University of Wisconsin Milwaukee
UWM Digital Commons

Theses and Dissertations

May 2015

Using Differential Item Functioning to Test for Inter-rater Reliability in Constructed Response Items

Tamara Beth Miller
University of Wisconsin-Milwaukee

USING DIFFERENTIAL ITEM FUNCTIONING TO TEST FOR INTER-RATER RELIABILITY IN CONSTRUCTED RESPONSE ITEMS

by

Tamara B. Miller

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Educational Psychology

at

The University of Wisconsin-Milwaukee

May 2015

ABSTRACT
USING DIFFERENTIAL ITEM FUNCTIONING TO TEST FOR INTER-RATER RELIABILITY IN CONSTRUCTED RESPONSE ITEMS

by

Tamara B. Miller

The University of Wisconsin-Milwaukee, 2015
Under the Supervision of Professor Cindy M. Walker

This study used empirical and simulated data to compare and contrast traditional measures of inter-rater reliability to a novel measure of inter-rater scoring differences in constructed response items. The purpose of this research was to investigate an alternative measure of inter-rater differences, based in modern test theory, which does not require a fully crossed design for its calculation. The proposed, novel measure of inter-rater differences utilizes methods that are typically used to examine differential item functioning (DIF). The traditional inter-rater reliability measures and the proposed measure were calculated for each item under the following simulated conditions: three sample sizes (N = 1000, 500, and 200) and 4 degrees of rater variability: 1) no rater differences; 2) minimal rater differences; 3) moderate rater differences; and 4) severe rater differences. The empirical data were comprised of 177 examinees scored on 17 constructed response items by two raters. For each of the twelve simulated conditions, plus the empirical data set, each item had four measures of inter-rater differences associated with it: 1) an intraclass correlation (ICC); 2) a Cohen's kappa statistic; 3) a DIF statistic when examinees were fully crossed with the two raters and Rater 1's scores comprise the reference group while Rater 2's scores comprise the focal group; and 4) a DIF statistic when members of the focal and reference groups were mutually exclusive and nested within one rater. All indices were interpreted first using a significance level to calculate the Type I error and power for each index and second using a cut value to determine a False Positive Rate and a True Positive Rate for each index. Comparison of the findings from the two different criteria revealed that the DIF statistic derived from the nested design had a large False Positive Rate, therefore it was determined that this index was best interpreted using a significance level criterion. Additional simulation study results found that across the three different sample sizes the ICC and the Cohen's kappa both had Type I error rates of 0%; the DIF statistic from the fully crossed design had Type I error rate from 2% to 12%; while the DIF statistic from the nested design had a Type I error rate from 2% to 10%. Further results of the simulation found that, pertaining to moderate and severe rating differences modeled as constant, pervasive rater severity, the ICC and Cohen's kappa both had uniform power of 0% across all three sample sizes; while the fully crossed DIF statistic had a range of power from 89% to 100% across all three sample sizes; and the DIF statistic derived from a nested-design had power ranging from 44% to 100% across all three sample sizes. This combination of adequate power and low Type I error for the DIF statistic from the nested design is notable not only because it shows an ability to detect poor rater agreement, but it achieved this by using a design that does not require both raters to score all of the items. Finally, results from the empirical study, with a sample size of 177 examinees, provided an example of the application of the DIF statistic derived from the nested design in relation to the other three inter-rater reliability indices which all required a fully crossed design. The results of this investigation indicate that the proposed measure may provide a more achievable and more powerful indicator of inter-rater reliability for constructed response items under the investigated conditions, than traditional measures currently provide.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Using Differential Item Functioning to Examine Inter-rater Reliability in Constructed Response Items

Test developers strive to develop tests with scores that are both reliable and valid. In order to achieve adequate test score reliability many tests are comprised of dichotomously scored (i.e., scored either correct or incorrect) selected response (SR) items. There are many advantages to items that can be scored dichotomously. Some of these advantages include that the examinee can answer a large number of SR questions in a fairly reasonable time period. A large number of points is associated with increased reliability. An additional advantage is that a test comprised of a large number of items makes it possible to cover a broad range of content thereby providing a thorough sample of the examinee's knowledge (McClellan, 2010).

However, SR items also have disadvantages. One disadvantage is that there are skills and abilities that simply can't be tested using the multiple choice or true/false format. In these cases test developers may utilize what have come to be known as constructed response items. Constructed response items are those items that require test takers to construct a response rather than select from a list of options (McClellan, 2010). Examples of constructed response items in educational settings include writing essays, writing out an equation, or describing a biological process. If the response cannot be made in a pencil and paper format, then Livingston (2009) goes on to specify that type of item as a performance assessment task. According to this definition provided by Livingston, examples of performance assessments in the educational realm may include conversing in a foreign language or conducting research on an assigned topic. Within a clinical realm, such as physical therapy, performance assessments have historically been used to assess physiological or neuro-motor functions. As examples, a patient's balance response or degree of shoulder function are attributes that are not able to be assessed by a SR item.

Although Livingston (2009) makes a distinction between constructed response and performance assessment items based upon whether or not they can be completed using pencil and paper, for the remainder of this paper, both types of items will be referred to as constructed response items. Using only one term will make the communication more concise and for the purposes of this investigation, the pertinent factor regarding the constructed response items is that they are all scored by raters. This factor will be discussed in greater detail in subsequent sections.

In addition to being scored by raters, another characteristic of constructed response items is that they are typically scored polytomously (Penfield & Lam, 2000). Items that are scored polytomously differ from dichotomously scored items in that the polytomously scored items have more than two possible score categories. While dichotomous items are scored as either correct or incorrect, polytomous items typically are scored by selecting one response from a set of ordered response categories. For example, a constructed response item may be scored using a 5-point rating scale (Penfield & Lam, 2000).

In the educational realm, constructed response items are considered to be suitable forms of assessment for assessing higher-order thinking and/or problem-solving skills (Muraki, Hombo & Lee, 2000). One advantage of constructed response items is that they are thought to reflect the examinee's ability to problem solve, reason and integrate information; rather than providing only isolated bits of knowledge (Muraki, Hombo & Lee, 2000).

Regardless of whether the setting is educational or clinical, a unifying characteristic of all constructed response items is that a rater is involved in generating the item score. In the case of dichotomously scored SR items, an item score is generated by the examinee, without interpretation and judgment by a rater. The reliance on raters to generate item scores in constructed response items is the primary disadvantage of this item type. A commonly cited disadvantage of constructed response items is that the reliability of a test composed solely of constructed response items, is notoriously low compared to a test composed of multiple choice items. Mehrens (1992) noted the following reasons for the lower reliability of tests consisting solely of constructed response items: 1) these tests typically have fewer items than a multiple choice test, 2) scoring of constructed response items tend to be subjective based upon raters' personal judgment, and 3) tests consisting of constructed response items have lower internal consistency compared to tests composed of multiple choice items. As noted in previous paragraphs another characteristic of items scored by raters, is that these items are usually polytomous, as opposed to dichotomous, in nature. The polytomous nature of rater scored items has also been implicated as a reason for lower reliability (Penfield & Lam, 2000). So, while many are in agreement that constructed response items offer additional information beyond that of a dichotomously scored multiple choice item, there is also agreement that inter-rater reliability poses a potential internal threat to the reliability of scores earned on a test composed solely of constructed response items.

Given the issue of inter-rater reliability, assessment tools comprised of constructed response items are often accompanied by a measure of inter-rater reliability, to provide evidence regarding the reliability of scores from the assessment tool. Traditional measures of inter-rater reliability, based in classical test theory, are often obtained subsequent to rater training, to promote improved rater agreement. However, this degree of training, to ensure agreement, seldom takes place in informal educational settings or in clinical settings, where time is often at a premium. Moreover, the calculation of the traditional measures of inter-rater reliability typically utilize only a small sample of examinees, due to the fact that all raters must provide scores for all examinees on all items (i.e., a fully crossed design). Despite the rater reliability concerns associated with constructed response items, developers continue to create tests with polytomously scored items due to their value in measuring a breadth of constructs (Welch, 2006; Lane & Stone 2006).

The purpose of this research was to investigate an alternative measure of inter-rater reliability, based in modern test theory that does not require a fully crossed design for its calculation. This makes it feasible for the proposed inter-rater measure to be used in a more authentic setting, as opposed to a research setting. That is, one rater can supply his or her scores that were assigned to a sample of examinees while another rater can supply his or her scores that were assigned to another independent sample of examinees. This design, involving examinees nested within raters, rather than fully crossed with raters, is more achievable in clinical and informal educational settings. It was proposed that this inter-rater reliability measure, rooted in modern test theory, may provide a more authentic measure of inter-rater scoring differences associated with constructed response items and the assessment tools made up largely of constructed response items.

In addition to offering a more authentic and achievable index of inter-rater reliability, the investigated measure was also proposed to provide a more precise and informative assessment of inter- rater scoring differences. The current investigation compared and contrasted traditional inter-rater measures with this proposed novel analysis of inter-rater scoring differences rooted in item response theory (IRT). Specifically, techniques that are typically used to evaluate differential item functioning (DIF) were used to examine scoring variability due to raters. Using DIF analyses to quantify inter-rater scoring variability resulted in a DIF statistic that can be interpreted in terms of inter-rater reliability. Walker (2011) notes that typically, the presence of DIF implies that a scale is not measuring the same thing for all respondents. This statement refers to DIF when DIF is conceptualized as the presence of a different dimension other than the trait of interest. However, if the source of DIF is from raters, rather than some additional dimension inherent to the test-taker, the presence of DIF still implies, as Walker (2011) noted, that a scale is not measuring the same thing for all respondents. Therefore, this proposal posits that the presence of DIF, as an indicator of inter-rater scoring differences, also represents a threat to the validity of the test scores.

DIF occurs when examinees, from two different population groups, with the same magnitude of the underlying trait measured by the test, perform unequally on an item. In DIF analyses, it is customary to refer to two different groups as the reference group and the focal group. Although DIF analyses are not typically used to evaluate inter-rater reliability or differences in inter-rater scoring, this can be accomplished by defining the reference group and the focal group as examinees nested within Rater 1 and Rater 2, respectively. If a DIF analysis is conducted using these groupings then item-level information pertaining to the disagreement of raters can be identified and quantified. This item-level information regarding rater agreement can provide information relevant to improving items, developing rubrics, as well as training and monitoring raters.

