relationship between examinees, item performance, and the set of latent characteristics underlying item performance can be described by an item characteristic curve (ICC). An example of a generic ICC can be found in Figure 1. The ICC relates parameters of the person and items to the probability of a particular response. In other words the ICC describes how changes in latent trait relate to changes in the probability of a specific response (Embretson & Reise, 2000). Thus, the ICC illustrates how well the item discriminates among the respondents with different levels of the latent trait or theta. In the figure, the x-axis represents theta and the y-axis represents the probability of the correct response.

The ICC illustrates how well the item discriminates among the respondents with different levels of the latent characteristic. The difficulty parameter (also called the location parameter) is located at the point of inflection on the ICC. At the point of inflection, the location parameter represents the level of theta necessary for a 0.50 probability of endorsing a specified response option. In Figure 1, the location parameter is approximately 1.0, which means that an individual with one standard deviation above them mean on the latent trait has a 0.50 probability of endorsing the correct response. Larger values of the item difficulty parameter indicate that a greater amount of the latent characteristic is necessary for a 0.50 probability of endorsing a particular response option. The values of the difficulty parameter can theoretically range from-∞ to +∞, but in practice usually range from -3 to +3 (Embretson & Reise, 2000). Items with positive values are difficult items with low-ability examinees having low probabilities of responding correctly to those items and items with negative values are easier items with most examinees having a moderate to high probability of answering the item correctly. Note that the direction of the IRT difficulty parameter is opposite of the CTT item difficulty parameter.

The discrimination parameter reflects how well an item recognizes differences between respondents of different levels of the latent characteristic. The larger the value for this parameter the better the item is at discriminating between individuals with different levels of the latent characteristic. The item discrimination parameter can theoretically range from -∞ to +∞ but in practice values are rarely higher than +2.0 (Embretson & Reise, 2000).

The ICC in Figure 1 is an example of what is called a Rasch model. This model is appropriate with scored items. In this specific ICC, the probability of answering the item correctly is a function of the latent characteristic (person parameter) and item difficulty (item parameter). This model does not contain a parameter for the item discrimination. The person

parameter is called theta (θ) in IRT which represents the examinee's estimate of the latent characteristic that is assessed by a measure. For example, a latent characteristic could be cognitive ability, verbal ability, or a personality trait. Often, the distribution of theta is a standardized distribution. Thus, an individual with $\theta = 0.0$ has an average level of the characteristic. Positive values indicate above average values and negative values indicate below average values on the latent characteristic. In Figure 1, as the level of theta increases, so does the likelihood of endorsing the item. Individuals with extremely low levels of theta ($\theta = -3.0$) have the lowest probability of responding to the item correctly while individuals with extremely high levels of theta ($\theta = 3.0$) have the greatest probability of responding to the item correctly.

The values of the difficulty parameter and theta are on the same scale of measurement. Thus, they are directly comparable, not dependent on one another, and a major advantage of IRT. This idea is completely unlike CTT where item stimulus features are sample-dependent. Test and examinee characteristics are confounded in CTT; therefore, one cannot compare the parameters or interpret item difficulty without knowing the ability level of examinees. In IRT, however, item parameters do not depend on examinee parameters and vice versa. One can directly interpret these values. This property is known as invariance and will be discussed below.

IRT represents a family of mathematical models that describe how individuals interact with test items (Embretson & Reise, 2000). There are a variety of IRT models that can be used to examine item response data. A primary distinction among the popular IRT models is the number of parameters used to describe items (Hambleton, Swaminathan, & Rogers, 1991). The three most popular IRT models are the one-, two-, and three-parameter logistic models which are all appropriate for binary items. The focus of the current research will be the linear logistic latent

trait model (Fischer, 1973) which is an extension of the Rasch model. The Rasch model is discussed in the next few paragraphs.

The Rasch model is one of the dominant models for binary items (e.g., success or failure on a test item) in most fields of research (Hambleton et al., 1991). The Rasch model is similar to other IRT models including the one-parameter logistic model. The one-parameter logistic model (1PL) is mathematically equivalent to Rasch's model (Hambleton et al., 1991). Both models do not model discrimination parameters; rather, the discrimination parameter is considered a constant. As discussed above, the discrimination parameter reflects how well an item recognizes differences between respondents of different theta levels. Both, the Rasch and 1PL models require that items have a constant value for the discrimination parameter. For the Rasch model this constant is 1 but for the 1 PL the constant does not have to equal 1 (de Ayala, 2009). However, mathematically both models are equivalent. The values from one model can be transferred into the other model with appropriate rescaling.

The Rasch model is represented mathematically as,

$$P(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad (1)$$

where $P(\theta)$ is the probability of selecting the response option scored as '1' for a particular person to a particular item, *b* is the location parameter, and *e* is an exponential function. The Rasch model defines probability where an examinee with a certain level of theta will solve item, *i*, with a certain difficulty, *b*. Thus, in the Rasch model the probability of endorsing an option is a function of the latent characteristic and item location. For example, in Figure 1 as the level of theta increases, so does the likelihood of endorsing the item. There are no other parameters in the Rasch model.

There are a number of assumptions that are required of the IRT models used in this study. These assumptions are invariance, unidimensionality, local independence, and model fit. Invariance of item and ability parameters is the cornerstone of IRT and a major distinction from CTT. Invariance means that parameters that characterize an item do not depend on the ability distribution of examinees and parameters that characterizes an examinee do not depend on the test items (Hambleton et al., 1991). Due to invariance, item parameters are independent of the sample used to calibrate them and theta estimates are independent of the test items. As discussed above, this is radically different from CTT indices that are defined by characteristics of items and the sample.

The property of invariance is crucial in utilizing IRT for understanding item difficulty. The purpose of the current research is to gain an insight into item difficulty and develop a difficulty framework that can be applied across different tests and item types. Given that CTT difficulty indices are defined by item and sample characteristics, utilizing CTT techniques will not yield difficulty factors that could be generalized to other tests, respondents, or situations. Thus, the property of invariance makes IRT essential for understanding item difficulty using the approach taken in this study.

Unidimensionality is another important assumption of many IRT models. Most commonly employed IRT models assume that a single latent-trait dimension underlies the item response (Embretson & Reise, 2000). Given that several factors always affect test performance, this assumption cannot be strictly met in many cases. However, for the unidimensionality assumptions to be met adequately there must be a presence of a dominant component that influences performance (i.e., sufficiently unidimensional; Drasgow & Lissak, 1983). If a

dominant component is present the assumption of unidimensionality is sufficiently satisfied (Reckase, 1983).

Local independence is another assumption of IRT. Local independence is obtained when, controlling for the latent trait, the probability of solving any item is independent of the outcome of any other item (Embretson & Reise, 2000). That is, after partialing out the latent trait, the items are uncorrelated. Local dependence occurs when examinee item responses depend on responses to other test items, not just their trait level (i.e., multidimensionality that is not included in the IRT model). Local independence also helps provide evidence for unidimensionality in some instances. When the data are sufficiently unidimensional, local independence is typically met (Embretson & Reise, 2000). However, testing factors may influence the occurrence of local dependence. For example, when multiple questions are embedded within a context of a reading passage, responses on later questions may be influenced by responses on preceding questions. Thus, certain testing situations may lead to local dependence. In such situations it is recommended that items displaying local dependence are combined into testlets or a set of items that are treated as a single test question (Yen, 1993). Alternatively, multidimensional IRT models can be applied.

The last assumption of IRT models is that the model fits the data. The advantage of IRT can only be obtained when there is a fit between the model and the type of test data, as a poorly fitting IRT model will not yield invariant parameters (Hambleton et al., 1991). In assessing the goodness-of-fit of a particular IRT model parameters are estimated to judge how well the model represents data at the item or person level (Embretson & Reise, 2000). Model fit can be evaluated by judging item fit. Any given test may consist of different items; therefore, there is no need for the same model to be applied to all items in a test (Embretson & Reise, 2000). Thus,

separate IRT models may be estimated for different test items. Model fit can also be assessed by person fit. Person-fit indices assess fit at the level of the individual examinee and the meaningfulness of a test score derived from the IRT model (Embretson & Reise, 2000).

Although the parameters obtained in IRT are not sample dependent, the definition of difficulty in terms of ability level does not provide information about what factors actually impact the difficulty of test items. Thus, neither CTT nor IRT outlines the specific factors that impact item difficulty. They simply estimate overall difficulty parameters. However, it is necessary to take into account the stimulus features of the items, not only the ability level of individuals responding to those items to understand item difficulty.

Taking into account item stimulus features is especially important from an IRT perspective, as item and person parameters are on the same scale. Furthermore, an understanding of what makes items difficult is essential to building knowledge about how individuals approach and solve items (Berk et al., 2001). One of the challenges in understanding what impacts item difficulty involves decomposing the difficulty parameter (*b* parameter in equation 1) to understand what factors lead to the parameter values that are observed. Although this type of item decomposition is not possible with CTT, it is possible with IRT using what is called the linear logistic latent trait model (Fischer, 1973).

Linear Logistic Latent Trait IRT Model

Linear logistic latent trait model (LLTM; Fischer, 1973) is a unidimensional IRT model in which the item difficulty parameter can be decomposed into the factors (called elementary parameters in LLTM) that are hypothesized to impact item difficulty (Embretson, 1983). As it applies in this study, the LLTM can be used to examine the impact of stimulus features on the difficulty of test items. LLTM makes it possible to test specific hypotheses about the impact of

item stimulus features, administration and testing effects, or cognitive processes on item difficulty (Kubinger, 2009).

This model is an extension of the Rasch model where the difficulty parameter, β_i, is decomposed into the factors that jointly determine the level of item difficulty. In LLTM difficulty is a combination of some components, known as the elementary parameters. Elementary parameters are theory-based components that are hypothesized to impact item difficulty. Specifically, β_i is decomposed into a linear combination of elementary parameters. The elementary parameters can be compared to a test battery. In administering a test battery researchers are usually interested in the total test score, which is similar to the overall difficulty parameter, β_i. However, in addition to the total test score researchers may also be interested in performance on individual tests. These tests are similar to elementary parameters. Thus, although in administering a test battery the ultimate goal may be a composite score researchers are usually interested in understanding the components of the composite score. Similarly, in LLTM the elementary parameters are components of the overall difficulty parameter, β_i.

In the present study, the elementary parameters will be item stimulus features. Thus, LLTM is an extension of the Rasch model with a linear constraint that describes the difficulty of a test item in terms of the stimulus complexity factors (Fischer & Formann, 1982). The main difference between LLTM and the Rasch model is that item difficulty is expressed in terms of underlying factors of stimulus complexity rather than individual parameters (Embretson, 1983). Essentially the β_i parameter of the Rasch model is decomposed into underlying difficulty components, or elementary parameters (Fischer, 1995). LLTM models item difficulty as a combination of some underlying components. Therefore, by modeling β_i, a more complex interpretation emerges in terms of item difficulty (Mislevy, Levy, Kroopnick, & Rutstein, 2008).

LLTM was developed because in some tests, item difficulty can be a function of certain processes involved in the problem solution, features of the test item or the conditions of the testing situations that impact test difficulty (Fischer, 1995). In such situations, item difficulty is described as an additive function of underlying elementary parameters. The core assumption of LLTM is that the difference between item parameters is due to cognitive operations, testing conditions, or stimulus complexity factors involved in one, but not in another item (Fischer, 1995).

In performing LLTM analysis, the first condition of fit is that the data must fit a traditional Rasch model. The model fit is determined by applying Andersen's Likelihood-Ratio test (LRT; Mair & Hatzinger, 2007). After the model fit is determined, parameters are estimated for the LLTM. Hypotheses are supported when the elementary parameters are statistically significant.

Thus, with LLTM it is possible to incorporate item content into predictions of item success (Fischer, 1973). When appropriate content factors can be specified for each item, then parameters that reflect impact on item difficulty can be estimated directly. For example, there may be numerous factors impacting the difficulty of items. The multivariate nature of LLTM allows researchers to see the impact of multiple factors simultaneously. LLTM can be applied to estimate the weights of each factor in determining item difficulty. Thus, in LLTM the difficulty of an item is characterized by the decomposition of difficulty parameters to understand the influence of each underlying factor (Embretson & Reise, 2000).

Thus, the Rasch model defines the probability that a test-taker of a certain ability level will respond correctly to an item of a certain difficulty level. In LLTM, all difficulty parameters β_i (i=1,2,…k) of the Rasch model are postulated as a linear combinations of certain hypothesized

elementary parameters, which in this study will be item stimulus features. LLTM is represented mathematically as

$$b_i = \sum_j^g q_{ij} n_j + c \qquad (2)$$

where β_i is item difficulty, n_j denotes the elementary parameters of the model, q_{ij} are fixed weights of the elementary parameter, n_j, and c is the normalization constant (Fischer & Formann, 1982). The weights are usually 1, if the j^{th} cognitive operation, or in this case stimulus feature, is assumed to be relevant for solving item I_i, and 0 if it is not. Thus, the weighted elementary parameters, n_j, are hypothesized to influence overall item difficulty, β_i. One examines the statistical significance of the elementary parameters to make conclusions about which factors are actually impacting item difficulty. As discussed above, the elementary parameters are interpreted as the impact of item stimulus features, administration and testing effects, or cognitive processes on item difficulty (Kubinger, 2009). In the current study the elementary parameters will be item stimulus features. Just as scores on individual tests on a test battery are components of a composite score, elementary parameters are components of the overall difficulty parameter.