One traditional measure of inter-rater reliability that was utilized for comparison to the newly proposed measure is the intraclass correlation coefficient (ICC). Although a variety of ICCs exist, all of which will be discussed in greater detail in the review of the literature, the ICC chosen for this proposal will allow generalization to a larger population of raters. Moreover, the ICC chosen for the purposes of this investigation provided a measure of inter-rater reliability that expresses the degree of agreement between the raters at the level of the individual rater, rather than averaged over all raters (Shrout & Fleiss, 1979). In the current study, the information gleaned about rater variability from the DIF analysis and the information from the ICCs was compared and contrasted.

The second traditional measure of inter-rater reliability utilized for comparison to the newly proposed measure was Cohen's kappa (Cohen, 1960). The kappa statistic is considered an appropriate measure of reliability of raters' ratings when the scores lie on a nominal or ordinal scale (Sim & Wright, 2005). Cohen's kappa statistic will be discussed in detail in the review of the literature. And, similar to the investigation conducted with the ICC, the information gleaned from the kappa statistic regarding inter-rater reliability was compared and contrasted to the information from the DIF analysis.

The primary purpose of this study was to determine if a DIF analysis offers advantages over traditional measures of inter-rater reliability, in terms of identifying and quantifying inter-rater scoring differences in constructed response items. Scoring differences of polytomous items by raters, that is poor inter-rater agreement, leads to larger variability due to raters and lower inter-rater reliability.

This study used both simulated and empirical data to investigate the use of DIF analyses as a measure of inter-rater reliability in constructed response type items. The research questions addressed in this study include:

1) Can DIF analyses be utilized to detect, analyze and quantify inter-rater reliability in constructed response items across different degrees of rater variability and sample sizes?

2) When using DIF detection methods to assess inter-rater reliability, is more precise item-level information obtained for constructed response items, when compared to the information offered by an intraclass correlation coefficient (ICC)?

3) When using DIF detection methods to assess inter-rater reliability, is more precise item-level information obtained for constructed response items, when compared to the information offered by a kappa coefficient?

To answer these research questions, polytomous data were simulated under a variety of conditions. The conditions differed based upon the amount of inter-rater differences present in the scoring as well as sample size. Specifically, the simulated data were generated to exhibit: 1) minimal, 2) moderate and 3) severe inter-rater scoring differences in a small, a medium, and a larger sample size. Inter-rater scoring differences in each of these sets of simulated, polytomous data were described using an ICC as a means of partitioning and quantifying the variance due to raters. Cohen's kappa was also used as a means to quantify rater agreement. Finally, two DIF analyses were used to describe the variance attributable to rater scoring differences in these nine sets of simulated data.

Additionally, contrasts and comparisons of the inter-rater agreement indices were performed using empirical, constructed response, polytomous data. As is often the case when using real data, as opposed to simulated data, once participants with missing data were removed the resultant data set was small, N = 177. This small, but realistic, sample size had to be considered when the DIF analyses were utilized to assess inter-rater reliability. Therefore, the nature and extent of these limitations were examined using the empirical data, as well as larger simulated data sets.

To summarize, the intent of this study was to examine inter-rater reliability using both classical and modern test theory indices. It was hypothesized that analyses typically used to detect DIF can be utilized as a measure of inter-rater reliability for constructed response items. Further, the DIF analyses were hypothesized to provide more precise measures of inter-rater reliability than traditional measures. Ultimately, the increased precision of identifying inter-rater scoring variability at the item level, through the use of analyses typically used to detect DIF, will assist in item development, rater consistency, and score validity.

The next section of this paper will review selections from the abundant literature regarding the estimation of inter-rater reliability, within both traditional and modern test theories. The review of the literature will also offer an overview of DIF, because the primary purpose of this study is to investigate the use of DIF analyses to evaluate inter-rater reliability.

Review of the Literature

Three major topic areas will be reviewed in this chapter. The first section of the literature review will explore measures of inter-rater reliability that are derived from classical test theory. Inter-rater reliability measures derived from classical test theory are often reported for assessment tools comprised largely of constructed response items. The assumptions, computations, and limitations of the various traditional inter-rater reliability measures will be reviewed. Subsequent to that, because this study explores inter-rater reliability analyses rooted in modern test theory, models rooted in modern test theory that incorporate a rater parameter will be explored. Finally, the topic of DIF will be reviewed. Although the topic of DIF may seem incongruent with the topic of rater reliability, the proposed method of evaluating inter-rater reliability utilizes DIF detection techniques. Therefore familiarity with the tenets of DIF will assist the reader in understanding the proposed use of DIF for inter-rater reliability assessment.

Inter-rater Reliability Measures in Classical Test Theory

In classical test theory, when the data are considered to be measured at least at an interval level, bivariate correlation coefficients are often used to estimate inter-rater reliability between a pair of raters. That is, scores assigned by one rater are correlated with the scores from a second rater. This results in a Pearson correlation coefficient (r_{xy}) as follows:

$$r_{xy} = \frac{cov_{XY}}{s_X s_Y} \qquad \text{Eq. 1}$$

where X values are the scores assigned by Rater 1 and Y values are the scores assigned by Rater 2. In Equation 1 *s* represents the standard deviation of the subscripted variable and cov$_{xy,}$, or the covariance of X and Y, is the degree to which the two variables, X and Y, vary together. The value of the cov$_{xy}$, can be found mathematically using:

$$cov_{XY} = \frac{\Sigma\,(X-\bar{X})(Y-\bar{Y})}{N-1} \qquad \text{Eq. 2}$$

where *N* represents the number of objects for which both an X value and Y value are recorded. In the case of inter-rater reliability *N* typically represents the number of interval level assessments that were scored by both raters.

The Pearson correlation coefficient calculated from Equation 1 can take on values only between -1 and 1 to describe the direction and the degree of the linear association between the interval level scores from two raters. Although its calculation and interpretation are fairly straightforward, the Pearson correlation coefficient does have its shortcomings. The value of r_{xy} can be influenced by outliers, non-normal distributions of the X and Y variables, non-linearity of the relationship as well as the range over which the X and Y data points are allowed to vary (Howell, 2007). Another known issue with the Pearson correlation coefficient is that it can show that two raters are highly correlated with one another when in fact the scores show little agreement. The term agreement, in the context of raters, refers to the degree to which two or more raters assign the exact same score to identical observable situations when the raters are using the same rating scale. To avoid these distortions to the Pearson correlation coefficient, or when the ratings are not considered to be measured at an interval level, Spearman's rho (ρ) is often used as an index of inter-rater reliability. Spearman's rho is a correlation between ranks. In fact, Howell (2007) recommends that Spearman's rho be calculated by simply applying Pearson's formula to the ranked data. In order to do this, interval level scores are first assigned to the examinees by the two raters. The interval level scores are then ranked from highest to lowest, within each rater individually. Figure 1 shows the interval level scores for 10 examinees assigned by two raters. Figure 2 shows these scores converted to rank order data.

	Examinee									
Rater	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Rater 1	98	36	45	58	84	76	33	59	55	63
Rater 2	89	77	62	55	57	46	47	71	68	59

Figure 1. Interval Level, Rater Assigned Scores for 10 Examinees Assigned by Two Raters.

	Examinee									
Rater	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Rater 1	1	9	8	6	2	3	10	5	7	4
Rater 2	1	2	5	8	7	10	9	3	4	6

Figure 2. Interval Level, Rater Assigned Scores Converted to Rank Order to Calculate Spearman's rho.

Once the raters' scores are converted to rank order data then the difference between the assigned rankings is calculated and used in the following equation:

$$\rho = 1 - \frac{6\,\Sigma\, d_i^2}{n(n^2-1)} \qquad \text{Eq. 3}$$

where d_i is the difference in paired ranks and n is the number of cases. In the case of tied rankings the pair of scores that are tied for a particular rank are given the rank equal to the average of the two involved rankings.

Similar to the Pearson correlation coefficient, the Spearman rank correlation coefficient can take on a value between -1 and 1 to describe the degree of association between the rank orders assigned by two raters. Whereas the Pearson correlation measures the strength and direction of the linear association between two sets of scores, Spearman's rho describes the strength and direction of the monotonic association between the rankings obtained from two raters. Theoretically speaking, Spearman's rho is less sensitive to outliers than the Pearson correlation coefficient and, because it measures a monotonic association rather than a linear association, it does not require the same assumptions as the Pearson correlation coefficient, such as that X and Y follow approximately normal distributions. Although the Spearman's reliability index requires less stringent assumptions than the Pearson correlation coefficient it does require rank order data. Therefore, data need to either be recorded as rank order data, which is not typical in educational ratings, or converted to rank order data. A possible scenario utilizing Spearman's rho entails each rater scoring multiple tests at the interval level, observing that the scores assigned do not meet the assumptions for Pearson's correlation coefficient then converting the total test scores into rank order data to evaluate inter-rater reliability. This was illustrated with a small sample size in Figures 1 and 2. Although possible, this scenario does not seem very practical in either an informal educational or clinical assessment realm.

While most discussions of traditional measures of inter-rater reliability include these indices, and Spearman's rho is cited as a non-parametric option when the assumptions of parametric indices are not met, practically speaking, rank ordering is seldom the means of scoring utilized in educational measures. Likewise, the results of clinical measures are seldom reported as a ranking. Moreover, while these two indices provide a measure of association between the scores assigned by two raters, they do not offer an index of agreement between the two raters. As was alluded to in the discussion of the Pearson correlation, although the correlation index may approach the maximum value of 1, this does not necessarily indicate a high level of inter-rater agreement. Moreover, these indices refer to reliability across an entire assessment tool. They do not refer to the amount of inter-rater association at the item level.

An index that reflects the amount of rater agreement is Cohen's kappa (κ, Cohen, 1960). Cohen's kappa is often used to evaluate inter-rater reliability (Howell, 2007) as it provides an index of rater agreement for nominal or ordinal level data in the case of two raters. Instead of dealing with rank orders, like Spearman's rho, κ is concerned with inter-rater reliability when raters are asked to place objects of measurement into nominal or ordinal level categories. Specifically, Cohen's kappa compares the degree of agreement between two raters to the degree of agreement that would happen by chance alone. That is, Cohen's kappa provides a measure of agreement which adjusts the observed proportional agreement to take into account the amount of agreement which would be expected by chance.

To calculate Cohen's kappa, a contingency table is set up to record the ratings provided from two raters, according to how many objects of measurement fall into each rating category. From the contingency table a simple proportion of agreement can be calculated by:

$$Proportion\ of\ Agreement = \frac{number\ of\ cases\ for\ which\ raters\ agreed}{total\ number\ of\ cases} \qquad \text{Eq. 4}$$

Howell (2007) points out that this measure expressed in Equation 4, does not take into account that a certain amount of agreement will take place merely by chance alone.

Therefore, Cohen's kappa adjusts for chance agreement between independent raters, using nominal or ordinal level data, by the following equation:

$$k = \frac{p - p_e}{1 - p_e} \qquad \text{Eq. 5}$$

where p equals the observed proportion of agreement and p_e equals the expected proportion of agreement. These proportions are calculated using the observed values and marginal values from the contingency table. Specifically, the proportion expected (p_e) value is calculated as in chi-square, by multiplying row total by column total and dividing by the grand total. The expected proportion represents the amount of agreement that is expected by chance in the case of statistically independent raters. The proportion observed (p) is calculated by adding the total number of cases in which the raters agreed on the classification, as gleaned from the contingency table, and dividing this total number by the grand total.

When there is perfect agreement between the raters, $p = 1$ and $\kappa = 1$. When the agreement between raters is equal to chance agreement, that is, $p = p_e$, then $\kappa = 0$. Kappa can take on values less than zero when there is an inverse relationship between the classifications. However, unlike a correlation coefficient which can range from -1 to 1, only κ values between 0 and 1 have any useful meaning.

Although widely used, kappa statistics are thought to have their own shortcomings and, therefore, some researchers advise using the kappa statistics only cautiously. Some have pointed out that, like correlation coefficients for continuous data, the value of kappa is not easy to interpret in the context of how precise a single observation is. Moreover, how large the value of kappa should be to indicate good inter-rater agreement depends largely on the intended use of the specific assessment. Another limitation, according to Uebersax (2010), pertains to the concept of chance agreement in the equation of the kappa statistics. Uebersax (2010) points out that the proportion of chance agreement, as calculated in the kappa statistic, is relevant only when the raters are statistically independent. He goes on to point out that raters, by virtue of the fact that they are rating the same object of interest, are clearly not independent. Therefore, Uebersax (2010) concludes that the proportion of chance agreement is not appropriate as a correction to the observed agreement levels.

To this point, the inter-rater measures that have been presented are limited to assessing the association or agreement of two raters. An intraclass correlation coefficient (ICC) is a reliability index, derived from classical test theory, that can be applied to measure inter-rater reliability when more than two raters are present (Shrout & Fliess, 1979). In their classic paper on ICCs Shrout and Fleiss (1979) describe six different forms of the ICC that are most appropriate for estimating inter-rater reliability. These authors indicate that, while many forms of the ICC exist, all take the same general form. This general form is that of the ratio of the variance of interest over a denominator equal to the total variance plus error.

$$ICC = \frac{variance\ due\ to\ examinees}{(variance\ due\ to\ examinees + variance\ due\ to\ raters + residual\ variance)} \qquad \text{Eq. 6}$$

As pointed out earlier, one advantage to this measure of inter-rater reliability is that an ICC can be used to calculate a reliability index for a situation with more than two raters. Shrout and Fleiss (1979) indicate that using ICCs to assess inter-rater reliability, given the ratings of multiple judges, is one example of the one-facet generalizability study and its resulting generalizability coefficient, discussed by Cronbach, Gleser, Nanda, and Rajaratnam (1972). Generalizability coefficients will be discussed later in this review.

The six different ICCs described by Shrout and Fleiss (1979) as measures of inter-rater reliability pertain to different reliability study designs. The mathematical model that corresponds to the design of the reliability study specifies the sources of variability that are estimable. Variance components are then estimated using an appropriate analysis of variance. It is the estimated variance components that are used to compute the ICC that ultimately describes inter-rater reliability. The specific ICC discussed in the following paragraphs corresponds to the fully crossed experimental design used in the current study.

The fully crossed portion of this investigation involves the scenario of two raters, randomly selected to represent the larger population of potential raters, both of whom provide scores for the same set of examinees. The examinees in this case are also assumed to be a sample from the total population of possible examinees. Both raters score all of the examinees one time. The linear model representing the sources of variance for this study design is as follows:

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij} \qquad \text{Eq. 7}$$

where x_{ij} represents the score (x) for examinee j from rater i. The value of μ is equal to the population mean and b_j is equal to the effect due to examinee j's true score compared to the population mean. When all of the raters score all of the examinees a rater effect, a_i can be calculated as well as the rater by examinee interaction effect, denoted in equation by ab_{ij}. Finally, the error term, e_{ij}, represents the random error component present in the score. In the situation where each rater provides only one score per examinee, the interaction term will be subsumed into the random error term (Shrout & Fleiss, 1979).

When the study design involves raters crossed with examinees, the expected mean squares for these components can be found from a random effects two way analysis of variance or ANOVA. The expected mean squares for this reliability study design are shown in Table 1 (adapted from Shrout & Fleiss, 1979).

Table 1.

Analysis of Variance and Mean Square Expectations for Two-Way Random Effects Model Design

Source of Variation	***df***	***MS***	***EMS***
Between examinees	*n-1*	*BMS*	$k\sigma_T^2 + \sigma_I^2 + \sigma_E^2$
Within examinee	*n (k-1)*	*WMS*	$\sigma_J^2 + \sigma_I^2 + \sigma_E^2$
Between Raters	*k-1*	*JMS*	$n\sigma_J^2 + \sigma_I^2 + \sigma_E^2$
Residual	*(n-1)(k-1)*	*EMS*	$\sigma_I^2 + \sigma_E^2$

Where *k* refers to the number of raters in the random sample, *n* refers to the number of examinees in the random sample, σ_T^2 represents the variance of b_j, σ_J^2 represents the variance of a_i, σ_I^2 represents the variance of the component of interaction and σ_E^2 represents the random error variance (Shrout & Fleiss, 1979).

This analysis partitions the within-examinee sum of squares into two portions: 1) a between-raters sum of squares and 2) a residual sum of squares. When the sum of squares is divided by the appropriate degrees of freedom, Shrout and Fleiss (1979) refer to these as JMS and EMS respectively. The between examinees mean squares is referred to as the BMS. The variance due to examinees can be obtained by subtracting EMS from BMS and dividing by the number of raters.

The ICC for this study design is then a ratio of variances:

$$ICC(2,1) = \frac{BMS-EMS}{BMS+(k-1)EMS+k(JMS-EMS)/n} \qquad \text{Eq. 8}$$

where BMS, EMS and JMS are explained in the preceding paragraph, *k* equals the number of raters and *n* equals the number of examinees.

According to Shrout and Fleiss (1979), when raters are considered a random effect, rather than a fixed effect, the variance component due to raters contributes to total variability so that the ICC becomes an index of rater agreement rather than rater consistency. In addition, when raters are considered a random effect, the ICC aids in answering the question as to whether raters are interchangeable. The alternative, that is, considering raters to be a fixed effect, according to Shrout and Fleiss (1979), results in a measure of rater consistency rather than agreement.

Therefore, ICCs as well as the Generalizability Theory (GT) derived counterparts, offer advantages over correlation coefficients. One reliability coefficient within GT is called a generalizability coefficient. One advantage of the generalizability coefficient and the ICC include the fact that agreement, or reliability, of multiple raters can be expressed, rather than limited to two raters, as is the case with Pearson's correlation coefficient and kappa. An additional advantage of ICCs and generalizability coefficients is that different study designs can be incorporated into the formulation and interpretation of the reliability coefficient (Brennan, 1992).

The current investigation also investigated a nested design, where examinees (*p*) are nested within raters (*r*) and these are crossed with all of the items (*i*). In GT the nested design can be written as *(p: r) x i*. As mentioned previously, one advantage to GT is that it takes study design into account when calculating a reliability value (Brennan, 1992). That being noted however, Brennan (1992) also pointed out that the ability of GT to account for the design of the study is maximized when an initial generalizability study is performed utilizing a fully crossed design. The estimated variance components derived from a fully crossed generalizability study allow the maximum number of design studies to be considered when examining variance components and the generalizability coefficients derived using the estimated variance components. So, while Brennan (1992) noted that any structure can be used for an initial generalizability study the initial study design will dictate which variance components can ultimately be estimated. Equation 9 below presents the formula for computing a generalizability coefficient ($E\hat{\rho}^2$) for a *(p: r) x i* design when a fully crossed generalizability design has been conducted previously, hence allowing estimation of variance components for all of the main effects:

$$E\hat{\rho}^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pr}^2}{n_r} + \sigma_{ip:r,e}^2/n_i n_r} \qquad \text{Eq. 9}$$

Similar to Equation 8, Equation 9 can be interpreted as the variance due to examinees in the numerator divided by the total variance. When the nested design described above is conducted without a preceding fully crossed design, the five variance components that can be calculated from the *(p: r) x i* design are limited to: σ^2_i, σ^2_r, $\sigma^2_{p:r}$, σ^2_{ir} and $\sigma^2_{ip:r,e}$. The variance due to persons is not one of the sources of variance that can be calculated from the nested study design. Because the nested design does not allow the variance due to persons to be calculated the result is that the generalizability coefficient in Equation 9 is not able to be computed without a preceding fully crossed design. Moreover, in the current investigation, in which the rater agreement is calculated at the item level, the variance due to items will drop out of the design leaving the item-level design to be described as *(p: r).* The result is a fully nested design in which people, or examinees, are nested within raters. Two variance components can be calculated from this fully nested design: $\sigma^2{}_r$, $\sigma^2{}_{p:r}$. Similar to the nested design crossed with the items facet, in this fully crossed design the variance due to persons cannot be calculated and ultimately the generalizability coefficient from Equation 9 can't be calculated.