Although the benefits of using LLTM in decomposing the difficulty parameter and gaining a richer understanding of stimulus complexity is apparent, LLTM has received little attention (Kubinger, 2008; Embretson & Daniel, 2008). Furthermore, most existing research utilizing LLTM has focused on cognitive operations in testing, not item stimulus features. For example, Sonnleitner (2008) utilized LLTM for testing a difficulty model for a German reading comprehension test. Looking primarily at cognitive demands of item processing, Sonnleitner (2008) found that certain components, such as propositional complexity- items referring to information scattered over the whole text and consisting of more propositions- increases

difficulty of items. The author found that item difficulty also increases with more response options as well as a greater number of correct response options.

Embretson and Wetzel (1987) also utilized LLTM to compare competing cognitive theories of text comprehension in explaining item difficulty on a reading comprehension test. The authors utilized 46 paragraph comprehension items from the Armed Services Vocational Aptitude Battery and administered them to a large sample of military applicants. Paragraph comprehension items involved reading a paragraph and responding to a multiple-choice question per paragraph. Utilizing LLTM to decompose item difficulty, the authors found that item difficulty of paragraph comprehension items depends on text representation variables and decision processing variables.

In another study, Embretson and Daniel (2008) investigated a cognitive complexity model for mathematical problem solving. The model consisted of problem representation (problem translation and integration), problem execution (solution planning and execution) and a decision stage which addresses the role of distracters in multiple-choice items. The proposed model was supported from the results of both the regression approach and the LLTM approach. However, LLTM estimates were more consistent. The authors argue that LLTM is rarely utilized for understanding the sources of item complexity. Instead, previously calibrated item parameters are modeled using regression techniques because raw item response data are often not available. However, Embretson and Daniel urge other researchers to use LLTM as LLTM estimates are more consistent.

One of the studies that utilized LLTM to research stimulus features of items is a study by Hohensinn and Kubinger (2008) who utilized LLTM to investigate the impact of different response formats on item difficulty in a mathematical competence test. The authors utilized a

conventional multiple-choice format with a single correct response and five distracters, as well as a multiple-choice alternative with two correct answer options and three distracters. In the third response format "grid" items were used in which the correct solution had to be found and filled into a provided grid on the answer sheet. The authors found that although there were some differences among items, overall the response option with two correct answers was more difficult than a conventional multiple-choice format with a single correct option. These findings are similar to that of Sonnleitner (2008), who found that difficulty increases with numerous correct response options. The authors found that constructed response or "grid" format was also more difficult than the conventional multiple-choice format.

LLTM has also been used by Gittler and Gluck (1998) to investigate the transfer of learning. Specifically, these authors investigated the effects of descriptive geometry instruction on spatial ability test performance. In investigating differences in item difficulty parameters before and after geometry instruction, Gittler and Gluck (1998) applied LLTM for the measurement of change. The authors found that geometry instruction stimulates the development of spatial ability. In other words, geometry instruction improves spatial ability. Additionally, sex differences, where males outperformed females, that were present during the first testing disappeared after the geometry instruction.

The benefits and the primary purpose of utilizing LLTM is to generate test items with specified item difficulty (Kubinger, 2009). However, as discussed above, we do not have a clear knowledge of the factors that influence item difficulty (e.g., Berk, Lohman, & Cassata, 2001; Carpenter et al., 1990; Embretson, 1995). As noted above, most research on item difficulty has focused on cognitive processes. The current paper is interested in developing and testing a general framework for the stimulus features of test items that impact difficulty and that can be

applied across different tests and item types. To the best of our knowledge, LLTM has not been used in an integrated theoretical approach such as that proposed in this study.

In utilizing LLTM to develop a broad stimulus-oriented difficulty framework, one needs to specify theoretical sources of item difficulty. As discussed above, much of interest in item difficulty centers on cognitive operations and properties of item response behaviors (e.g., Carpenter et al., 1990). At the same time, there is little research on stimulus features of items. Furthermore, within the frameworks that discuss item features, most address a particular item type (e.g., sentence completion items) with little focus on difficulty factors that can be applied across different item types. In the next sections, theoretical sources of item difficulty will be reviewed.

Chapter 3: Theoretical Approaches to Understanding Item Difficulty

As discussed above, in identifying theoretical sources of item difficulty one can address cognitive process and item stimulus features. Traditionally in conceptualizing item difficulty greater attention has been paid to cognitive processes rather than stimulus features of items. Cognitive process theories focus on cognitive operations and the role of memory in performance of examinees. Rather than examining item stimulus features as sources of difficulty, cognitive theories look at cognitive processes that set high scorers apart from low scorers. Two examples of cognitive process sources of item difficulty are working memory and mapping and decision processes.

Cognitive Sources of Item Difficulty

Working memory is among the most widely researched cognitive sources of item difficulty. For example, working memory was a focus of Carpenter et al. (1990) who utilized the Raven Progressive Matrices Test to study analytic intelligence and identified which processes distinguish between higher and lower scoring test-takers. The authors found that difficult problems required test-takers to posses sophisticated problem-solving goal management. In other words, in answering the problems that elicited high error rates, test-takers had to be able to produce goals and sub-goals, as well as be able to track their progress to ensure that they are on the path to satisfying the higher level goals (Carpenter et al., 1990).

This type of research involves "mapping" the cognitive operations involved in connecting information, and decision processes. Cognitive maps are used to construct and accumulate spatial knowledge in order to reduce cognitive load and enhance recall and the learning of information. These are a commonly researched cognitive source of item difficulty, which focuses on information processing of examinees and connections they make during the

test. For example, in identifying the difficulty of GRE reading comprehension items, Gorin and Embretson (2006) found that difficulty of these items is explained primarily by decision processes necessary for mapping information between passages and response alternatives. According to the authors, processing response alternatives with difficult vocabulary requires more cognitive resources than processing low-vocabulary alternatives. Thus, respondents may not consider response alternatives with high-level vocabulary to save resources and time (Gorin & Embretson. 2006).

This research on cognitive processes as sources of item difficulty has increased our knowledge of cognitive underpinning of solving test items. However, sole focus on cognitive processes does not help with understanding how item stimulus features impact difficulty. Furthermore, as previously discussed, understanding what item stimulus features impact item difficulty is an important step in mapping the cognitive processes that are involved (Carpenter et al., 1990; Embretson, 1995). A better understanding about the factors that impact item difficulty could also lead to a more efficient test generation and improve the fairness quality of standardized tests. Therefore, the current paper will focus on item stimulus features in proposing and testing a general item difficulty framework.

Stimulus Features as Sources of Item Difficulty

In high-stakes testing programs, there is a need to develop alternate forms of tests with items of varying difficulty. The need for alternate tests is motivated by concerns that reusing the same test will lead to a decline in validity. These concerns are due to the possibility that individuals re-taking an exam after an initial failure are at unfair advantage by being already exposed to test items. Additionally, a possible security breach would also undermine test validity (Sackett, Burris, & Ryan, 1989). Thus, generation of multiple tests is necessary in mass testing

situations. However, in order to generate multiple tests without item-pretesting one needs to know which item stimulus features impact item difficulty (Lievens & Sackett, 2007).

Although cognitive difficulty factors provide insight about what differentiates high scorers from low scorers, the cognitive approach does not explain what aspects of an item make it more or less difficult. To understand actual item difficulty, one needs to address the stimulus features of test items. The stimulus features of items, as they impact difficulty, have been examined to a lesser degree then cognitive operations. Furthermore, research that has investigated item stimulus features focused mostly on specific item types (e.g., sentence completion items) while virtually no research has focused on item features that could be applied across a variety of item types. Filling this gap is particularly important for mass testing situations, as an understanding of difficulty factors could aid in generation of parallel tests consisting of multiple item types. In developing general item difficulty factors, one needs to find an organizational framework of item stimulus features. In the current study, the framework will be item generation theory (e.g., Irvine et al., 1990).

Chapter 4: Item Generation Theory Approach to Item Difficulty

Item generation theory (e.g., Irvine et al., 1990) represents a relatively new research area in which specific cognitive and psychometric theories are applied to test construction practices for the purpose of producing test items (Gierl & Leighton, 2004). Item generation theory was developed for mass testing and allows test developers to produce comparable forms of tests without extensive item pre-testing (Lievens & Sackett, 2007). A major task of researchers working on item generation theory was to provide parallel tests that do not require equating by item response theory methods (Irvine, 2002).

Item generation theory is particularly appropriate to the task outlined in this paper as it can be used to specify item stimulus features that contribute to item difficulty for a variety of different item types. One advantage of item generation theory is that it forces test developers to articulate the theoretical factors that may contribute to item difficulty (Lievens & Sackett, 2007). This determination includes radicals and incidentals. For example, if test developers wish to vary test difficulty in a construct-congruent way, without using test banks, then the elements of items that change their difficulty have a root-causal or *radical* function (Irvine, 2002). According to Irvine (2002), radicals are the basis of valid and reliable parallel test forms. Radicals are structural elements of an item which cause statistically significant changes in item difficulties, which may be measured by error rates and/or completion time (Irvine, 2002).

Researching difficulty of a German reading comprehension test, Sonnleitner (2008) highlighted a number of factors that we would identify as radical elements in item generation theory. For example, among other radicals, the author found that item difficulty increases with a greater number of response options. In other words, items that have more response options are

more difficult. Thus, test developers wishing who want to set a higher difficulty level for a test can utilize a higher number of response options.

Radicals are distinguished from *incidentals* in item generation theory. Incidentals are item surface characteristics that ensure item independence in serial tests, but they do not determine item difficulty (Irvine, 2002). In other words, in parallel testing there is a need to generate multiple items of similar difficulty that are not identical to one another. Incidentals allow researchers to generate multiple items which are different on the surface, but those variations do not determine the actual difficulty of those items. For example, Sheehan and Mislevy (2001) found that "clothing" changes (changes that alter surface characteristics) do not alter item parameters. In their research the authors utilized parent and variant sentences. Parent sentences are original sentences while variant sentences try to preserve parent item parameters while appearing different on the surface. For example, a parent sentence focused on "political commentators," while a variant focused on "presidents of major corporations." The authors found that item parameters of the variant are undistinguishable from that of the parent sentence. Thus, incidentals allow the generation of multiple items without altering item difficulty level.

For example, in Sonnleitner's (2008) test of difficulty models of a German reading comprehension exam, it was found that text coherence has no impact on item difficulty. A text is considered coherent when there are referent words connecting new information with information already presented. Sonnleitner proposed that items relating to highly coherent texts are easier to solve. However, the author found no relationship between coherence of the text and item difficulty. Therefore, in this situation, text coherence is an incidental which bears no impact on difficulty level. This incidental, therefore, can be varied in producing parallel versions of test items.

In another study, after evaluating different approaches to developing alternative situational judgment tasks in high-stakes setting when item pretesting is not feasible, Lievens and Sackett (2007) found that incidentals in situational judgment tasks are linguistic and grammar changes to item stem and options. These incidentals merely provided variation on alternative test forms and did not undermine psychometric properties of the exam.

Once the radicals and incidentals are determined, they are varied to produce item variants (Lievens & Sackett, 2007). It is important to note that radicals are not intuitively determined. Initial empirical support is necessary before a factor can be considered a radical and applied across item variants. In generating item variants some items could be isomorphs because they have the same radical but different incidentals (Lievens & Sackett, 2007). Thus, such items look different even though they are psychometrically equivalent. Other items are true variants as they differ in radicals and incidentals (Lievens & Sackett, 2007). Item generation theory forces test developers to think a priori about factors that contribute to difficulty. Therefore, item generation theory may serve as a meta-item writing approach where the focus is on item specification, rather than the item itself (Lievens & Sackett, 2007). The result is a set of rules that developers follow in generating alternate test forms.

A deeper understanding of the stimulus factors that impact item difficulty could be very useful in effectively generating test items and parallel tests through item generation theory. Such knowledge can translate into empirically sound radicals and incidentals that can be used to produce item variants. As discussed previously there is a lack of a general framework for the stimulus features of test items that impact difficulty and that can be applied across different tests and item types. In an effort to reduce this gap, the current paper will utilize item generation theory (e.g., Irvine et al., 1990), which allows items to be decomposed into the factors that are

hypothesized to impact difficulty and integrate this approach with LLTM. As previously mentioned, LLTM expresses item difficulty in terms of underlying factors of stimulus complexity and, therefore, is well suited to be utilized with item generation theory in that items can be coded on the radicals and incidentals that they possess and subjected to LTM analysis.