Therefore, in order to calculate a generalizability coefficient for a nested study design, first a fully crossed design has to be conducted. Without the fully crossed design independent estimates of the variance due to persons and the variance due to raters are confounded because the persons facet (i.e., examinees) is nested within the raters facet (Brennan, 1992). This puts us back at the same requirement that is needed to calculate the ICC and the kappa coefficient. However, utilizing a DIF analysis allows an analysis of rater agreement despite the fact that examinees are nested within raters. A preceding, fully crossed design is not necessary for a DIF analysis.

As noted by other authors, neither the ICC, nor the generalizability coefficient, addresses individual discrete item responses or ratings (Patz, Junker, Johnson, Mariano, & Louis, 2002). These coefficients require that examinees be fully crossed with the raters in order to calculate the inter-rater coefficient. Therefore, the ICC and the GT counterparts have limited ability to quantify the relationships between raters, individual items and examinees (Patz et al., 2002).

The subsequent sections will introduce models that have been developed within modern test theory that include a rater effect component. These are statistical models rooted in modern test theory that attempt to account for the subjectivity introduced through the introduction of human raters when an assessment tool contains constructed response items.

Inter-rater Reliability Models in Modern Test Theory

Item response theory (IRT) is a statistical procedure that allows one to model and investigate examinee, item, and rater characteristics associated with performance assessments. One model within IRT is the Rasch Model (Rasch, 1980/1960; Rasch, 1961). The basic Rasch Model describes the expected performance of examinees on an item as a function of two factors: 1) the examinee's ability level and 2) the difficulty of the item. An extension to this basic Rasch Model includes the addition of other variables or facets, such as rater severity, when estimating examinees' ability levels. This class of measurement models, which model multiple variables that potentially have an impact on assessment outcomes, is known as the Many-Facet Rasch Models (MFRM) or FACETs models (Eckes, 2002). The MFRMs have been characterized as being analogous to an ANOVA-based approach to understanding the sources of measurement error (Myford & Wolfe, 2003). Utilizing the MFRM approach, and including a parameter to capture rater severity, allows one to detect group-level rater effects (i.e., a rater main effect), interactions that include the rater and another source of variability, as well as individual-level rater effects.

In the case of polytomous items, with multiple, ordered categories, the MFRM takes the following mathematical form to describe the log-odds of each transition between adjacent rating scale categories:

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = B_n - D_i - C_j - F_k \qquad \text{Eq. 10}$$

where,

P_{nijk} = the probability of examinee *n* receiving a rating *k* on item *i* by rater *j*,

P_{nijk-1}= the probability of examinee n receiving a rating k-1 on item i by rater j,

B_n = the ability level of examinee *n*,

D_i = the difficulty of trait *i*,

C_j = the severity of rater *j*,

F_k = the difficulty of category *k* relative to *k*-1.

After transforming the observed ratings to a logit scale, or log odds scale (ln = natural logarithm), the MFRM is an additive linear model describing the probability of an examinee *j* receiving a rating of *k* rather than *k-1* on trait *i* by rater *j* (Myford & Wolfe, 2003).

Myford and Wolfe (2003) explain that, when a MFRM analysis is conducted, the facets in the model are analyzed simultaneously but independently. This joint calibration on the logit scale allows the facets of rater severity, examinee ability and trait difficulty to be on a single linear scale.

Wang (2012) noted that the MFRM is an improvement over the IRT approach, because it includes the facet of rater severity. However the MFRM approach only provides information about how raters make use of response categories, not how well the raters discriminate among test takers. Therefore, the MFRM implicitly assumes that rater discrimination is the same for all raters. As a result, in situations where there is significant variation in rater discrimination, the application of the MFRM will not be appropriate, according to Wang (2012).

Moreover, Patz et al. (2002) indicated that the MFRM considers all observed item ratings, *Xijr*, (examinee *i*, item *j,* and rater *r*) as locally independent, conditioned on examinee proficiency. This conceptualization ignores the dependence between ratings of the same item, given examinee proficiency, *Θ*. This leads to distortions in the calculation of standard errors for ability estimates, as well as the other parameters in the model. This distortion was investigated by Patz (1996) and Junker and Patz (1998) in studies using standard test information function equations and reported in Patz et al. (2002). In these studies, it was determined that when the number of raters per item increases, under the MFRM, it appears that the estimates of the examinee's latent proficiency, *Θi* are infinitely precise, even though the examinee answers no more items.

Furthermore, while the MFRM can include facets other than examinee ability and item difficulty, it accounts only for item difficulty and does not consider item discrimination, because it is an extension of the Rasch model. As the reader will see in the subsequent sections introducing differential item functioning (DIF), some types of DIF are caused by variability in the item discrimination parameter. This type of DIF is not addressed in the MFRM framework. However, if DIF analyses can be used to detect differences in rater scoring and if some types of DIF pertain to discrimination at the item level, then using DIF analyses to describe inter-rater reliability allows discrimination to be considered in the investigation of rater variability. This can offer an improvement to the MFRM framework.

Patz et al. (2002) developed the Hierarchical Rater Model (HRM) for polytomous items that can address the shortcomings of the MFRM. The original HRM, developed by Patz in 1996, was expanded (Patz et al., 2002) to pertain to polytomous data scored by multiple raters. The expanded version can be used to scale examinees and items, as well as model aspects of consensus among raters and individual rater severity and consistency effects. According to HRM, the rating process is conceptualized as involving two phases. In the first phase the examinee's performance is taken to be an example of his or her ideal rating. The HRM labels this ideal rating ξ, which is at the item level. In the second phase the rater assigns a score k to the performance. If the rater is accurate then $k = \xi$. The HRM uses IRT methods to estimate the parameters of the model. HRM estimates rater shift as a function of the probability that the observed rating given by the rater is equal to the test-taker's ideal rating. In the HRM framework the rating process is modeled using a matrix of rating probabilities. The rating probabilities can reflect rater severity and rater reliability. Moreover, interaction effects, and/or dependence of ratings on various rater covariates, can also be modeled. According to Patz et al. (2002), the HRM makes corrections to how the MFRM estimates examinee proficiency. Specifically, Patz et al. (2002) indicated that the HRM uses information from multiple raters more appropriately, because the HRM takes into account marginal dependence between different ratings of the same examinee's work.

The HRM partitions the data into two stages. In the first stage examinees' ideal ratings (ξ_{ij}) are generated, which describe examinee *i'*s performance on item *j* according to the designated IRT model. Being ideal, and following the proposed model, these are unobserved per-item latent variables. Barr and Raju (2003) pointed out that these ideal ratings are analogous to the classical theory true score or the Generalizability Theory universe score. Patz et al. (2002) indicated that the presence of the ideal rating ξ_{ij} accounts for the dependence between multiple ratings of same piece of work by examinee *i*. Including this ideal rating, according to Patz et al. (2002), corrects for the negative bias found in the MFRM framework.

The second stage involves the observed ratings from one or more raters for examinee *i*'s performance on item *j*. Observed rating(s) categories assigned by the raters, according to Patz et al. (2002), may or may not be the same as the ideal rating category generated in the first stage. That is, in the second phase the rater assigns a score k to the performance. If the rater is accurate $k = \xi$ (Barr & Raju, 2003). This rating process is then modeled using a matrix of rating probabilities indicating the probability of rater r assigning the ideal rating (as derived from the chosen IRT model). HRM estimates rater characteristics, such as severity and reliability, as functions of the probability that the observed rating given by the rater is equal to the test-taker's ideal rating. Specifically, rater severity can be represented by probabilities in the matrix by location of the mode, and rater reliability can be represented in the matrix through the spread of the distribution. Estimates of the ideal rating for examinee i on item j can be viewed as consensus rating by the raters (Patz et al., 2002).

Patz et al. (2002) indicated that the sources of variability are linked through modifications to the distributions of the various sources or facets. These modifications to the distributions are meant to reflect the discrete nature of the IRT rating data. A possible description of the distributions is below (Patz et al., 2002):

Θ_i ~ i.i.d., Normally distributed (0 ,1), $i = 1\ldots, N$ Eq. 11

ξ_{ij} ~ distributed as described by an appropriate IRT model, $j = 1\ldots., J$ for each i

X_{ijr} ~ a signal detection model described in a probability matrix, $r = 1, \ldots, R$, for each i, j

Where Θ_i represents examinee proficiency modeled as being randomly sampled from the population of interest and *i.i.d.* indicates that the theta values are independent of one another and identically distributed that is, they are taken from the same distribution. ξ_{ij} represents for each examinee i, the ideal (unobservable) response on item j is distributed according to the IRT model chosen by the researcher and X_{ijr} represents the observed rating that rater r assigned to examinee i's response on item j.

The HRM offers the flexibility of selecting specific models to describe the different levels in Equation 11 above. By designating a model that will describe the rating probabilities the model can then be sensitive to each specific rater's severity, or leniency, as the observed ratings deviate from those predicted by the model. For instance, the probability that rater *r* assigns response category *k* given the ideal rating of ξ, can be made proportional to a normal density in *k* with the mean equal to $\xi + \phi$ and a standard deviation of ψ. This can be expressed by (Patz et al., 2002):

$$p_{\xi kr} = P[X_{ijr} = k|\xi_{ij} = \xi]\ porportional\ to \exp\left\{-\frac{1}{2\psi_r^2}[k-(\xi+\phi_r)]^2\right\}$$

$i = 1,\ldots, \mathrm{N}$; $j = 1, \ldots, \mathrm{J}$; $r = 1,\ldots, \mathrm{R}$. Eq. 12

Where the parameter ϕ is a measure of rater *r*'s bias and ψ is a measure of variability. From Equation 12, when $\phi_r = 0$, rater *r* is most likely to rate in the ideal rating category, that is, $k = \xi$. When $\phi < -0.5$, rater *r* is most likely to rate in some category $k < \xi$. This value of ϕ indicates rater severity in relation to the ideal category. Alternatively, when $\phi > 0.5$, rater *r* is most likely to assign some category $k > \xi$. This larger value of ϕ indicates rater leniency in relation to the ideal category.

Another parameter in the HRM, ψ provides a measure of the variability associated with rater *r*. As the value of ψ moves away from zero, this indicates more variability, and less reliability, associated with rater *r*. The most likely category to be assigned by rater *r* is $\xi + \phi_r$. A small value of rater variability, ψ_r, indicates small to zero probability that the rater will assign a category other than the most likely category. Therefore, a large value of ψ_r is associated with a larger probability of a rater assigning in any of the categories near the most likely category, $\xi + \phi_r$. Therefore, under the HRM, reliable consensus among a group of raters is said to exist when both ψ and ϕ approach the value of zero across all of the raters. The bias and variability measures associated with individual raters are values derived from item level data, according to the probabilities described by the model. The bias index offers information regarding the rater's assignment of a category, in relation to the most likely category, given the examinee's ability, or theta. The rater variability index measures a rater's consistency in assigning the most likely category or another category.

Comparing HRM to MFRM, Patz et al. (2002) noted that while both assess rater severity, the severity parameters are estimated using nonequivalent scales: the MFRM estimates rater bias as an additive shift in the adjacent rating category logits and the HRM estimates rater bias as a shift in the modal rating category used by the rater. In addition, there is a sign change: severe raters get positive bias parameters under MFRM, and negative bias parameters under the HRM. Perhaps more importantly, Patz et al. (2002) noted that the HRM approach to estimating rater severity and leniency does not suffer from the downward bias of the standard error found in MFRM. While the MFRM distorts the estimates of examinee's latent proficiency when the number of raters per item increases, HRM avoids this. By incorporating the estimation of the examinee's ideal rating ξ_{ij}, the dependence between multiple ratings of same piece of work by examinee i is accounted for in HRM. This modeling of the ideal rating corrects for the downward bias of the standard error found in the MFRM.

These are two examples of models based upon IRT that attempt to model rater behavior on items that require rater scoring. The next sections will discuss DIF. DIF analyses, although not traditionally used to examine inter-rater scoring differences, allow an item-level analysis of rater agreement without the requirement of a fully crossed design as required by the traditional inter-rater measures.

Introduction to DIF

DIF refers to item- level analyses that answer the question of whether or not an item is free from bias. In this context, item bias pertains to the fact that individuals matched on proficiency level, but belonging to different groups, have different expected responses to an item (Penfield & Gattamorta, 2009). Using IRT as a framework, DIF is defined as a difference in the conditional probabilities of answering an item correctly (in the case of dichotomous items) in two or more groups (Hildago & Lopez-Pina, 2004). Some authors differentiate statistical DIF from substantive DIF (Myers, Wolfe, Feltz & Penfield, 2006). These authors state that once an item is statistically found to have DIF and function differently for different groups, a construct other than the construct of interest is posited to be responsible for the between groups differences, after conditioning upon degree of the construct of interest. This second step of identifying a possible second construct results in substantive DIF. Other authors refer to these differences as DIF and item bias; item bias being the case when a designated, unintended second construct is identified that accounts for performance differences (Penfield & Lam, 2000) after examinees are conditioned on the construct of interest. Items that are found to be biased, and therefore unfair, are typically edited or removed from the test, so as not to subtract from the fairness of the entire test.

Historically, DIF analyses were mainly concerned with the detection of dichotomous items that exhibited DIF. This is simply due to the fact that many large scale, high stakes, assessments were comprised of primarily dichotomous items. To perform a DIF analysis, examinees are divided into two groups. The two groups would be matched, based upon ability, and then compared to one another to analyze for DIF. In such an analysis, assuming dichotomous data, an item is said to have DIF if there is a difference in the conditional probabilities of a correct response based upon group membership. Using the terminology typically associated with DIF the two groups are identified as either the *reference group* or the *focal group*. The groups may be different genders (male and female) different races (black and white) or any other identified groups that the psychometrician hypothesizes may show differences after being matched on ability. The focal group is the group hypothesized to be disadvantaged by the item. The reference group is the comparison group (de Ayala, 2009).

Often, two types of DIF are referred to: uniform DIF and non-uniform DIF. To describe uniform versus non-uniform DIF the IRT item characteristic curve or item response function (IRF) will be used. The IRF describes the probability of correctly answering a dichotomous item across the range of abilities of the latent trait of interest. Typically, in IRT, the IRF is thought to be sample invariant. DIF violates this assumption; therefore the IRF differs for the focal group and the reference group. The nature of how the IRF differs between the two groups determines if the DIF is uniform or non-uniform. Specifically, uniform DIF is present if the IRF for the focal and reference group differs uniformly across all ability levels of the latent trait of interest (Figure 3). This uniform difference is achieved when the item is equally discriminating for the focal and reference group but differs in difficulty for the two groups of examinees.

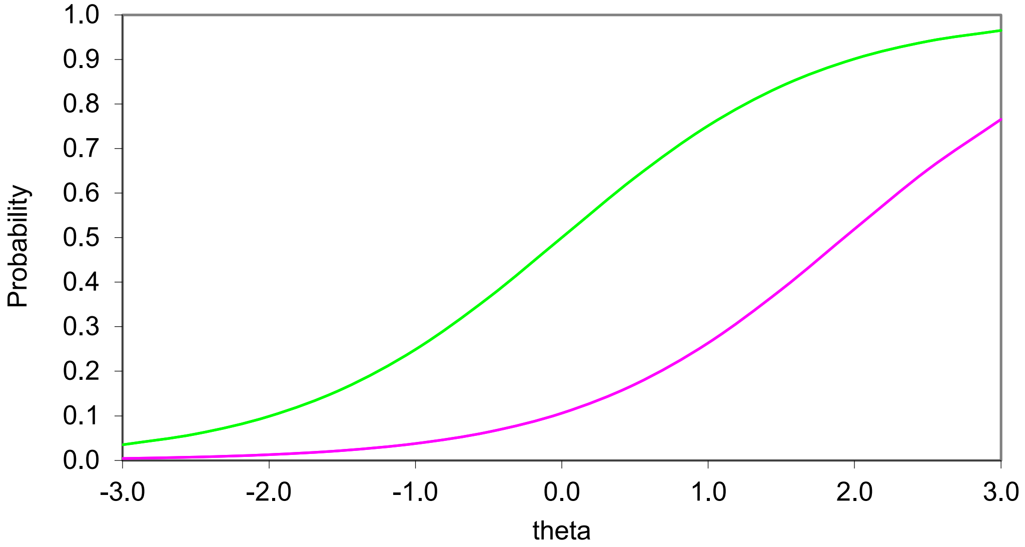

Figure 3. Item response functions for reference group and focal group when uniform DIF is present.

Alternatively, non-uniform DIF is said to be present when the difference in the IRF between the focal and reference group is not uniform across the range of all ability levels of the latent trait of interest. Non-uniform DIF can be described as one of two types: ordinal non-uniform DIF and disordinal non-uniform DIF. In both cases the difference between the focal group and reference group is not constant across all levels of ability of the latent trait of interest.

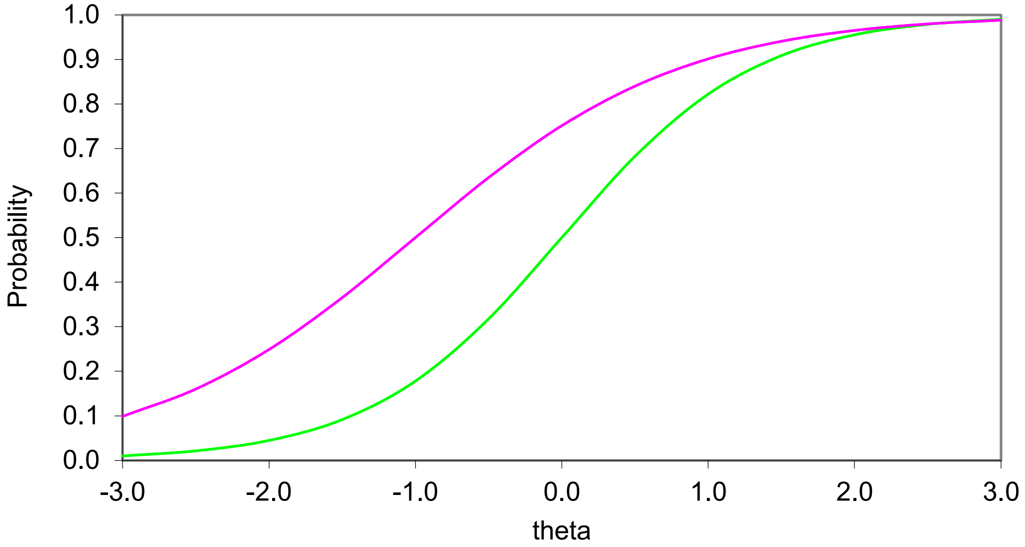

Figure 4. Item response functions for reference and focal group when ordinal non-uniform DIF is present.

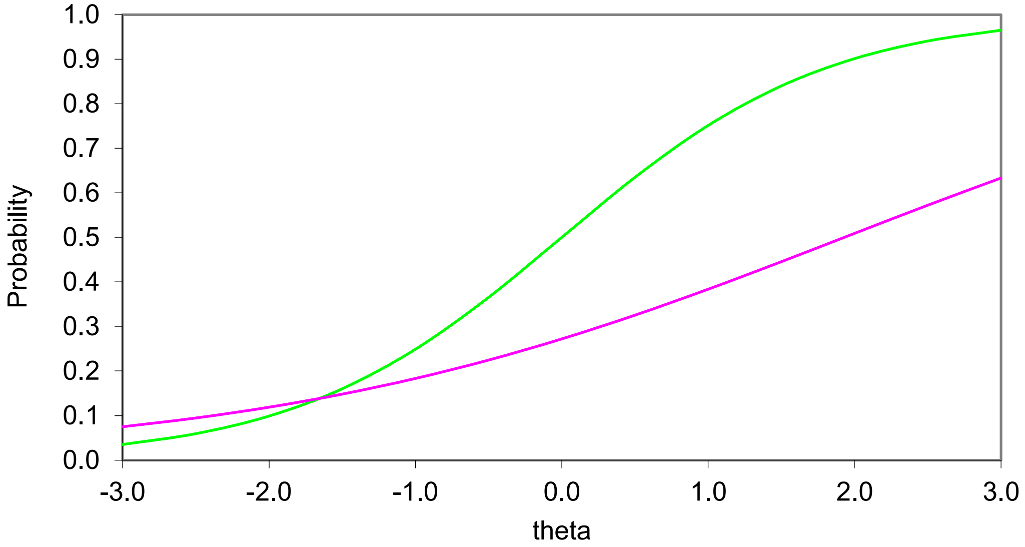

Figure 5. Item response functions for reference and focal group when disordinal non-uniform DIF is present.