In testing the item generation framework and proposing incidentals one would be hypothesizing the null. That is, they do not impact item difficulty according to item generation theory. For that reason, incidentals will not be hypothesized as sources of item difficulty in this study. Incidentals have not been hypothesized in prior research investigating item difficulty within the item generation framework (Sonnleitner, 2008). However, they have been suggested by some researchers to impact other aspects of test performance (e.g., Freedle and Kostin, 1997). Rather than testing null relationships, incidentals will be those factors that do not contribute to item difficulty.

Radicals as a Source of Item Difficulty

Item generation theory allows item decomposition and requires researchers to think and specify what factors contribute to test difficulty (Irvine, 2002). Radical elements allow researchers to vary test difficulty without test banks or item pre-testing. At the same time, incidentals allow researchers to preserve serial independence of multiple tests without altering the difficulty level. The item generation approach is appealing in its simplicity and versatility, as item generation theory could be applied to any type of test.

As discussed above, to understand actual item difficulty, one needs to address the stimulus features of test items. There is virtually no research on general item stimulus features that could be applied across a wide range of tests. In developing a general item difficulty framework, the current paper draws on existing item difficulty stimulus features (e.g., Freedle &

Kostin, 1997) as well as research on cognitive processes which could also be viewed through the prism of stimulus features (e.g., Pellegrino & Glaser, 1980). There is potentially an infinite number of difficulty factors that could be examined and utilized in test generation. The current paper is interested in broad stimulus difficulty factors that have been utilized in prior research and which have been supported as having impact on item difficulty. Specifically, the current research will examine the impact of rule complexity, word rarity, language ambiguity, sentence length, the presence of a negative component, the number of response options, and item content as potential item stimulus difficulty factors or radicals. These factors have received prior attention and have been shown to impact item difficulty. There is a potentially infinite number of difficulty factors, general and test-specific, that could be evaluated in the future within the item generation framework. The present study selected factors with substantial prior research. Other potential factors are possible and can be pursued in future research endeavors.

In the next paragraphs existing research on item stimulus difficulty factors will be discussed and the hypotheses of the current research will be proposed. The stimulus features of item difficulty discussed below are also listed in Table 1. The actual coding of the factors will be discussed below in the Method section.

Rule complexity. Pellegrino and Glaser (1980) found that item difficulty on an inductive reasoning task is influenced by rule complexity, which depends on working memory capacity and representational variability. Although Pellegrino and Glaser's work focuses on cognitive processing rather than item demands, the inclusion of rule complexity expands the focus to item stimulus features. Arguably, rule complexity could be viewed from a cognitive point of view, in that the examinee must infer and process various rule elements. However, rule complexity could also be viewed as an item stimulus feature in that rule complexity or the number of rules could

impact the difficulty of a particular item. If rule complexity is viewed as an item stimulus feature impacting difficulty, rule complexity could have a radical function. Thus, although Pellegrino and Glazer's theory on of individual differences emphasizes working memory (Embretson, 1995), it includes a stimulus feature in addition to cognitive processes in addressing item difficulty.

According to Embretson (1995) differences in maintaining an absolute level or resources and directing them is important in tasks that cannot be performed automatically. Therefore, items with more rules may be more difficult than items with fewer rules because they require greater control and more resources (Embretson, 1995). When faced with difficult items, ones consisting of difficult vocabulary, for example, examinees may choose not to respond rather than exert their resources (Gorin & Embretson, 2006). Thus, similarly, when faced with multiple or complex rules, examinees may choose to avoid the item rather than exert effort and resources on its solution.

Given that rule complexity can impact item difficulty and this difficulty factor could be applied across a wide range of items, rule complexity as a stimulus feature will be examined in the current paper. Therefore, it is proposed that rule complexity will be a statistically significant elementary parameter contributing to item difficulty.

H1: Rule complexity is a statistically significant elementary parameter contributing to item difficulty.

Word Rarity. In the current paper, the issue of difficult vocabulary will be addressed as word rarity. Word rarity is often used as a marker for difficult vocabulary, and word frequency indices (e.g., Kucera & Francis, 1967) have been used to investigate word rarity. In approaching

and solving test items, understanding of the question is very important, which can be complicated by the use of difficult language.

Research highlighting word rarity was conducted by Roccas and Moshinsky (2003). Examining the correlation between the rarity of words used in verbal analogies and difficulty of items, the authors found that item difficulty is positively correlated with word rarity. The research of Roccas and Moshinsky (2003) confirmed earlier work of Bejar, Chaffin, and Embretson (1991) who examined the relationship between the difficulty of analogies and the familiarity of words used in an item stem. The authors found that word familiarity was negatively related to item difficulty. Thus, items with difficulty vocabulary may be more difficult than items with easier vocabulary.

Another study highlighting word rarity was conducted by Embretson and Wetzel (1987). The authors found that the vocabulary level of the response alternatives affected the likelihood that an examinee would consider the alternative as a potential correct response. Distracters with rare or difficult vocabulary were less likely to be processed by examinees and were processed faster, as measured by response time. Examinees were also less likely to choose a correct response alternative with high vocabulary level. Thus, the presence of a correct response option with high vocabulary negatively affected the likelihood of an examinee choosing that response. Essentially, a correct response option with rare words decreased the likelihood of an examinee responding to the item correctly, thereby increasing item difficulty.

These results were confirmed by Gorin and Embretson (2006). In identifying the difficulty of GRE reading comprehension items, Gorin and Embretson (2006) found that difficulty of these items is explained primarily by decision processes necessary for mapping information between passages and response alternatives. According to the authors, processing

response alternatives with difficult vocabulary requires more cognitive resources than processing low-vocabulary alternatives. Thus, items with low-vocabulary correct responses were easier than items with high-vocabulary correct response. Alternatively, one may think of other additional factors in addition to cognitive load discussed by Gorin and Embretson (2006), including anxiety, which may explain processing of difficult vocabulary.

Thus, processing items with rare words may require greater cognitive resources. However, although the authors concentrate on cognitive resources in processing response alternatives, one may argue that word rarity is also a stimulus feature that could potentially be utilized in parallel generation of tests. Therefore, one could propose that word rarity has a radical function in impacting item difficulty. Given, that the use of rare words in a test item can impact its difficulty and given that this difficulty element could be applied across a wide range of items beyond analogies, it warrants further investigation. Thus, it is proposed that word rarity is a statistically significant elementary parameter contributing to item difficulty.

H2: Word rarity is a statistically significant elementary parameter contributing to item difficulty.

Language ambiguity. Related to language difficulty is language ambiguity. Many words can be interpreted in different ways, depending on a test-taker's knowledge, prior experience, and the text context. When an ambiguous word, one that can have multiple meanings, is presented, it takes a longer time for a test-taker to process and select the optimal meaning (Just & Carpenter, 1980). In examining item difficulty of a reading comprehension test, Sonnleitner (2008) found that ambiguity in test questions increases item difficulty.

Ambiguity in an item can also impact item properties and the quality of one's data. In examining the impact of item ambiguity Coombs and Coombs (1976) found that ambiguity in

wording or item content increases the number "no opinion" responses. Additionally, when facing ambiguity in test items, examinees may leave the item unanswered or select either a positive or negative response option (Velez & Ashworth, 2007). Furthermore, in responding to Likert-type items with a midpoint response option, item ambiguity may lead individuals to select the midpoint, not because of their neutral opinion, but due to confusion (Velez & Ashworth, 2007). Thus, item ambiguity can adversely impact item reliability as well as validity of one's finding. Additionally, according to Coombs and Coombs (1976), ambiguous items pose more cognitive difficulty and, therefore, many examinees may choose to notrespond to such items.

Thus, item ambiguity can increase item difficulty. Ambiguity can be a factor in many different items beyond reading comprehension, including sentence completion items and analogies. Given that this difficulty factor could be applied across a wide range or items it is proposed that ambiguous language, operationalized as the number of alternative meanings for a word, is a statistically significant elementary parameter contributing to item difficulty.

H3: Ambiguous language is a significant elementary parameter contributing to item difficulty.

Negative wording. The presence of negative wording can increase item difficulty and, for this reason, test developers often refuse to use negative test items (Gorin, 2005). Negative wording increases item difficulty by having an impact on processing and comprehension, as negative statements have been shown to be more difficult to comprehend than affirmative statements (Carpenter & Just, 1975; Gorin, 2005; Fodor, Fodor, & Garrett, 1975). Processing negatively worded statements could lead to an interference of new information integration into an existing knowledge structure (Gorin, 2005). This interference, in turn, can lead to incorrect comprehension of the question. Indeed, in investigating item generation of verbal comprehension

items, Gorin (2005) found that negative wording significantly increases item difficulty. In a separate study examining the correlation between the rarity of words used in verbal analogies and difficulty of the item, Roccas and Moshinsky (2003) also found that negative wording increases item difficulty.

Thus, past research has shown that it is more difficult to process texts that include negative wording (e.g., Gorin, 2005) as negative wording adds complexity to the processing of the text. However, the impact of negative wording has been viewed through the cognitive processing theory, as a task of processing negative statements could be considered an additional cognitive operation. According to Gorin and Embretson (2006), additional cognitive operations require more processing and effort, thereby making test items more difficulty. At the same time, negative wording could be viewed from an item stimulus perspective in that the presence of negative wording increases item difficulty.

Thus, prior research suggests that negative wording impacts item difficulty. If negative wording is a radical that impacts difficulty, this knowledge can lead to a more careful generation of parallel items. Test developers could avoid negatively worded variants if original items are not negatively worded, thereby not increasing item difficulty for parallel items. Therefore, it is proposed that the presence of negative wording is a statistically significant elementary parameter contributing to item difficulty.

H4: The presence of negative wording is a statistically significant elementary parameter contributing to item difficulty.

Item length. In an effort to develop item difficulty sources that could be used for mass test generation Berk et al. (2001) conducted a study examining item difficulty of sentence completion items. Utilizing both self-reports and IRT analyses, the research of Berk and

colleagues (2001) sought to understand difficulty factors which would allow the generation of parallel tests for different ability levels of test-takers. The results of examinee self-reported information and IRT analysis (3PL model) indicated that the number of words in a sentence is not related to item difficulty ratings. Thus, the length of the sentence did not contribute to item difficulty. Nonetheless, the authors are cautious about making generalizations from these findings. The examinees performed well on the test with 25 out of 36 students responding correctly to 14 of the 20 items. Therefore, had the test items represented a larger span of difficulty, a different conclusion regarding the nature of a sentence length would be reached. The authors also point out that the sentence length could have a positive relationship with informational cues about the correct answer (Berk et al., 2001).

In an earlier study, Green (1984) investigated the effects of item characteristics on item difficulty on a multiple-choice exam. One of the item characteristics was language difficulty, which was varied by increasing sentence length, syntactic complexity, and substituting uncommon terms for familiar words. Green did not find a significant effect of language on difficulty across all test items. Examination of individual items revealed mixed results. On some items increasing language difficulty resulted in additional information making items easier while on others increasing language difficulty led to increased item difficulty. Thus, this is another study offering inconclusive results about the impact of sentence length on item difficulty. It should also be noticed that Green (1984) varied three factors, sentence length, syntactic complexity, and vocabulary in examining impact of language difficulty on item difficulty, not just sentence length. It may be that these factors operate differently and, if one was to only vary sentence length without increasing vocabulary load and altering the syntax, different results would be obtained. It should be noted that difficult vocabulary or word rarity has been

investigated in prior research and has been shown to impact item difficulty (e.g., Gorin & Embretson, 2006). It is, therefore, important to gain a better understanding of the impact sentence length has on item difficulty without contamination of other difficulty factors.

If it is indeed found that number of words in a sentence has no impact on item difficulty, this information could be efficiently used in generating item variants without altering item difficulty. However, if the number of words in an item as a radical, it would signal that test developers should take caution in varying the length of test item variants. It is proposed that item length is a statistically significant elementary parameter contributing to item difficulty.

H5: Item length is a statistically significant elementary parameter contributing to item difficulty.

Response options. According to Embretson and Wetzel (1987) and Gorin and Embretson (2006) every single response option is processed and either confirmed or rejected. The likelihood of a correct response option confirmation is dependent on a number of factors, including word rarity, which is addressed above. However, given that every single response option is processed and requires cognitive effort on the part of the examinee (Gorin & Embretson, 2006), the number of response options could be a difficulty factor. Therefore, items with a greater number of response options should be more difficult than items with fewer response options.

In examining item difficulty of a reading comprehension test, Sonnleitner (2008), found that difficulty increases according to the number of response options. The author found that difficulty further increases when there are numerous correct response options. The presence of more than one correct response options is not applicable to many tests and therefore will not be examined in the current study. However, the number of response options as a radical can be applied across a wide range of tests and items. If the number of correct response options impacts item difficulty,

this radical should be controlled in generating item variants. Therefore, this stimulus feature will be included in the present study. It is proposed that the number of response options is a statistically significant elementary parameter contributing to item difficulty.

H6a: The number of response options is a statistically significant elementary parameter contributing to item difficulty.

It is also proposed that constructed-response items are more difficult than selected-response or multiple-choice items. It has been shown that selected-response items are easier than same content constructed-response items (Downing, 2004). The absence of choices may require test-takers to exert greater resources. Instead of processing existing choices, an individual must construct an answer and go through a process of solving and possibly falsifying his or her answer. This process may require greater cognitive resources than the process of selecting existing responses. Therefore, it is proposed that constructing a response is a statistically significant elementary parameter contributing to item difficulty.