Specifically, in ordinal non-uniform DIF the IRFs differ with respect to discrimination, for the focal group and the reference group, as well as difficulty (as was the case in uniform DIF). Figure 4 illustrates an example of ordinal, non-uniform DIF. The different discrimination values for the reference and focal group is how ordinal non-uniform DIF differs from uniform DIF. The item does not discriminate equally for the two groups of examinees. In the case of disordinal, non-uniform DIF the item discrimination for the focal group differs from the item discrimination for the reference group and the difficulty may or may not differ for the two groups. However, what makes it disordinal is that the two IRFs cross over the range of ability of the latent trait of interest. Figure 5 offers an illustration of disordinal, non-uniform DIF IRFs. Another way of stating this is that disordinal non-uniform DIF implies a crossing interaction between group membership and level of ability of the latent trait of interest when determining the probability of correctly answering an item (Walker, 2011). An example of this phenomenon would be that members of the reference group have a higher probability of correctly answering the item over some range of ability while members of the focal group have a higher probability of correctly answering the item over other ranges of ability.

DIF in Polytomous Data

As more tests include items with polytomous formats, methods have been developed that allow assessment of DIF in polytomous items. Tests of DIF in polytomous items pertain to whether examinees, conditioned on ability level, who belong to different groups, have the same chance of obtaining each score level of the polytomous item (Penfield & Gattamorta, 2009).

The nature of polytomous items, compared to dichotomous items, brings inherent benefits and challenges. Penfield and Lam (2000) indicated three conditions of polytomous items which are not present in dichotomous items: 1) polytomous scores are typically lower in reliability; 2) it is more challenging to define an estimate of ability to match examinees who comprise the different groups; and 3) it is more challenging to create a measure of item performance that incorporates the multiple categories of polytomous scores. They cited these three conditions as constraints to the advancement of measuring DIF in polytomous items.

Penfield and Lam (2000) explained that the low reliability of performance scores is partially due to the fact that performance assessments have fewer items than a test of ability constructed with dichotomous items. The lower number of items lends itself to lower reliability. Additionally, Penfield and Lam (2000) pointed to rater variability on scoring of the polytomous items as a factor in lower reliability. Finally, lower reliability of the performance tests can also be attributed to the fact that the ability of interest in a performance test is broader than the domain that is tested by dichotomous items which tend to address a more narrowly-defined domain. In fact, Penfield and Lam (2000) noted that, even when responses to a large number of polytomous items within a performance assessment are available, the total test score may lack internal consistency due to multidimensionality within the responses.

Despite the known limitations associated with polytomous items, developers continue to create tests with polytomously scored items (Welch, 2006). The extension of DIF to polytomous items, from dichotomous items, is typically described as more complex, but uses the same general concept, and because polytomous items persist in large scale tests, DIF analysis in polytomous items is necessary.

One type of polytomous model is the Graded Response Model (GRM, Samejima, 1969). This model is appropriate when the response categories for a polytomous item are ordered. The GRM assumes one discrimination parameter per item that applies across all of the response categories. The GRM also assumes *k* - 1 difficulty parameters per item, where *k* is the number of response categories for that particular polytomous item. For example, if a polytomous item can be scored using one of five ordered response categories the GRM will estimate *k* - 1, or four, difficulty parameters. Using these item discrimination and difficulty parameters the GRM specifies the probability that an individual of a given ability will score any particular response category or higher. The result is that it divides a polytomous item into a series of binary items. For instance, for a five category (0,1,2,3,4) polytomous item the difficulty, or b values, represent the ability level needed to have a 50% probability of scoring in that category or higher:

b_{1j} = Prob 0 vs 1234

b_{2j} = Prob 01 vs 234

b_{3j} = Prob 012 vs 34

b_{4j} = Prob 0123 vs 4

Samejima's GRM uses all of the response data to model the 50% probability that an individual with ability *θ*, scoring in a category or higher, using the following mathematical relationship:

$$P *_{X_{ij}} (\theta_i) = \frac{e^{Da_j(\theta_i - b_{x_{ij}})}}{1+e^{Da_j(\theta_i - b_{x_{ij}})}} \qquad \text{Eq. 13}$$

where, for each item *j* there is one item discrimination value, a_j; *D* is the scaling constant, θ_i is the ability level of person *i* and b_{xij} is the category boundary associated with a particular category x. As mentioned above, Equation 13 describes the probability that an individual with trait level θ_i will receive a category score, equal to *x* or higher on item *j*. A second step to determine the item characteristic curve under the Graded Response Model involves subtracting adjacent category characteristic curves. This is expressed in Equation 14 below:

$$P_{X_{ij}}(\theta_i) = P*_{X_{ij}}(\theta_i) - P*_{X_{ij}+1}(\theta_i) \qquad \text{Eq. 14}$$

Using Equation 14 to obtain the probability that an individual will respond to the lowest category *x*, within item *j*, the first category characteristic curve is subtracted from 1; and to obtain the probability that an individual will respond to the highest category, zero is subtracted from the highest characteristic curve (Dodd, Koch, & Ayala, 1989).

If a polytomous item is functioning differentially, similar to DIF in dichotomous items, this indicates that individuals of the same proficiency, but belonging to different groups, will have different expected responses to that item (Penfield & Gattamora, 2009). However, because of the multiple score categories for polytomous items the DIF can be rather complex in polytomous items.

In an attempt to bring some order to the multiple types of DIF possible in polytomous items, Penfield, Alvarez, and Lee (2009) developed a two dimensional system for categorizing polytomous DIF. The first dimension is labeled pervasiveness and addresses whether DIF is present in all possible score categories. If all score categories of the polytomous item contain DIF then the DIF is described as pervasive. Alternatively, if DIF is present in only some of the score categories of the polytomous item then it is termed non-pervasive DIF. The second dimension in their proposed taxonomy is termed consistency. This dimension pertains to the direction and the magnitude of DIF found in the score categories of all possible items. Consistency is labeled constant when the magnitude of DIF present is the same across all of the item score categories and favors the same group across all of the item score categories. When DIF favors the same group in differing amounts across the polytomous item score categories, the consistency is termed convergent. Finally, the third possible label for the consistency dimension is known as divergent and describes polytomous item DIF when DIF is present in differing amounts across the score categories in favor of different groups across the different item score categories. According to these authors, the presence of divergent polytomous DIF may be evidence for more than one cause of DIF.

Overview of Techniques Used to Detect DIF in Polytomous Data

All of the techniques for detecting DIF in polytomous items have the same aim, which is to assess between-group differences in item measurement properties after examinees are matched on ability. Not surprisingly, a myriad of techniques have been proposed to achieve this goal. Potenza and Dorans (1995) have provided one way to organize the techniques to analyze DIF in polytomous items. Their classification system involves grouping polytomous DIF detection techniques based upon two of the characteristics of the technique. The first grouping characteristic in the classification scheme is the means by which the polytomous DIF detection method generates the ability measure. The ability estimate is used to match examinees from the reference group and the focal group to ultimately determine if items display different measurement properties for examinees from different groups. The variable utilized to estimate ability is referred to as the matching variable. The matching variable can be derived by using a latent method or an observed score method. That is, polytomous DIF detection techniques can be divided into two mutually exclusive groups based upon whether the ability estimate is derived from an estimate of the latent ability, such as IRT, or a true score estimate under classical test theory, or if the ability estimate is derived from an observed score.

An example of a technique that uses a latent method of estimating ability is the generalized partial credit model (GPCM). The GPCM is an extension of the two parameter logistic model to the polytomous case (Davis, 2003). Poly-SIBTEST (Chang et al., 1996) is also an example of a technique that employs a latent method of estimating ability. In poly-SIBTEST the score on the matching criterion is the sum of scores on items that are presumed to be DIF-free (Woods, 2011). The ability estimate in poly-SIBTEST is based on true score estimation under classical test theory (Penfield & Lam, 2000). Alternatively, an example of a technique that uses observed scores to estimate ability is the Mantel Method. The Mantel Method uses a matching variable which is typically an observed sum of item scores (Woods, 2011).

Secondly, under the Potenza and Dorans (1995) classification system, polytomous DIF detection techniques can also be divided into two other mutually exclusive groups based upon whether the technique calculates item performance at each level of ability using parametric techniques or using techniques that are considered non-parametric (Penfield & Lam, 2000; Potenza & Dorans, 1995) Those techniques that are considered to be parametric use an IRT function to determine item performance at each level of ability (Penfield & Lam, 2000). The IRF that describes the expected probability of a correct response across the range of abilities is a mathematical function that uses item parameters to describe item performance. Therefore, if the IRF's for two groups differ for an item this is indication that item performance differs for the two groups and provides evidence that DIF exists using a parametric approach.

On the other hand, nonparametric methods do not use a mathematical model to determine item performance at each level of ability. Rather a nonparametric approach is used to compare the observed item performance at each ability level for the two groups. If the observed item performance at each level of ability differs for the two groups then this is evidence that DIF may be present (Penfield & Lam, 2000).

A method may be categorized as parametric or non-parametric regardless of whether or not the ability estimate (discussed above) is derived from a latent method or an observed score method. For instance, the GPCM mentioned above as an example of a latent variable method, is considered to be a parametric technique. In GPCM, differences in either item location and/or item discrimination provide evidence for the presence of DIF (Penfield & Lam, 2000). Therefore the item parameters are utilized to determine item performance making this a parametric method of polytomous DIF detection. On the other hand, poly-SIBTEST, which was also cited above as an example of a latent variable method, uses a nonparametric method to determine and compare item performance at each ability level. In Poly-SIBTEST DIF is detected by comparing differences in the observed item performance at each ability level for focal and reference groups (Penfield & Lam, 2000). Based upon Potenza and Dorans classification system, Poly-SIBTEST is a latent variable, non-parametric approach,

The different methods of DIF detection each have various strengths and weaknesses. Research comparing the different methods often draws conclusions about the methods based upon various criteria such as Type I error rate, the ability to detect uniform and nonuniform DIF, power and computational complexity (Penfield & Lam, 2000). Results from the comparative research of the DIF analyses allows the following generalizations to be made about the strengths and weaknesses of the various types of DIF detection according to which category the technique falls into. Those techniques considered to be observed score nonparametric analyses have been found to have high power and low Type I error rate when the groups involved in the analyses have equal mean abilities and the DIF is uniform. Latent score nonparametric DIF analyses, such as Poly-SIBTEST, performs similarly to the observed score nonparametric analyses with the added advantage that Poly- SIBTEST performs better when the mean abilities of the groups of interest differ. While the observed score parametric methods perform better at detecting non uniform DIF, these techniques require large sample sizes for adequate parameter estimation (Penfield & Lam, 2000).

Poly-SIBTEST.

Bolt (2002) points out that the different polytomous DIF methods, some of which were mentioned in the preceding paragraphs, use different definitions of DIF when determining the presence of DIF. The definition of DIF used in the latent trait, nonparametric Poly-SIBTEST method implies that DIF occurs when the expected item scores, conditional on the latent trait ability level of interest, differs across groups (Bolt, 2002). Because Poly-SIBTEST is non-parametric it does not use an IRT model to determine the conditional expected scores in each group. Instead, the expected scores for the studied item for the reference group and the focal group are estimated by conditioning on total scores taken from a subtest of items considered to be DIF-free. Therefore, the

DIF-free subtest scores can be used to represent the ability level of the latent trait of interest. The expected item scores can be estimated as:

$$ES_R(t) = \sum_{k=1}^{m} kP_{Rk}\,(t) \qquad \text{Eq. 15}$$

and

$$ES_F(t) = \sum_{k=1}^{m} kP_{Fk}\,(t) \qquad \text{Eq. 16}$$

where P_{Rk} and P_{Fk} are equal to the proportion of examinees in the reference group (subscript *R*) and focal group (subscript *F*) that obtain score *k* on the studied item and have subtest score *t*. In Poly-SIBTEST the expected item scores are adjusted using a regression correction procedure that corrects for measurement error in the subtest (Bolt, 2002). Specifically, Chang et al. (1996) indicated that Poly-SIBTEST adjusts the expected item scores using a linear transformation to correct for the effect of impact, which is defined as true differences between the two groups. The regression-correction, which is used to determine the expected item score, is a defining feature of Poly-SIBTEST. The slope of the regression line for each group separately is equal to the reliability of the test minus the studied item, that is, Cronbach's alpha with a guessing correction (Woods, 2011; Chang et al., 1996). This results in adjusted expected item scores, $ES*_R(t)$ and $ES*_F(t)$ which eliminate the bias due to the measurement error in the DIF-free subtest, especially when impact is present.

A DIF index is estimated using an average difference of these adjusted scores, weighted by the proportion of examinees obtaining DIF-free subtest score equal to *t*. The estimated DIF index is represented below by $\hat{\beta}_{UNI}$ where,

$$\hat{\beta}_{UNI} = \sum_{t=1}^{T} \left([ES_R^*(t) - ES_F^*(t)] \frac{N_R(t) + N_F(t)}{N} \right) \qquad \text{Eq. 17}$$

where *T* equals the maximum possible score on the DIF-free subtest, $N_R(t)$ and $N_F(t)$ are equal to the numbers of examinees obtaining DIF-free subtest score *t* from the reference group and the focal group respectively, and *N* is the total number of examinees. Bolt (2002) reports that for large samples this DIF index, $\hat{\beta}_{UNI}$, can be assumed to be approximately normally distributed when the null hypothesis of no DIF is assumed. Therefore, $\hat{\sigma}_{\hat{\beta}_{UNI}}$ can be found using the following formula:

$$\hat{\sigma}_{\hat{\beta}_{UNI}} = \sqrt{\left[\frac{N_R(t)+N_R(t)}{N}\right]^2 \left[\frac{\hat{\sigma}^2_{Rt}}{N_{Rt}} + \frac{\hat{\sigma}^2_{Ft}}{N_{Ft}}\right]} \qquad \text{Eq. 18}$$

Then a test statistic can be obtained by dividing the DIF effect measure by its standard error (Zwick, 1996; Bolt, 2002):

$$SIB = \frac{\hat{\beta}_{UNI}}{\hat{\sigma}_{\hat{\beta}_{UNI}}} \qquad \text{Eq. 19}$$

This test statistic can be evaluated against a standard normal distribution. Therefore, the null hypothesis of no DIF is rejected whenever $|SIB| > 1.96$ using a significance level of $\alpha = 0.05$ and an alternative hypothesis stating that DIF is present against either group.

In this study, Poly-SIBTEST was used to perform an exploratory single-item DIF analysis to determine if any items function differentially, which would imply that participants were rated differently by the two different raters. In this initial exploratory analysis, each item was analyzed for DIF while the remaining items were considered to comprise the matching subtest. Items exhibiting large DIF were excluded from the matching subtest. The criterion for large DIF follows the guidelines proposed in Gierl (2002); that is a polytomous item with a β value ≥ 1.00 and a corresponding $\alpha < 0.5$ is designated as functioning differently for the focal and reference group when conditioned on examinee's ability. Remaining items that were found to have fairly comparable item characteristic curves across the two different raters comprised the matching subtest which was assumed to be DIF-free. Once a DIF-free matching subtest was identified using statistical criteria, a second exploratory single item analysis was re-run. This re-run allowed each item to be retested for DIF using the statistically-identified set of items that performs equally for the two different groups as a matching subtest. The results of the second analysis using the subtest comprised of statistically equally performing items was reviewed. Again, statistical criteria were applied to determine if an item was exhibiting DIF using the guidelines outlined above for DIF detection in polytomous items.

This literature review has offered explanations of different techniques for measuring and reporting variability that is inherent in ratings generated from rater-scored observations. Inter-rater reliability and agreement indices rooted in classical test theory were presented. Subsequent to that models derived from IRT that incorporate rater effect were presented. Finally, the review of the literature changed trajectory from the discussion of inter-rater reliability, and information about DIF and DIF in polytomous data was presented. While DIF analysis, to date, has not been used to examine scoring differences attributed to raters the following section will describe a scenario whereby DIF detection techniques were used to measure and examine rater differences.

Methods

Both simulated and empirical data were utilized in order to examine measures of inter-rater reliability. Initially, a description of the simulated data will be presented. Subsequent to that, a description of participant and item information comprising the empirical dataset will be described. Finally, the analyses applied to the simulated and empirical data will be described.

Simulated Data Description

SAS was used to simulate item response data sets using the Graded Response Model (Samejima, 1969), with five response categories for each item. The Appendix displays the item parameters used for the reference group (i.e., Rater 1) as well as for the focal group (i.e., Rater 2) under the four different conditions of rater scoring differences. The reference group item parameters were taken from the Patient Reported Outcomes Measurement Information System (PROMIS) Sleep Disturbance Item Bank (Buyse, Yu, Moeul, et al., 2010). The item parameters for the focal group under the four rater scoring conditions will be discussed in greater detail in the subsequent section.

Factors manipulated.

Two factors were manipulated in the simulated portion of the study. These two manipulated factors were: 1) degree of rater scoring differences and 2) sample size. Both of these factors are described in greater detail below.

Rater scoring differences.

Examinees' item scores were simulated under four conditions meant to represent four different levels of rater scoring differences (none, minimal, moderate, and severe). When the amount of rater scoring differences were minimal, moderate or severe, the scoring variability was modeled by systematically adjusting the *b* (difficulty) parameters of the response categories in the items chosen to exhibit poor rater agreement. The items exhibiting poor rater agreement, also referred to as rater scoring differences, were modified by changing the *b* parameters of the focal group. By altering the *b* parameters for each possible score category in all of the items chosen to exhibit differences in rater scoring, the result was a simulation of differences in rater severity and greater variability across raters' scores for these items. Therefore, the simulation of differences in rater severity resulted in greater scoring differences among the raters and ultimately less inter-rater reliability on the items selected to model poor rater agreement. The next paragraphs will describe, in detail, how the simulated rater scoring differences were modeled in the selected items.

Constant, pervasive rater scoring differences (Penfield, Alvarez & Lee, 2009) were simulated to mimic 1) a minimal difference in item scoring between raters in the items displaying rater scoring differences, 2) a moderate difference in item scoring between raters in the items displaying rater scoring differences, and 3) a severe difference in item scoring between raters in the items displaying rater scoring differences. Specifically, the *b* parameters associated with minimal rater scoring differences were made higher by a value of 0.1 across all score categories within the item for the focal group, compared to the reference group. To simulate moderate rater scoring differences the focal group's *b* parameters of all the possible score categories were 0.3 greater than the *b* values of the reference group. The level of severe rater scoring differences, was simulated by utilizing *b* parameters 0.6 greater for the focal group across all possible score categories for the items displaying differences in rater severity (Penfield, Gattamora

& Childs, 2009). The level of none, in the condition of rater scoring differences, simulated no rater severity and therefore no value was added to the *b* parameters. Instead, for this level of rater scoring differences, the reference group item parameters were utilized to generate both the reference group responses as well as the focal group responses. This level of no simulated rater variability allowed comparison to a condition without rater scoring differences, as described in the *Analyses* section below.

The reference and focal group item parameters under the none, minimal, moderate and severe rater differences conditions are listed in the Appendix. Because the item difficulty values are larger for the focal group when a value was added to the *b* parameters in all items displaying differences in rater scoring, the reference group scored by Rater 1, is more likely to receive higher item scores than the focal group scored by Rater 2, when conditioned on ability level. This rationale applied to the simulation of minimal, moderate and severe levels of the rater scoring differences condition.

The items selected to display rater scoring differences were chosen because they offered a range of item discrimination values. This range of discrimination values allowed examination of inter-rater reliability with differing item discrimination levels. Moreover, the number of items displaying rater scoring differences was selected in order to leave a relatively large number of items available in the matching subtest which is free of items that have rater scoring differences.

Sample size.