H6b: Response construction is a statistically significant elementary parameter contributing to item difficulty.

Item content. According to Sonnleitner (2008), the text of an item can impact difficulty. Each text requires construction of a model and therefore depends on content-specific knowledge and motivation to work with a particular topic. Therefore, the content of the text can make items more or less difficult. Furthermore, several researchers have suggested that surface characteristics can, in fact, impact performance.

The major concern with this potential source of item difficulty is that it differentially impacts individuals from different racial/ethnic groups and gender. Numerous researchers have reported that scores on cognitive tests are typically 0.7 to 1.0 standard deviations higher for

White test-takers compared to Black test-takers (e.g., Roth, Bevier, Bobko, Switzer & Tyler, 2001). Hough, Oswald, and Ployhart (2001) have reported this disparity for a variety of cognitive ability tests, including verbal, quantitative, memory For example, researchers have shown that seemingly construct-irrelevant aspects, including math and science content, reliance on prior knowledge, and presence of a cultural context, which may not be easily recognizable to all test-takers, impact test performance (e.g., Freedle & Kostin, 1997; Fagan & Holland, 2007) and spatial.

Freedle and Kostin (1997) offer a theory on seemingly construct-irrelevant features of test questions, such as content, which can actually impact item difficulty. According to the authors there are significant differences in how Black and White examinees, who are matched on verbal scores, respond to verbal test items. Utilizing differential item functioning (DIF), Freedle and Kostin (1997) investigated ethnic differences in analogy item responses on SAT and GRE questions. Among other factors Freedle and Kostin (1997) found that science content accounts for differences between Black and White test-takers. Although the content on these tests should not impact difficulty and should function as an incidental, science content was found to impact difficulty and function as radical. Essentially, the presence of this factor in test items increases difficulty for Black test-takers.

According to Scheuneman and Gerritz (1990) Black examinees perform worse than White examinees on science-related SAT analogy items. Freedle and Kostin (1997) suggest that worse performance among Black examinees may be because they do not consider science relative to their experiences. Based on prior research Freedle and Kostin (1997) proposed and found that Black examinees perform more poorly on analogies having science content than White examinees.

The present study is not investigating racial predictions; rather, it would be interesting to see if, regardless of race, item content impacts difficulty. The Wonderlic test items that will be examined in the present study do not include science-related items. However, there is a wide variety of items that one can code and investigate potential impact on item difficulty. The test provides a broad range of problem types, including analogies, definitions, disarranged sentences, mathematical calculations, analyses of geometric figures and word problems. At a basic level despite the variety of item-types, there are math items, verbal items, and items requiring colloquial or cultural knowledge. Most of the different item types can be categorized into these broad categories. As discussed above, each item test requires a construction of a model and depends on content-specific knowledge and individual motivation to work with a particular topic (Sonnleitner, 2008). Therefore, certain item contents can be more difficult. Some text contents may be more accessible than others and test-takers may have greater motivation to invest effort into answering these questions. The content of other items, however, may require greater cognitive resources and processing and, as a result, be more difficulty. Thus, item content may have an impact on item difficulty.

According to Freedle and Kostin (1997) factors such as presence or absence of math or science content, cultural content, and presence of social/personality content influences how Black and White examinees respond to individual cognitive ability test items. Essentially, according to Freedle and Kostin (1997), the presence of these factors makes the item difficult for Black test-takers. Although group differences in test taking is not the primary focus of the current paper, it would be interesting to see if some aspects of Freedle and Kostin's theory (1997), such as presence of a math or colloquial (cultural) content increases item difficulty.

The present study is not outlining specific predictions concerning racial differences. Rather it is proposed that item content is a statistically significant elementary parameter contributing to item difficulty. If it is found that these contents are radicals, it carries implication for test-item generation. For example, if colloquial content is found to be a radical, it would signal that parallel item generation should avoid colloquial or cultural knowledge items, as it could alter item difficulty level. In following Freedle and Kostin's conclusions (1997), it proposed that math and colloquial content are statistically significant elementary parameters contributing to item difficulty.

H7a: Math content is a statistically significant elementary parameter contributing to item difficulty.

H8b: Colloquial content is a statistically significant elementary parameter contributing to item difficulty.

Summary

There remains a gap in understanding general stimulus features of item difficulty that could be applied across different item types. In an effort to reduce this gap, the current paper proposes to test general difficulty factors that are focused on stimulus features of items. The proposed difficulty framework will utilize item generation theory (e.g., Irvine et al., 1990), which allows items to be decomposed into the factors that are hypothesized to impact difficulty as well as examine the impact of different item features on difficulty. The proposed difficulty factors will be analyzed together within the proposed set. To test the proposed framework, the study will utilize data from a study by Ferreter et al. (2008) and LLTM (Fischer, 1973), an IRT-based analytical approach that expresses item difficulty in terms of actors of stimulus complexity rather than individual parameters (Embretson, 1983).

Chapter 5: Method

Dataset

The data used in this study were drawn from Ferreter et al. (2008). That study collected data from 851 undergraduate students at a large Northeastern university who participated in the study for credit in their psychology course. The total sample consisted of 48% males and 52% females. The majority of participants (84%) were between the ages of 18 and 22. There was varied ethnic/racial representation in the sample. The largest self-reported group were Asian (38%) followed by White/Caucasian (21%), Hispanic (20%), and African-American/Black (11%). The remaining 10% selected other racial/ethnic categories (e.g., Aleutian/Pacific Island, White/Non-Caucasian, and Native American).

There is little consensus about the requisite sample size for IRT analyses (Embretson & Reise, 2000). Rather, the recommendations depend on the type of items and the number of parameters being estimated. The Rasch model estimates fewest parameters and, therefore, smaller sample sizes are adequate for stable parameter estimates. A sample of 100 may be sufficient for Rasch analyses if only item or person parameters are being estimated (Linacre, 1994). For non-Rasch models, it has been shown that a Graded Response Model, which is a generalization of the 2-PL model (Embretson & Reise, 2000), can be estimated with as few as 250 participants. However, around 500 participants are recommended for stable parameters (Reise & Yu, 1990). Thus, given that LLTM is a Rasch-based model a sample size of 851 is sufficient to obtain stable parameters.

Measures

Wonderlic Personnel Test. The Wonderlic Personnel Test (Wonderlic Associates, 1983) is a traditional test of cognitive ability. The test contains 50 items and has a combination of multiple-choice and constructed-response items. The test has a 12-minute time limit. The score is

calculated as the number of correct answers given in the allotted time. Based on normative information, a score of 20 is intended to indicate average intelligence. The test provides a broad range of problem types, including analogies, definitions, analysis of geometric figures, and disarranged sentences. The problems are arranged to become increasingly difficult. Estimates of internal consistency for the scores on the Wonderlic have been good. For example, Ferreter et al. (2008) reported an estimate of internal consistency of 0.79.

Coding of Item Difficulty Factors

With the exception of language ambiguity, all difficulty factors were coded by two independent and calibrated raters. Raters were trained on three trial tests with items similar to the Wonderlic. Item types on the trial tests included the same item-types as on the Wonderlic. Raters were trained until their agreement reached 80%. Test items were coded one factor at a time. Raters started with one factor and coded all test items before moving to the next factor.

For all the factors, interrater agreement reached acceptable levels (Landis & Koch, 1977). For any item where agreement was not reached, ratings of a third rater were used to resolve any disagreements.

Rule complexity. Rule complexity refers to the number of steps required to solve an item. For example, a rule on an item may be "Arrange the following words so that they make a complete sentence. Is it a true statement?" In this example, there are two rule elements. The first instructs the test-taker to arrange the words, while the second asks if arranged words create a true statement. In the current study, each item was coded according to the number of rule elements or the number of total steps needed to solve an item. For example, items with one rule element were coded 1 and items with two or more rule elements were coded 2. All item types were coded by two independent raters. Interrater reliability analysis using Cohen's Kappa was performed to

determine consistency among raters. The interrater agreement for rule complexity was 0.76. Kappa values over 0.70 are considered acceptable for interrater agreement (Landis & Koch,1977). A total of 29 items on the Wonderlic were coded as items with one rule and 21 item were coded as items with two or more rules.

Word rarity. Items with rare words should be more difficult than items with more common words. In previous research, a word frequency index by Kucera and Francis (1967) has been utilized to investigate word rarity. The word frequency index developed by Kucera and Francis remains the norm for American English. High- frequency words occur 40 times or more per million (range is 40-240 million, $M = 90$) and low-frequency words occur less than 10 times per million (range is 1-7 million, $M = 7$). The present research utilized the Kucera and Francis index to identify items which have rare words in them, those occurring at a low frequency of 1 to 7 times per million. All items on the Wonderlic, including item instructions, were analyzed by two independent coders. Items with rare words were coded 1 and items with no rare words were coded 0. Interrater reliability analysis using Cohen's Kappa was performed to determine consistency among raters. The interrater agreement for word rarity was 0.72. The lower Kappa rating reflects rater differences in looking up singular versus plural word forms. Out of all the items, 24 were coded as items with rare words.

Language ambiguity. Given that many words have multiple meanings, this study will define word ambiguity according to the number of meanings item words have. Every noun, verb, adverb, and adjective on the Wonderlic was entered into online Mirriam-Webster dictionary (www.merriam-webster.com) to evaluate how many multiple meanings each of these words have. Following that, a median value for the ambiguity of the whole test was determined. The median ambiguity for all test items was 9 meanings. Items with median ambiguity of 9 and

above were coded as ambiguous (1) items with median ambiguity below 9 were coded as non-ambiguous. All items and item instructions on the Wonderlic were coded. To avoid errors in identifying the correct words in an item to code, coders were not used for language ambiguity. Rather every noun, verb, adverb and adjective in each item was entered into the online dictionary to evaluate the number of multiple meanings for each word. Out of all the items, 36 were coded as ambiguous.

Negative wording. The present study investigated this stimulus factor with a broad range of item types of the Wonderlic. All item types on the Wonderlic were coded. The presence of a negative component pertains to the item stem as well as to the response choices. Two independent coders evaluated each item for presence of negative wording. Items that have a negative component, which included the presence of such words as never, not, or don't, were coded 1 and items that do not have a negative component were coded 0. Interrater reliability analysis using Cohen's Kappa was performed to determine consistency among raters. The interrater agreement for rule complexity was 1.0. About half of the test items, 24 out of 50, had a negative component.

Item length. The current research investigated the impact of item length on item difficulty. Item length was conceptualized as the number of words in an item. Two operationalizations of item length were utilized. The first operationalization was the length of the item stem. The second operationalization was the length of the total item, including response options. Thus, items were coded according to the number of words in an item stem and the total number of words in an item, including response options. Median item lengths were calculated for the entire test. The median item stem length was 15 words. For the first indicator of length, items with stems of 15 or more words were coded as long items (1) and items with less than 15

words were coded 0. Out of 50 items, 26 were coded as long items. The median overall item length for the entire test was 18 words. For the second indicator of length, the length of the total item, items with 18 words or more were coded as long items (1) and items with less than 18 words were coded as 0. For overall item length, 25 out of 50 items were coded as long items.

Two independent coders evaluated item length. Interrater reliability analysis using Cohen's Kappa was performed to determine consistency among raters. The interrater agreement for both operationalizations of item length was 1.0.

The number of response options. The present research investigated the impact of the number of response options on item difficulty. The items on the Wonderlic are either constructed-response items or have between 2 and 6 close-ended response options. This difficulty factor was coded in two ways. The first operationalization differentiated items between multiple-choice and constructed-response items. It has been shown that selected-response items are easier than same content constructed-response items (Downing, 2004). The absence of choices may require test-takers to exert greater resources. Therefore, to investigate the impact of constructed-response items on difficulty, construct responses were coded 1 and multiple-choice items were coded 0. Out of 50 items, 19 were coded as constructed-response items. The second operationalization of this difficulty factor sought to understand the impact more choices have on item difficulty. Previous research has suggested that difficulty increases according to the number of response options (Sonnleitner, 2008). Given that multiple choice items had at most 6 response choices, items were split and those with 3 multiple choices or less were coded 0 while those item with 4, 5, or 6 choices were coded 1. Out of 50 items, 14 had 4 or more response options.

Two independent coders evaluated item response choices. Interrater reliability analysis using Cohen's Kappa was performed to determine consistency among raters. The interrater

agreement for both operationalizations was 1.0. All items on the Wonderlic were coded with this factor.

Item content. The current paper is interested in exploring the impact of item content on item difficulty. There are three broad content themes that appear in the Wonderlic. There are questions requiring some type of verbal knowledge (analogies, sentence construction items) and items requiring math knowledge (geometry, word problems). Another factor is common knowledge/U.S. colloquial expression items. The following is an example of an item involving U.S. colloquial expression. "Are the meaning of the following sentences similar, contradictory, neither similar nor contradictory? Elbow-grease is the best polish. The work proves the worker."

The item content factor was coded two ways. The first operationalization defined item content in terms of the three broad categories described above. Two independent coders identified each item as math, verbal, or colloquial/common knowledge items. Interrater reliability analysis using Cohen's Kappa was performed to determine consistency among raters. The interrater agreement for this operationalization was 1.0. Out of 50 items, 22 were coded as math items, 22 as verbal items and 6 as colloquial items.

The second operationalization of item content involved greater detail. Each math item was coded as geometry items (3 items), word problems (11 items), or items requiring calculations only (8 items). Items coded as geometry items had geometric figures and asked participants to manipulate those figures in some way to calculate an answer. Items coded as word problems had a problem that required reading, understanding an item and performing calculations. For example, the following item would be coded as a word problem. "A person's car traveled 18.5 miles in 30 minutes. How many miles per hour was it traveling?" Items requiring calculations only had no other requirements but some arithmetic computation. The

following item would be coded as calculations only item. “Which is the next number in the series 100 10 1 0 .1 .01.”

Verbal items were coded as items requiring verbal knowledge (13 items), sentence construction items (3 items), or analytical items (6 items). The following item will be coded as an item requiring verbal knowledge, “Abduct Abet------Do these words have similar meanings, contradictory meanings, neither similar nor contradictory. Verbal items coded as sentence construction items require the respondent to create a sentence from words. For example, the following item would be coded as a sentence construction item. “Arrange the following words so that they make a complete sentence: planet is Mars a.” Finally, analytical items required test-takers to come up with an answer based on two or more statements. The following is an example of an analytical item: “Assume the first 2 statements are true. Is the final one true, false, not certain? All explorers are risk-takers. Most explorers are introverted. Some risk-takers are introverted.”

Interrater reliability analysis using Cohen’s Kappa was performed to determine consistency among raters. The interrater agreement for this operationalization was 0.75. All items on the Wonderlic were coded according to their item content.

Overview of the difficulty factors. Based on the operationalizations described above, there are 11 broad difficulty factors: rule complexity, word rarity, language ambiguity, negative wording, length of item stem, length of the whole item , response option types-multiple choice versus construct response, number of response options, math item content, verbal item content and colloquial/common knowledge items. Thus, this difficulty model consists of 11 proposed factors. The second model attempts a more in-depth evaluation of item content. Specifically, the math and verbal operationalizations are further decomposed into more detailed disaggregated

operationalizations. The result is a 15-factor model: rule complexity, word rarity, language ambiguity, negative wording, length of item stem, length of the whole item , response option types-multiple choice versus construct response, number of response options, geometry content, word problem content, calculation content, analogy content, sentence construction content, analytical content, and colloquial content. Both models were estimated to determine if more information about item content could lead to a better understanding of its impact on item difficulty.

Chapter 6: Results

Descriptive analyses and classical item analyses. Descriptive statistics, including minimum and maximum values, means and standard deviations are presented in Table 2. As can be seen, descriptive statistics are not reported for items 46 and 50. All participants responded incorrectly to these items. As will be seen in subsequent analyses, these items were removed from model estimation. Following examination of the distribution it was found that the Wonderlic Personnel Test has a positively skewed platykurtic distribution. The positive skew suggests that the test is difficult and the platykurtic distribution means that the distribution has a smaller peak around the mean.

Classical item analyses were conducted prior to hypothesis testing. The internal consistency of the responses to the items on the Wonderlic Personnel Test was 0.80. Item difficulty was also estimated using classical item analysis, where item difficulty is defined as the proportion of test-takers who correctly respond to test items. Item difficulty in classical item analysis is conceptualized as a group-level item mean and *p*-values, defined as proportion of persons responding to the item correctly, define difficulty. Difficulty indices range from 0.0 to 1.0 with 0.0 reflecting very difficult items and 1.0 reflecting very easy items. However, items with difficulties at the extremes provide little useful information about the differences between examinees. Information about differences among examinees is maximal with $p = 0.50$. Classical item difficulty analysis revealed that the item means on the Wonderlic Personnel Test range from 0.0012 to 0.9612 and the average item mean on the test is .4205

Tests of dimensionality. The test of unidimensionality was conducted on the existing data by Ferreter et al (2008). Sufficient unidimensionality was concluded by the authors of the study.

Additionally, the Rasch model tests unidimensionality as part of the examination of model fit (Demitrov & Raykov, 2003), which is discussed below.

LLTM parameter estimation and model fit. In LLTM the difficulty parameters β_i (i=1,2,…k) of the Rasch model are postulated as linear combinations of certain hypothesized elementary parameters n_j (j=1,2…p), which in this study will be item stimulus features. LLTM is represented mathematically as

$$\beta_i = \sum_j^p q_{ij} n_j \qquad (2)$$

The data set was analyzed using the software *LPCM-Win* 1.0 (Fischer & Ponocny-Seliger, 1998). LPCM-WIN estimates the conditional maximum likelihood item and person parameters and supports the formulation and testing of hierarchical hypotheses about the item difficulty factors. In step 1, the assumptions of the Rasch model are tested. Once the fit of the Rasch model is established, step 2 is to test the linear restrictions of LLTM. All program defaults will be used in the analyses.

LLTM is an extension of the Rasch model. As with other Rasch-type models, LLTM assumes that all items are equivalent in terms of discrimination. Rasch-type models do not permit each item to have a different discrimination. Rather, it is assumed that discrimination is uniform for all items. However, many models based on the core Rasch model have incorporated an item discrimination parameter, including the LLTM. However, a limitation of the LPCM-Win program is that it does not allow one to model a discrimination parameter for any item. This is a limitation of the software and not the model, which does not assume a uniform discrimination for all items.

Given that LLTM is a Rasch model with linear restrictions, testing the fit of LLTM requires two steps: testing the fit of the Rasch model (which also tests unidimensionality) and

testing the linear restrictions in equation 2 (Demitrov & Raykov, 2003). In step 1, the assumptions of a Rasch model-fitting item pool were tested using Andersen's Likelihood-Ratio test (ALR; Andersen, 1973). The ALR tests the assumption that a test is actually measuring the same ability in different subpopulations. The ALR statistic is an asymptotic Chi-square test (x^2) with degrees of freedom equal to the number of parameters estimated in the subgroups minus the number of parameters in the total data set. The ALR examines whether the model fits data significantly better when the item parameters are estimated separately for subgroups of subjects than when they are estimated on the whole sample. For the parameter estimates in the current study, the sample was divided between the low and the high scorers. An insignificant ALR implies that a Rasch model fits the data. The fit of the Rasch model is a pre-requisite for conducting LLTM analysis. Application of the Rasch model may entail deletions of items that do not fit the model (Kubinger, 2008). Once the fit of the Rasch model is established, step 2 is to test the linear restrictions of LLTM.

In step 1, all 50 Wonderlic Personnel Items were entered to estimate the Rasch model. The results indicated that the Rasch model does not hold. The ALR ratio was significant, as can be seen in Table 3. An examination of item standard errors showed that although all items had very high standard errors at 82, standard errors of items 46 and 50 were 1983.5 suggesting something peculiar about these items. With such large standard errors estimates of item parameters cannot be obtained (Fischer & Ponocny-Seliger, 2003.)

An examination of the dataset revealed that these two items received no correct responses. In other words all test respondents provided incorrect answers for items 46 and 50. Therefore, due to high standard errors and the fact that not a single test respondent answered correctly, items 46 and 50 were removed and a second Rasch model was estimated. The results

of the second Rasch model indicate that the model holds. As can be seen in Table 4, the ALR is low and insignificant at 55.4246.

The fit of the Rasch model is a pre-requisite for conducting LLTM analysis. With the fit of the Rasch model established, the second step is to test the linear restrictions of LLTM with the ALR (Kubinger, 2008). Given that the Rasch model holds it acts as a saturated model, where possible parameters are estimated, and a goodness-of-fit test is applied by using the ALR test. The data's likelihood in LLTM is contrasted against data's likelihood in the Rasch model using the formula:

$$-2Ln\left\{\frac{L_{RM}}{L_{LTM}}\right\}\sim x^2 \qquad (3)$$

where LRM is the likelihood of the data estimated by the Rasch model and LLTM is the likelihood of the data estimated by the LLTM, with *df=k-p* (Kubinger, 2008). A low x^2 indicates that the proposed model does not significantly differ from the data's likelihood under assumptions of the saturated model, a theoretical model that is fully specified (Sonnleitner, 2008). In other words a non- significant x^2 indicates that the hypothesized elementary parameters or radicals are able to explain the observed item parameters. As discussed above, elementary parameters are theory-based components that are hypothesized to impact item difficulty.

There is also a graphical test for the relative goodness-of-fit of the LLTM and Rasch model. If the LLTM fits well, the points with coordinates that represent estimates of item difficulty with the LLTM and the Rasch model should be scattered around the 45-degree line (Fischer, 1995). An example of the graphical test can be seen in Figure 2, which graphically represents the fit estimated by the ALR and visible in the accompanying tables. Fischer also

noted that often ALR tests turn out significant in empirical research and lead to rejection of the LLTM even when the graphical goodness-of-fit indicates a good match between Rasch and LLTM item difficulties. The graphical test indicated a good fit with coordinates scattered around the 45-degree line.

After establishing the fit of the Rasch model, LLTM was applied. A weight matrix was generated and tested by applying the LLTM. The structure matrix consisted of 11 difficulty factors. Table 5 shows the 11-factor structure matrix for the items of the Wonderlic. The result of ALR testing the fit of the model is shown in Table 6. According to the low $\chi 2$-value, data given the tested model does not significantly differ from the data's likelihood under the assumption of the saturated model, which is a theoretical fully-specified model. The low $\chi 2$-value means that the postulated 11 radicals are able to explain the observed item parameters. The graphical goodness- of- fit test also indicated a good fit with coordinates scattered around the 45-degree line.

Table 7 shows each radical's contribution to item difficulty. As one can see, 6 of the 11 proposed incidentals contribute significantly to item difficulty. Specifically, results indicate that language ambiguity, negative wording, response option type, math content, verbal content, and colloquial content are significant radicals contributing to item difficulty. The significance of language ambiguity, the use of negative wording and colloquial content on item difficulty are particularly interesting due to the construct-irrelevant nature of these factors. According to item generation theory construct-irrelevant factors or incidentals should not impact item difficulty. Neither of these factors taps the intelligence construct, yet these factors impact the difficulty of these items. These findings will be further addressed in the next section.

In coding the Wonderlic on item content, general as well as specific codes were created. The 11-factor model utilized general content codes that identified verbal and math items. However, the items content was also operationalized in greater detail. Math items were coded as geometry items, items requiring calculations only, and items with word problems. Verbal items were coded as items requiring knowledge, sentence construction items, and analytical items. In trying to gain a deeper understanding of item difficulty the 15-factor LLTM was applied. The results of ALR testing the fit of the model are shown in Table 8. According to the low $\chi 2$-value data given the tested model does not significantly differ from the data's likelihood under the assumption of the saturated model. Similarly, the graphical goodness-of-fit test indicated a good fit with coordinates scattered around the 45-degree line. Thus, the 15 radicals are able to explain the observed item parameters. Table 9 shows each radical's contribution to item difficulty. As evident, in this model 4 of the proposed 15 radicals are significant. Specifically, language ambiguity, negative wording, response option type, and colloquial content are significant radicals contributing to item difficulty.

It is interesting to note that in the 11-factor model math and verbal contents were significant. However, in the 15-factor model, none of the sub-factors emerged as significant. At the same time, the other factors, including language ambiguity, negative wording, response option type, and colloquial knowledge are significant factors or radicals impacting item difficulty. Perhaps content can impact difficulty on a broad level and when decomposed loses its significance. It is interesting to note that the four significant factors are construct-irrelevant. They are not related to the test content but yet they impact item difficulty.

Chapter 7: Discussion

The purpose of the current paper was to examine the impact of general stimulus features on item difficulty within the context of item generation theory. The goal of this work was to gain an understanding about the impact of item stimulus features that could improve the quality and fairness of standardized tests as well as enable more efficient generation of parallel test items. Item generation theory utilizes existing theoretical content knowledge and applies it to test construction for the purposes of producing test items. Through the radical and incidental approach, item generation allows test developers to produce parallel test forms without massive item pretesting. Thus, item generation offers an alternative to the daunting test construction approach. However, to utilize item generation, test developers must have a concrete knowledge about which factors are radicals and which are incidentals. Item generation theory forces researchers to think a priori about theoretical factors that contribute to difficulty.

The present study examined the impact of item stimulus features on item difficulty. Prior research has offered a number of frameworks about the sources of item difficulty (e.g., Carpenter et al., 1990; Freedle & Kostin, 1997; Gorin & Embretson, 2006; Irvine, Dann, & Anderson, 1990; Roccas & Moshinsky, 2003). Most of the frameworks, however, are focused on cognitive processes and not on the stimulus features of test items. These frameworks define difficulty in terms of cognition of the individual completing test items. In other words, the amount of cognitive processing necessary to respond correctly to a test item is what makes items difficult from the point of view of cognitive processing. These frameworks do not define difficulty based on item properties. Understanding the cognitive processes involved in approaching and solving test items is of utmost importance in understanding the full complexity of item difficulty. However, understanding the impact of item stimulus features on item difficulty is also crucial.

Gaining a deeper understanding of the impact item properties have on item difficulty can aid the goals of the item generation theory, which is to create better test items and to simplify the process of creating parallel tests.

The present research proposed and tested a general item difficulty framework that was focused on stimulus features of items. The proposed framework utilized the approach of item generation theory, which allows items to be decomposed into the factors that are hypothesized to impact difficulty. To test the proposed framework, the study utilized LLTM, an analytical approach that expresses item difficulty in terms of underlying stimulus complexity.

Summary of Findings

The results of the study indicate that certain item stimulus features impact item difficulty. From the examined item stimulus features, it was found that language ambiguity, negative wording, constructed-response items, and colloquial knowledge impact item difficulty. Specifically, item difficulty increased as the presence of each factor increased. For example, greater language ambiguity resulted in greater item difficulty. However, it is important to note that the findings discussed below are limited to cognitive ability tests. The current study was conducted utilizing the Wonderlic, which is a test of general cognitive ability. Thus, the following discussion is limited to the cognitive ability domain. The generalizability of these findings to other cognitive ability tests and tests of other constructs is not known, but given the high correlations among cognitive ability tests and similarity in content, it is likely that there will be some degree of generalizability. Item generation theory forces test developers to articulate radicals and incidentals that may contribute to item difficulty (Lievens & Sackett, 2007). Given the theoretical nature of item generation theory it is expected that radicals are not universal but

are particular to a given test or a construct of interest. However, there may be some universal aspects for different tests of the same construct that use similar test designs.

Prior research on language ambiguity has examined its role and impact from a cognitive point of view. It has been reasoned that multiple-meanings of ambiguous words leads to longer processing (Just & Carpenter, 1980). It has also been found that ambiguity negatively impacts the quality of one's data, as it increases the number of "no opinion" responses and skews a response into positive or negative directions (Coombs & Coombs, 1976; Velez & Ashworth, 2007). As a stimulus feature, Sonnleitner (2008) examined the impact of language ambiguity on a reading comprehension test and found that this stimulus feature increases item difficulty. The definition of ambiguity utilized by Sonnleitner was not clear. Therefore, the present study extended the inquiry of language ambiguity with a clear operationalization and a test, the Wonderlic, which includes different types of items, beyond reading comprehension. It was found that language ambiguity, defined as the number of alternative meanings for a word, is a statistically significant elementary parameter contributing to item difficulty. In the present study, the presence of this stimulus feature increased item difficulty. In the context of item generation theory, language ambiguity should be viewed as a radical that impacts item difficulty, not an incidental that can be used to create item variants. Therefore, in creating test item variants, test developers should be aware that when words on item variants are more ambiguous than on the original test item, defined as having more alternative meanings, the difficulty of that test item increases.

The present study also found that negative wording is a statistically significant elementary parameter contributing to item difficulty. Including negative words on test items makes items more difficult. Past research has investigated negative wording from the cognitive

processing perspective. Negative wording impacts processing and negatively worded statements have been shown to be more difficult to process than affirmative statements (e.g., Carpenter & Just, 1975). Due to an added layer of complexity in processing negatively worded statements, test developers have long avoided the use such items (Gorin, 2005). The present research, however, extended the investigation of negative wording from a cognitive process that occurs when individuals face such items to a stimulus feature. The present study found that negative wording is a statistically significant elementary parameter contributing to item difficulty. The presence of negatively worded statements increases item difficulty. Thus, as with language ambiguity, negative wording carries a radical function. Due to cognitive processing research, many researchers try to avoid negatively–worded items. That, however, may not always be possible. Showing that negatively-worded statements are radicals underscores the caution one must use in item development. Therefore, if original test items do not have negatively worded statements, item variants should not have negative wording either. Including negatively worded items in item variants would increase the difficulty of those variants. Negative wording is a radical function and should not be used as an incidental in generating parallel versions of test items.

Both language ambiguity and negative wording can artificially increase item difficulty. Thus, within the item generation framework both factors are radicals. Therefore, knowing that language ambiguity and negative wording impact item difficulty can aid in test development and prevent the utilization of radicals as incidentals in generating item variants. This knowledge can aid in the process of developing test items that do not include these factors. As a result, test developers will be able to create better test items and test item variants that vary true incidentals, ones that do not impact item difficulty.

The present study has also found that response construction is a statistically significant elementary parameter contributing to item difficulty. Prior research has shown that constructed-response items are more difficult than selected-response items (Downing, 2004). The absence of choices requires a test-taker to recall information and construct a response. In a study investigating subgroup differences on a multiple-choice and a constructed-response test of scholastic achievement, Edwards and Arthur (2007) found that both groups score lower on a constructed-response test. The present research showed that response construction is a statistically significant elementary parameter contributing to item difficulty as the presence of a constructed response option increases item difficulty. Within the item generation theory framework the response construction can be considered a radical as it impacts item difficulty. Thus, when creating item variants for parallel tests, test-developers should not alternate between constructed response items and multiple-choice items.

At the same time, constructed-response items are often necessary in testing as they can measure more complex skills than cannot be evaluated by multiple-choice items. For example, in literature a test-taker may be asked to write an essay comparing and contrasting two stories and in mathematics a test-taker may be asked to write a mathematical equation or a diagram (Livingston, 2009). When such knowledge is required a multiple-choice item falls unsatisfactorily short of measuring the necessary skills. Thus, the construct of interest is important in one's decision to utilize constructed-response items.

Furthermore, although response construction is a statistically significant elementary parameter contributing to item difficulty, it is also possible that multiple-choice items actually decrease item difficulty. It is also possible that multiple-choice items decrease difficulty for certain groups. For example, Edwards and Arthur (2007) found that that constructed-response

items result in lower test scores for both African-American and White test-takers. However, although both groups had lower mean scores on the constructed-response test, there was a substantial 39% reduction in subgroup differences compared with the multiple-choice test. Based on these results, it is possible that multiple-choice items may decrease difficulty differentially for White test-takers and that constructed-response items may offer a more fair testing situation for all groups. Further investigation of the impact of construct-response versus multiple-choice items on difficulty as well as the impact of item-type choice on subgroup differences is warranted to gain a better understanding of this issue. At the same time, it is important for test-developers to be aware of this issue and be driven by the construct of interest in generating test items with appropriate response options. It is also important to remember that response construction as a radical. Thus, test-developers should not vary response options on item variants.

Another factor that can impact item difficulty is item content. The present research illustrated that colloquial knowledge is a statistically significant elementary parameter contributing to item difficulty. According to Sonnleitner (2008), the text of an item can impact difficulty. Each text requires a construction of a model and therefore, depends on content-specific knowledge and motivation to work with a particular topic. The current study predicted that math content and colloquial knowledge content impact item difficulty.

The hypotheses in this study were based, in part, on research by Freedle and Kostin (1997) who found that African-American test-takers performed worse on items that have science content. The results of this study indicated that colloquial knowledge is a statistically significant radical contributing to item difficulty. Items that include colloquial US expressions and cultural knowledge measure something that is not relevant to the construct of interest, but which impacts item difficulty. This finding fits well with prior research indicating that seemingly construct-

irrelevant aspects of test items, including reliance on prior knowledge, and presence of a cultural context, which may not be easily recognizable to all test-takers, impact test performance (e.g., Freedle & Kostin, 1997; Fagan & Holland, 2007; Goldstein, Scherbaum & Yusko, 2009).

The inclusion of US colloquial expressions in a test of intelligence makes one think about what these questions can measure. For example, one of colloquial the questions on the Wonderlic asks a test-taker to decide whether the meaning of two sentences is similar, contradictory or neither. The two sentences are two separate colloquial expressions and a test-taker is asked to evaluate their similarity. To respond correctly to this item, a test-taker must be able to process information, which is one definition of intelligence (Fagan, 2000). However, one also needs to have a solid knowledge of English colloquial expressions to respond to the question correctly. Arguably, a person may be well able to do the first task, but without the requisite knowledge, a test-taker will not be able to answer such a test item correctly. Thus, this item, in addition to intelligence, measures the knowledge of English colloquial expressions, which is irrelevant for the purposes of assessing intelligence. Furthermore, not only does the item have construct-irrelevant content, such an item places individuals for whom English is not a first language at a disadvantage over native speakers. There could also be additional regional, class and racial differences in the way test-takers respond to such test items.

Thus, given that colloquial knowledge has a radical function, cultural sensitivity should be important in test development. Of course, there may be situations where a deep knowledge of language, culture, and idioms may be required. In such cases, questions that gauge respondents' knowledge of US colloquial expressions are relevant. However in intelligence testing, knowledge of colloquial expressions is not relevant. Rather such questions tap construct-irrelevant factors that impact item difficulty. Thus, in developing test items, colloquial

expressions, even common idioms should be avoided. If US colloquial knowledge is not the construct of interest, including such items impacts difficulty.

In testing the content hypotheses, interesting results were found. It was proposed that colloquial and math content would be statistically significant predictors of item difficulty. Items were coded as verbal, math or colloquial test questions. It is interesting to note that in the 11-factor model math, verbal, and colloquial contents were significant elementary parameters contributing to item difficulty. However, in the 15-factor model, which further broke down item content, only colloquial content emerged as significant. None of the math contents emerged as significant, contrary to the proposed hypothesis. One reason for this may be that when decomposed, item content loses its significance, while at the broad level (i.e., math) it remains significant. It is also possible that better or different operationalizations are needed to further understand the impact of math content on item difficulty.

In testing other hypotheses, it was found that rule complexity, word rarity, item length, and number of response options were not statistically significant predictors of item difficulty. If future research continuously shows that these factors are incidentals within the cognitive ability testing domain, these factors can be varied to produce item variants. It is important to keep in mind that this study was conducted with a cognitive ability test-the Wonderlic. Therefore, results and implications are discussed within the framework of cognitive testing. However, further inquiry is necessary to establish that these factors are indeed incidentals for in this domain.

It may also be worthwhile to further investigate these factors in future research. For some factors, different operationalizations may be necessary to fully understand the impact of these factors. For example, it has been found that word rarity does not impact item difficulty. However, it has also been shown that vocabulary level impacts difficulty of test item when

difficult words are needed to convey the semantic relation of the test items (Sheehan & Mislevy, 2001). Therefore, difficult words in parts of the sentence that are not necessary to convey semantics of the item may not increase item difficulty. Future research, for example, may investigate this operationalization of word rarity. It would be worthwhile to examine the impact word rarity of semantically important word has on test item difficulty. The present study showed that colloquial knowledge content impacts item difficulty. Future research can look into operationalizaitons that are different from the math and verbal taxonomy discussed in this study to see if other item content impacts item difficulty.

Thus, there are many other stimulus factors that future research can explore within the framework of cognitive ability testing and with other test-types, such as personality. It is important to keep in mind that the item generation theory is theoretically driven. Thus, it is expected that different radicals and incidentals will be appropriate for different item content. The results of this study, however, show the utility of the radical and incidental approach in test item generation.

Implications

The results of this study show that factors one may consider as potential incidentals and vary to create parallel test items are, in fact, radicals that impact item difficulty. Therefore, gaining a deeper understanding about stimulus factors that impact difficulty is important as it can prevent the utilization of these radicals as incidentals in generating item variants. Within the item generation framework this knowledge can be utilized to create better test items and generate parallel test versions without extensive item pretesting. Understanding the impact of item stimulus features on item difficulty is important from the point of view of test-item development. Knowing which item stimulus features contribute to item difficulty can aid in constructing items

without these features. Thus, further research in this area can be applied to generating better test items and gaining more appreciation for the impact seeming incidentals could have on item difficulty.

Item generation theory forces researchers to think theoretically about factors that may impact item difficulty and vary radicals and incidentals in producing item variants. The conceptual emphasis of item generation theory is crucial in motivating future research to further investigate potential difficulty factors that are focused on item stimulus features. Discovering what drives item difficulty and utilizing the radical and incidental approach to create parallel tests can eliminate the need for the long laborious process involved in item development for mass testing.

This study is among the few to examine the impact of item stimulus features on item difficulty. The factors selected for this study have been studied in past research endeavors, most within the cognitive-processing frameworks of item difficulty. Potentially, there is an infinite number of stimulus factors that one can examine. Gaining a better understanding of item stimulus features will improve the categorization of these features into radicals and incidentals.

The purpose of item generation is not to alter test difficulty for nefarious purposes. The current paper makes no recommendations about using the factors that have been found to impact item difficulty to manipulate it. However, it should be acknowledged that the manipulation of difficulty is a potential caveat of learning and gaining a deeper understanding about the factors that impact item difficulty. Nonetheless, the findings of the current paper can be utilized to provide test developers with the necessary information to generate parallel tests and reduce construct-irrelevant variance. Ultimately, despite this caveat, further knowledge of this area can lead to more efficient test generation and the development of better test items.

The goal of this study was to shed a light on the impact of item stimulus features on item

difficulty. This knowledge can result in the development of better test items and more efficient test-item generation. Gaining a deeper knowledge about the impact item stimulus features can have on item difficulty and understanding how item stimulus features fit into the radical and incidental framework of item generation theory will allow test developers to develop better items and item variants by varying incidentals without altering item difficulty level.

Limitations

Of course, there are limitations with the results of this study. As discussed above, the factors investigated from this study were limited to those that have been researched in the past, mostly within the cognitive processing framework. These factors are not all encompassing and there can potentially be an infinite number of stimulus factors that may impact difficulty. Thus, stimulus factors evaluated in this study should be viewed as a starting point. Furthermore, the salience of stimulus features will depend on the construct of interest. Therefore, future investigations into the radical functions of item stimulus features should be guided by the construct of interest.

Another limitation concerns the design the Wonderlic and the archival nature of this study, which precluded experimental manipulation of factors. The items on the Wonderlic were coded according to the difficulty factors. However, most questions had numerous factors coded within them. In the future, it would be interesting to develop test items that only have one difficulty factor to evaluate and compare an impact of such difficulty factor to the impact of multiple factors. Tests items can be designed that individually have only one difficulty factor, such as word rarity. Then, one could examine the impact of word rarity without the presence of other factors.

Finally, item difficulty factors investigated in this study with test items from the

Wonderlic Personnel Test, a test of traditional cognitive ability. The discussion of radicals and construct-irrelevancy was based on intelligence testing. Although test developers should be aware of ambiguous language and negative wording in test item development, other stimulus factors may be more relevant for other types of testing, regardless of their impact on item difficulty. An obvious example is ability to write short essays on an exam. Such ability cannot be tested with multiple-choice items. Although response-construction impacts item difficulty, in a situation where multiple-choice items are not appropriate, response construction should be used. Thus, there may be numerous item stimulus features that impact item difficulty. However, the construct of interest should be an important consideration in evaluating item stimulus features, their impact on item difficulty, and the decision to utilize them in test item generation. Thus, it is important to evaluate the utility of the radical and incidental approach of item generation theory with stimulus features of items with other aptitude and non-aptitude tests.

The question posed by this study is important in mass testing situations where multiple parallel test versions are necessary. Test developers typically write large number of items and pretest them to identify items of similar difficulty levels (Lievens & Sacket, 2007). Item generation theory offers an alternative approach whereby test developers can create parallel items by varying incidental factors, while maintaining difficulty-causing radical factors unchanged. Utilizing the radical and incidental approach forces test developers to think about factors that impact item difficulty. A deeper understanding about the sources of item difficulty can provide test developers with sufficient knowledge to approach test construction through the radical and incidental approach, without item pretesting. Prior research has examined item difficulty mostly from a cognitive processing point of view. Very few studies have evaluated the impact of general item stimulus features on item difficulty. Understanding difficulty from the

cognitive processing point of view is important in developing theories of information processing and in training individuals to approach and solve test items efficiently. However, understanding the impact of item stimulus features on item difficulty is important from test-development and item writing development. The present study investigated the impact of general item stimulus features on item difficulty. The results of this study show that gaining such understanding is important as factors one may consider as potential incidentals and vary to create parallel test items are, in fact, radicals that impact difficulty. Therefore, it is important to continue this line of inquiry and further investigate the impact of other item stimulus features on item difficulty. A better knowledge of this content area can result in simpler item development process and the generation of better test items.

Table 1. *Highlighted Item Difficulty Sources*

Difficulty Factor C (cognitive element) S (stimulus element)	Research Evidence	Difficulty Element	If Stimulus feature: General (G) or Item Specific (S)
Rule Complexity (C) & (S)	Pellegrino & Glaser (1980)	Rule complexity	(G)
Language (S)	Gorin & Embretson (2006)	Word rarity	(S) Can be applied to different item-types
	Roccas & Moshinsky (2003) Embretson &Wetzel (1987)	Word rarity Word rarity of response options	(S) Can be applied to different item-types (G)
	Sonnleitner (2008)	Language ambiguity	(S) Can be applied to different item-types
	Gorin (2005 Roccas & Moshinsky (2003	Negative component	(S) Can be applied to different item-types
Sentence Length (S)	Berk et al. (2001) Green (1984)	Sentence length is not a predictor of difficulty	(S) Can be applied to different item-types
Number of Response Options (S)	Sonnleitner (2008)	Number of response options	(S) Can be applied to different item-types
Item Content (S)	Freedle & Kostin (1997) Sonnleitner (2008)	Item Content	(G)

Note. Item Specific (S) refers to a particular item type.

Table 2. *Descriptive Statistics for All Variables*

	M	Min	Max	*SD*	Skeweness	*SE*	Kurtosis	*SE*
Item 1	0.83	0	1	0.37	-1.77	0.83	1.15	0.17
Item 2	0.73	0	1	0.45	-1.03	0.83	-0.95	0.17
Item 3	0.04	0	1	0.19	4.99	0.83	23.05	0.17
Item 4	0.83	0	1	0.38	-1.71	0.83	1.17	0.17
Item 5	0.96	0	1	0.19	-4.78	0.83	20.96	0.17
Item 6	0.60	0	1	0.49	-0.41	0.83	-1.83	0.17
Item 7	0.63	0	1	0.48	-0.54	0.83	-1.71	0.17
Item 8	0.92	0	1	0.27	-3.20	0.83	8.23	0.17
Item 9	0.78	0	1	0.41	-1.38	0.83	-0.10	0.17
Item 10	0.43	0	1	0.50	0.27	0.83	-1.93	0.17
Item 11	0.81	0	1	0.39	-1.59	0.83	0.54	0.17
Item 12	0.94	0	1	0.23	-3.79	0.83	12.42	0.17
Item 13	0.41	0	1	0.49	-0.42	0.83	-1.87	0.17
Item 14	0.93	0	1	0.26	-3.29	0.13	8.85	0.17
Item 15	0.87	0	1	0.34	-0.47	0.83	2.69	0.17
Item 16	0.84	0	1	0.36	-2.16	0.83	1.60	0.17
Item 17	0.60	0	1	0.49	-1.89	0.83	-1.83	0.17
Item 18	0.42	0	1	0.49	-0.41	0.83	-1.91	0.17
Item 19	0.31	0	1	0.46	0.32	0.83	-1.33	0.17

Table 2.

Item 20	0.76	0	1	0.43	-0.54	0.83	-0.58	0.17
Item 21	0.66	0	1	0.48	-0.66	0.83	-1.57	0.17
Item 22	0.01	0	1	0.10	9.66	0.83	91.55	0.17
Item 23	0.80	0	1	0.40	-1.47	0.83	0.15	0.17
Item 24	0.51	0	1	0.50	-0.32	0.83	-2.00	0.17
Item 25	0.52	0	1	0.50	-0.70	0.83	-2.00	0.17
Item 26	0.45	0	1	0.50	0.19	0.83	-1.97	0.17
Item 27	0.22	0	1	0.42	1.35	0.83	-0.19	0.17
Item 28	0.50	0	1	0.50	0.20	0.83	-2.00	0.17
Item 29	0.46	0	1	0.50	0.15	0.83	-1.98	0.17
Item 30	0.63	0	1	0.48	-0.53	0.83	-1.72	0.17
Item 31	0.22	0	1	0.41	1.36	0.83	-0.14	0.17
Item 32	0.10	0	1	0.31	2.60	0.83	4.75	0.17
Item 33	0.20	0	1	0.40	1.53	0.83	0.34	0.17
Item 34	0.25	0	1	0.43	1.15	0.83	-0.68	0.17
Item 35	0.18	0	1	0.38	1.68	0.83	0.84	0.17
Item 36	0.26	0	1	0.43	1.19	0.83	-0.60	0.17
Item 37	0.04	0	1	0.19	4.60	0.83	19.16	0.17
Item 38	0.20	0	1	0.40	1.53	0.83	0.34	0.17
Item 39	0.02	0	1	0.15	6.35	0.83	38.45	0.17
Item 42	0.04	0	1	0.19	4.83	0.83	21.35	0.17

Table 2- Continued

Item 43	0.12	0	1	0.03	29.39	0.83	864.00	0.17
Item 44	0.07	0	1	2.48	3.50	0.83	10.20	0.17
Item 45	0.06	0	1	0.08	13.05	0.83	168.79	0.17
Item 46	0.00	0	1	0.00	---	---	---	---
Item 47	0.04	0	1	0.19	5.00	0.83	23.05	0.17
Item 48	0.04	0	1	0.06	16.91	0.83	284.66	0.17
Item 49	0.01	0	1	0.03	29.39	0.83	864.00	0.17
Item 50	0.00	0	1	0.00	----	---	---	---

Table 3. *Results of Andersen's Likelihood-Ratio test for Rasch Model1*

Model	*df*	x^2(LRT)	x^2α=0.01
Model1	49	628.41	74.94

Table 4. *Results of Andersen's Likelihood-Ratio test for Rasch Model2*

Model	*df*	x^2(LRT)	x^2α=0.01
Model2	49	55.42	74.94

Table 5. Structure matrix for LLTM

Item #	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11
1	0	1	0	1	0	0	0	0	0	1	0
2	1	1	1	1	0	1	0	1	0	1	0
3	0	0	0	0	0	0	1	0	1	0	0
4	1	1	1	1	0	0	0	1	0	1	0
5	1	0	1	0	1	1	0	0	0	0	1
6	0	0	1	0	1	1	1	0	0	1	0
7	0	1	1	1	0	0	0	0	0	1	0
8	1	1	1	0	0	0	1	0	1	0	0
9	1	1	1	1	0	0	0	1	0	1	0
10	1	1	0	1	0	1	0	1	0	1	0
11	0	1	1	0	0	0	1	0	1	0	0
12	0	0	1	1	1	1	0	0	0	1	0
13	0	0	1	0	1	0	1	0	1	0	0
14	1	0	1	1	1	1	0	0	0	1	0
15	1	0	1	0	0	0	1	0	1	0	0
16	1	1	1	1	0	1	0	1	0	1	0
17	0	0	0	1	1	1	0	0	0	0	1
18	0	1	1	1	0	0	0	0	0	1	0
19	0	0	0	0	1	1	1	0	0	1	0
20	0	0	0	0	0	0	0	1	1	0	0
21	0	0	0	1	1	1	0	1	1	0	0
22	1	1	0	0	0	0	1	0	1	0	0
23	1	0	1	0	0	0	1	0	1	0	0
24	0	1	1	1	1	1	0	0	0	1	0
25	0	0	1	0	1	0	1	0	1	0	0
26	0	0	1	0	1	1	1	0	1	0	0
27	0	1	1	1	1	1	0	0	0	0	1
28	0	1	1	0	0	0	0	1	0	1	0
29	0	1	1	1	0	0	0	0	0	1	0
30	0	0	1	1	1	1	0	0	0	1	0
31	0	1	1	0	0	1	0	1	0	0	1
32	0	1	0	0	1	1	1	0	1	0	0
33	0	0	1	1	0	0	0	0	0	1	0
34	0	1	1	1	1	1	0	0	0	0	1
35	1	1	1	0	1	1	1	0	1	0	0

Table 5.

Item #	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11
36	0	0	0	0	0	0	0	1	1	0	0
37	1	0	1	0	1	0	1	0	1	0	0
38	0	0	1	1	1	1	0	0	0	1	0
39	1	1	0	1	0	1	0	1	0	1	0
40	0	1	1	0	1	1	0	1	1	0	0
41	0	0	0	1	1	1	0	0	0	1	0
42	1	1	1	0	0	0	0	1	0	1	0
43	0	0	1	0	1	1	1	0	1	0	0
44	0	0	1	1	1	1	0	0	0	0	1
45	1	0	1	1	1	0	1	0	1	0	0
46	1	1	1	0	1	0	1	0	1	0	0
47	1	0	1	1	0	1	0	1	0	1	0
48	1	0	0	0	1	0	0	0	1	0	0
49	1	0	1	0	0	0	1	0	1	0	0
50	1	1	0	0	1	0	1	0	1	0	0

Note: Note: Factor 1 is rule complexity, Factor 2 is word rarity, Factor 3 is language ambiguity, Factor 4 is negative wording, Factor 5 is item length, Factor 6 is total length, Factor 7 is response option-type, Factor 8 is number of response choices, Factor 9 is math content, Factor 10 is verbal content, and Factor 11 is colloquial content.

Table 6. *Results of Andersen's Likelihood-Ratio test for LLTM-Wonerlic11*

Model	*df*	x^2(LRT)	x^2α=0.01
Wonderlic-11	38	47.97	61.18

Table 7. *Contribution of radicals to item difficulty-Wonderlic11*

Radical	Basic Standard Parameters	Significance
Rule Complexity	-0.004	n.s.
Word Rarity	0.013	n.s.
Language Ambiguity	-0.050	$\alpha < 0.05$
Negative Wording	-0.105	$\alpha < 0.01$
Item Length	0.019	n.s.
Total Length	0.037	n.s.
Response Options (multiple choice versus construct responses)	-0.131	$\alpha < 0.01$
Number of Response Choices	-0.048	n.s.
Math Content	-0.146	$\alpha < 0.05$
Verbal Content	-0.167	$\alpha < 0.05$
Colloquial Content	-0.114	$\alpha < 0.05$

Table 8. *Results of Andersen's Likelihood-Ratio test for LLTM- Wonderlic15*

Model	*Df*	x^2(LRT)	x^2α=0.01
Wonderlic-14	34	36.28	56.08

Table 9. *Contribution of radicals to item difficulty-Wonderlic15*

Radical	Basic Standard Parameters	Significance
Rule Complexity	-0.046	n.s.
Word Rarity	0.034	n.s.
Language Ambiguity	-0.061	$\alpha < 0.05$
Negative Wording	-0.093	$\alpha < 0.05$
Item Length	0.056	n.s.
Total Length	0.071	n.s.
Response Options (multiple choice versus construct responses)	-0.134	$\alpha < 0.01$
Number of Response Choices	-0.007	n.s.
Geometry	-0.151	n.s.
Calculation	-0.953	n.s.
Word Problems	-0.045	n.s.
Verbal Knowledge	-0.025	n.s.
Sentence-Completion	-0.016	n.s.
Analytical	-0.006	n.s.
Colloquial Content	-0.187	$\alpha < 0.01$

Figure 1. *Example of a Rasch Model Item Characteristic Curve where the difficulty parameter, b=1.0*

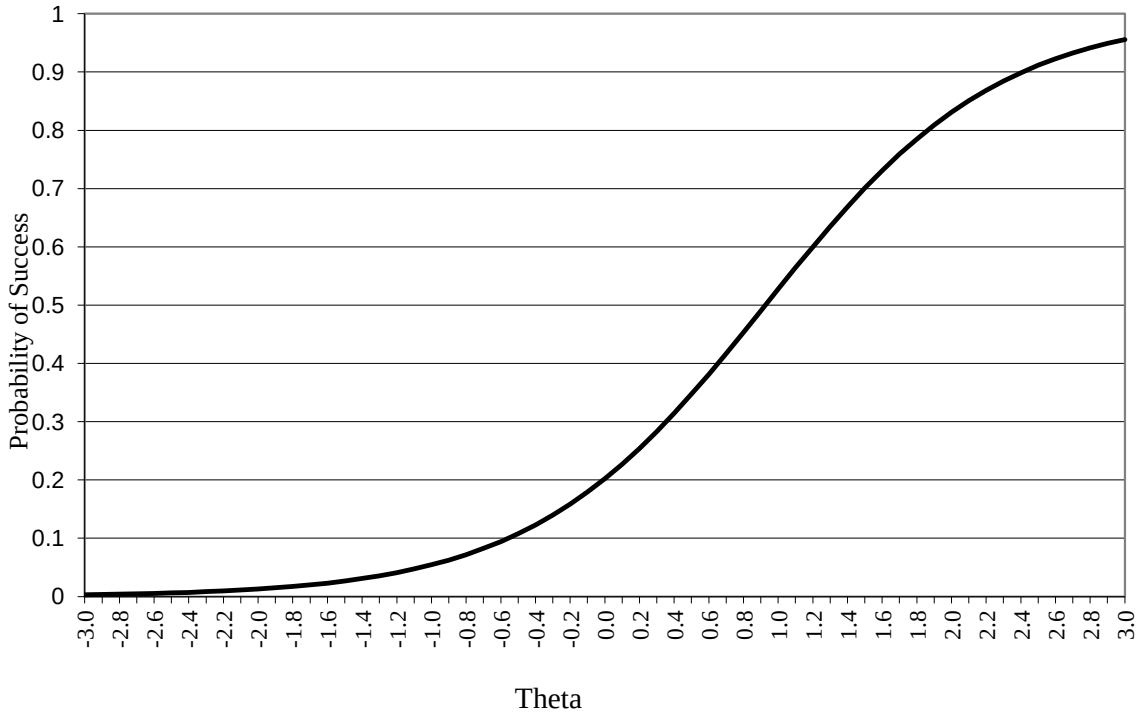
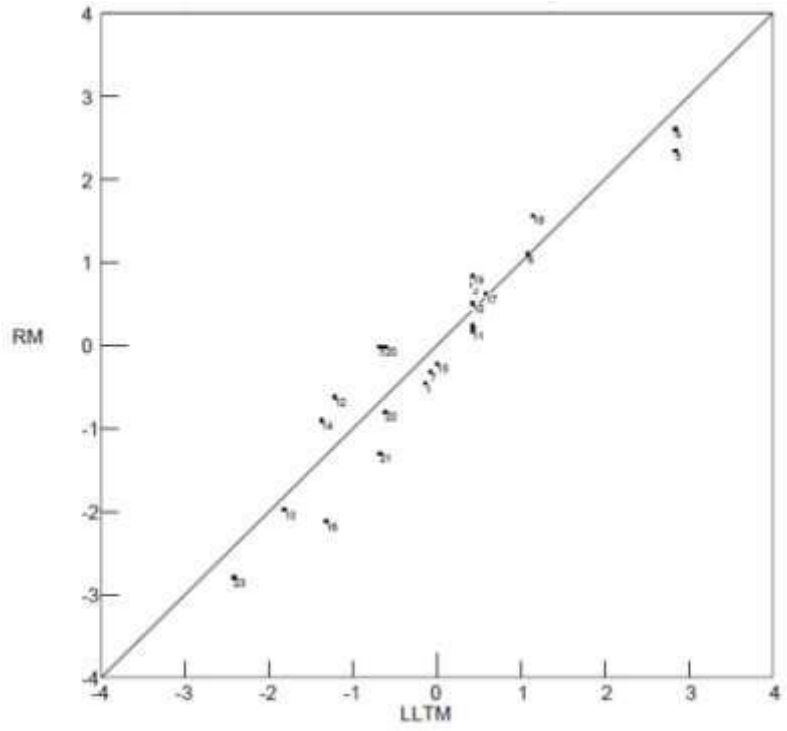


Figure 2. *Example of a Graphical Goodness of Fit Test*



References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika,* 38, 123-40.

Bejar, I.I., Chaffin, R., & Embretson, S. (1991). *Cognitive and psychometric analysis of analogical problem solving.* New York: Springer-Verlag, 1991.

Berk, E.J.V., Lohman, D.F., & Cassata, J.C. (2001). What does a verbal test measure? A new approach to understanding sources of item difficulty. Paper presented at annual meeting of the American Educational Research Association, Seattle, WA.

Carpenter, P.A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic model of sentence verification. *Psychological Review, 82,* 45-73.

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

Coombs, C. H., & Coombs, L. C. (1976). ”Don’t know”: Item ambiguity or respondent uncertainty? *Public Opinion Quarterly,* 40, 497-514.

De Ayala, R.J. (2009). *The theory and practice of Item Response Theory.* New York, NY: Guilford Press.

Demitrov, D.M., & Raykov, T. (2003). Validation of cognitive structures: a structural equation modeling approach. *Multivariate Behavioral Research, 38,* 1-23.

Downing, S.M (2004). Selected-response item formats in test development. In Downing, S. M., & Haladyna, T. M. (Eds.). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses.

Journal of Applied Psychology, *68*, 363-373.

Edwards, B.D., & Arthur, W. (2007). An examination of factors contributing to reduction in subgroup differences on a constructed-response paper-and-paper test of scholastic achievement. *Journal of Applied Psycholgy,92,* 794-801.

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

Embretson, S.E. (1995). The role of working memory capacity and general control processing in intelligence. *Intelligence, 20,* 169-189.

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

Embretson, S.E., & Reise, S.P. (2000). *Item response theory of psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Fagan, J.F. (2000). A theory of intelligence as processing: Implications for society. *Psychology, Public Policy, and Law, 6, 168-179.*

Fagan, J.F., & Holland, C.R. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence, 35,* 319-334.

Ferreter, J., Goldstein, H., Scherbaum, C., Yusko, K., & Jun, H. (2008). *Reducing adverse impact using a nontraditional cognitive ability assessment*. Poster presented at the Society for Industrial & Organizational Psychology, San Francisco, CA.

Fischer, G.H. (1973). The linear logistic test model as an instrument of educational research.

Acta Psychologica, 36, 359-374.

Fischer, G.H. (1995). The linear logistic latent trait model. In G.H. Fischer & I.W. Molenaar (Eds.). *Rasch Models: Foundations, recent developments, and applications*. New York, NY: Springer-Verlag Inc.

Fischer, G.H., & Formann, A.K. (1982). Some applications of logistic latent trait models with linear constraints on parameters. *Applied Psychological Measurement, 6,* 397-416.

Fischer, G. H. & Ponocny-Seliger, E. (1998). Structural Rasch modeling: Handbook of the usage of LPCM-WIN 1.0. Groningen: ProGAMMA.

Fodor, J. D., Fodor, J. A., & Garrett, M. F. (1975). The psychological unreality of semantic representations. *Linguistic Inquiry, 4,* 515–531.

Freedle, R., & Kostin, I. *(*1997*)*. Predicting Black and White differential item functioning in verbal analogy performance. *Intelligence, 24,* 417-444.

Gierl, M.J., & Leighton, J.P (2004). Review of Item Generation for Test Development. *Journal of Educational Measurement, 41,* 69-72.

Gittler, G., & Gluck, J. (1998). Differential transfer of learning: Effects of instruction in descriptive geometry on spatial test performance. *Journal of Geometry and Graphics, 2,* 71-84.

Goldstein, H. W., Scherbaum, C. A., & Yusko, K. (2009). Adverse impact and measuring cognitive ability. In J. Outtz's (Ed.) *Adverse impact: Implications for organizational staffing and high stakes testing* (pp. 95-134). New York: Psychology Press.

Gorin, J.S. (2005). Test design with cognition in mind. *Education and Measurement: Issues and Practice, 25,* 21-35.

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

Green, K. *(*1984*).* Effects of item characteristics on multiple-choice item difficulty. *Educational and Psychological Measurement, 44,* 551-561.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park: Sage.

Hohensinn, C., & Kubinger, K.D. (2008). On varying item difficulty by changing the response format for a mathematical competence test. *Austrian Journal of Statistics, 38,* 231-239.

Hough, L.M., Oswald, F.L., & Ployhart, R.E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9,* 12-194.

Irvine, S.H. (2002). The foundation of item generation for mass testing. In S.H. Irvine & P.C. Kyllonen (Eds.). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Irvine, S.H., Dann, P.L., & Anderson, J.D. (1990). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology, 81,* 173-195.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329-354.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50,* 311-327

Kubinger, K.D. (2009). Applications of the Linear Latent Trait Model in psychometric research. *Psychological Measurement, 69,* 232-244.

Kucera, H., & Francis, W.N. (1967). *Computational Analysis of Present-day American English*. Providence: Brown University press.

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

Lievens, F., & Sackett, P.R. (2007). Situational judgment in tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology, 92,* 1043-1055.

Linacre, J.M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328.

Livingston, S.A. (2009). *Constructed-response test questions: Why we use them; how we score them* (ETS Research Report). Princeton, NJ: Educational Testing Service.

Mair P., & Hatzinger, R (2007). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science, 49,* 26-43.

Mirriam-Webster Online Dictionary. Retrieved from http://www.merriam-webster.com/

Mislevy, R. J., Levy, R., Kroopnick, M., & Rutstein, D. (2008). Evidentiary foundations of mixture item response theory models. In G. R. Hancock & K. M. Samuelsen (Eds.). *Advances in latent variable mixture models.* Charlotte, NC: Information Age Publishing.

Pellegrino, J.W., & Glaser, R. (1980). Components of inductive reasoning. In R.E. Snow, P.A. Federico, & W.E. Montague (Eds.). *Aptitude, learning and instruction; Cognitive process analyses of aptitude* (Vol. 1). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.). New horizons in testing: *Latent trait theory and computerized adaptive testing.* New York: Academic Press.

Reise, S.P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133-144.

Roccas, S., & Moshinsky, A. (2003). Factors affecting the difficulty of verbal analogies. *Applied measurement in education, 16,* 99-113.

Roth, P.L., Bevier, C.A., Bobko, P., Switzer, F.S.I., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54,* 297-330.

Sackett, P. R., Burris, L. R., and Ryan, A. M. (1989). Coaching and practice effects in personnel selection. In C. L. Cooper and I. T. Robertson (Eds.). *International Review of Industrial & Organizational Psychology*. West Sussex, England: John Wiley and Sons.

Scheuneman, J.D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27,* 109-131.

Sheehan, K. M., & Mislevy, R. J. (2001). An inquiry into the nature of the sentence completion task: implications for item generation. *Research Report RR-01-13.*Princeton, NJ: Educational Testing Service.

Sonnleitner, P. (2008). Using the LLTM to evaluate an item generating system for reading comprehension. *Psychology Science Quarterly, 50*, 345-362.

Velez, P., & Ashworth, S .D. (2007). The impact of item readability on the endorsement of the midpoint response in surveys. *Survey Research Methods, 1,* 69-74.

Wonderlic Associates, Inc. (1983). *Wonderlic Personnel Test Manual.* Northfield, IL: Author.