Three different sample sizes were considered. In the first condition responses from 1000 examinees were generated for each group. This means that traditional inter-rater reliability measures (described below in greater detail) were calculated using two scores per examinee for 1000 examinees. That is, score data from Rater 1 was simulated for 1000 examinees on all items using the reference group item parameters. Likewise, score data from Rater 2 for the same 1000 examinees using the focal group item parameters was simulated. Two different DIF analyses were performed on each sample and the different DIF analyses will be described in greater detail in the Analysis section. Briefly, for the first DIF analysis the simulated data from Rater 1 (representing the reference group) and the simulated data for Rater 2 (representing the focal group) were treated as two ratings for the same group of 1000 examinees. For the second DIF analysis the examinees were nested within raters so the examinees who receive scores from Rater 1 were treated as different examinees from those who receive scores from Rater 2. Therefore, half of the simulated 1000 examinees were randomly assigned to be nested within (i.e., rated by) Rater 1 and the other half of the simulated examinees were nested within (i.e., rated by) Rater 2. Using a sample size of 1000 examinees means that, for the second DIF analysis, the reference group had 500 examinees and the focal group had 500 examinees. This sample size constituted the larger group size (Woods, 2010).

The second sample size used in the simulation was 500. Again, data were simulated to mimic 500 examinees who were rated on the same items by two different raters. Therefore, for the traditional inter-rater reliability measures and for the first DIF analysis, two scores for each of the 500 examinees were utilized in the computation of the inter-rater reliability statistic. The resultant group sizes for the second DIF analysis, when examinees are nested within the two different raters, were a reference group and focal group each consisting of 250 examinees (Woods, 2010).

Finally, a third, a small sample size of 200 was used. This size mimicked the sample size of the empirical data set. Moreover, this very small sample size is potentially the most easily achieved sample size in an applied setting. The analysis involved 200 examinees for the traditional measures of inter-rater reliability as well as for the group sizes in the fully-crossed DIF analysis. However, this small sample size resulted in 100 examinees each for the reference and focal groups in the DIF analysis that utilized an examinees nested within raters design.

Factors that remained constant.

The test length was fixed at twenty seven items for all conditions, with five of those items exhibiting minimal, moderate or severe inter-rater scoring differences. Ability distributions for the focal group (i.e., the group of examinees whose scores were assigned by Rater 2) and reference group (i.e., the group of examinees whose scores were assigned by Rater 1) both followed the standard normal distribution, N (0, 1). No impact, or difference in group mean ability, was present in the simulated data.

Therefore, the summary of the simulated data can be described as three sample sizes and four degrees of inter-rater scoring differences to produce twelve conditions of data sets. Each condition was replicated 100 times.

Empirical Data Description

The empirical data are a subset of items, examinees, and raters who were involved in a reliability and validity study of constructed response items meant to mimic the items that make up Washington's State educational test, the *Washington Assessment of Student Learning* or WASL. The constructed response items were developed by Pacific Education Institute (PEI) as a means to align environmental education goals with the state curriculum standards and the state tests. Each component of the empirical data will be described in detail below.

Participants.

Participants included both examinees and raters. The examinees consisted of a subset of eighth graders and fifth graders, from the state of Washington, as well as the raters who participated in the March 2003 administration of a reliability and validity study of classroom-based constructed response tasks which pertain to environmental issues. The larger pool of examinees and raters who participated in the original validity and reliability study will be described initially. Following that, the criteria that were used to create the subsets of examinees and raters in the current investigation will be explained and the resultant subsets of examinees and raters will be presented.

Examinees.

In the original reliability and validity study the examinees came from thirteen elementary schools and nine middle schools from throughout the state of Washington. All of the participants were in grade 8 or grade 5 and were students in the state of Washington public school system in March, 2003, when the data were collected. According to the technical report summarizing the results," Students represented rural, suburban and urban populations in Washington State, as well as geographically diverse regions of the state." (Taylor, Kurtz Smith, Tudor, Ferguson, & Bartosh, 2003, p. 7). The original number of students who completed the constructed response tasks differed for each task but on average, 148 fifth graders completed the performance tasks and an average of 187 eighth graders also completed the performance tasks.

For the purposes of this study, a subset of participants was created from this original sample, with the goal of obtaining the largest possible sample size of examinees that had been scored by the same two raters with no missing data. Therefore the following requirements were applied to create the subsets of participants: 1) all examinees comprising the subset had to have completed the same items and 2) each item completed needed to be graded by the same two raters. The result is a fully crossed sample with no missing data. Maintaining the same two raters across all of the items and examinees and eliminating examinees and raters with missing data, resulted in 177 examinees that responded to 17 constructed response items, all of which were rated by two raters, Rater 1 and Rater 2. The participants who comprise the examinees in the current study were all in 8th grade at the time of the original reliability and validity study described above.

Raters.

The raters went through scoring training in hopes of achieving inter-rater consistency for the scoring of items for the original reliability and validity study. The rater training is mentioned as it describes an important property of the rater-participants. The following description of rater training is quoted from the PEI technical report (Taylor et al., 2003):

> Prior to scoring training, EEAP coordinators selected student work to represent each score point on a scoring rubric. Three student responses were selected for each score point. These pre-scored responses are called ‘anchor papers’.
>
> Volunteer teachers/raters attended a scoring session. First, they completed the task they were to score. Then they reviewed the scoring rubric for the items

> within the task. They discussed the scoring rubric and examined the anchor papers. Next, the teachers practiced scoring student work that had been previously scored by expert scorers. They compared their scores with the criterion scores and discussed any discrepancies. The goal was for all teachers to consistently apply the scoring rubrics to students' responses (p. 8).

In the original reliability and validity study three raters rated each item. In the case of large disagreement between the three raters, a fourth rater also rated an item. However, for the purpose of the current investigation, only ratings from two raters were retained. As described previously, the two raters that were included in the current study were chosen based upon the criterion of attaining the largest fully crossed sample of examinees with no missing data.

Instrumentation.

The assessment tool is a set of constructed response items that were created in an attempt to measure Washington state educational standards using environmental contexts. The educational disciplines being assessed include reading, writing, mathematics, science, and social science knowledge and skills. The item format on the studied assessment tool is meant to mirror the item format used on Washington's State educational test, the *Washington Assessment of Student Learning* or WASL.

These constructed response items required the examinee to answer questions, propose solutions, graph data, identify information, or use other similar means to convey the answer. For the purpose of the current study 17 constructed response items were included for analysis. The items chosen for analysis in this investigation fit the criteria of being constructed response items with the largest number of examinees and raters with no missing data. Again, as was described previously, the goal of the current proposal was to attain the largest, fully-crossed empirical data set crossing items with examinees and raters.

Procedures.

For the original reliability and validity study, each student was assigned an identification (ID) number. The student ID number was used on the task booklets and no names were collected in association with student work or test scores. Therefore no student or school names were present in the data utilized in this investigation. The examinees were administered the constructed response items between February 15th and March 15th of 2003. Teachers administering the tasks received written instructions on how to administer the test. Directions for Administration included oral directions for completing the demographic information on the test booklets, oral directions for completing the tasks, and directions for return of the materials (Taylor et al., 2003).

Trained raters scored the items assigning scores from a minimum of zero to a maximum score of four. The raters were provided with rubrics to assist in scoring the items.

Analyses

In order to examine the stated hypotheses and to determine if an analysis typically used to identify DIF can detect, analyze and quantify rater variability in constructed-response items the statistic generated from a DIF analyses, $\beta_{Ne,}$ was compared to other item-level indices. Ultimately, four indices of rater agreement were generated for each item, in both the simulated and empirical data. The four indices describing the amount of rater agreement for each item are: ICC, Cohen's kappa, and two β statistics. Each index is explained in greater detail in subsequent paragraphs.

Intraclass correlation coefficients.

One of the traditional inter-rater agreement measures, the ICC, was calculated for each item. Rater agreement at the item level, in both the simulated and empirical data, was described using an ICC as a means of partitioning and quantifying the variance due to raters (please see Eq. 7). The ICC provides a measure of how much of the overall variability in the scores assigned on each item is due to the variability attributable to raters. Larger rater variability indicates more disagreement among raters, or put another way, less agreement among raters. Because the rater variability term is in the denominator of the ratio, larger rater variability leads to a lower overall ICC value.

The ICC values were interpreted in two ways to determine which items appeared to have large rater variability indicating low rater agreement. The first means by which the ICC was interpreted used the raw ICC value to determine if an item exhibited rater scoring disagreement. Cicchetti (1994) provides cutoffs for ICC values that pertain to "poor rater agreement." Based upon these guidelines, an item-level ICC value ≤ 0.4 was designated as an ICC value that indicated rater variability was large and therefore rater agreement was poor. The raw ICC values were scored such that a score of "1" designated that the ICC was less than or equal to 0.4 and therefore the item was designated as an item exhibiting poor rater agreement. If the raw item level ICC value was greater than 0.4 the value of "0" was assigned indicating adequate rater agreement. These 0/1 ICC scores were used in the simulation portion of the study to compare the proportion of items identified as those with poor rater agreement to the proportion identified by the other indices.

The second means of interpreting the ICC involved significance testing of the value of the item level ICC to determine if the ICC was significantly different than zero (McGraw & Wong, 1996). To achieve this, the *F* statistic was used to determine if the calculated ICC inferred a population value significantly different than zero; that is, the hypothesis that ICC = 0 was tested against the alternative hypothesis that ICC > 0. Using $\alpha = .05$, the null hypothesis that ICC = 0 was rejected when the *p* value associated with the $F_{observed}$ statistic was $\leq \alpha$. A raw ICC value of zero would be indicative of poor rater agreement and therefore if the observed ICC value was significantly different than zero that indicated that the item did not exhibit poor rater agreement. These significance tests for ICC values were conducted using SPSS version 19.

In the empirical data, raw ICC values were used to examine agreement with the other rater agreement indices and to examine the performance of the ICC in the empirical dataset. Significance testing of the ICC was also utilized to interpret the item level ICC values in the real data. These analyses will be described in greater detail in subsequent sections.

Cohen's kappa statistic.

The second traditional inter-rater agreement measure calculated for each item was Cohen's kappa statistic. This index provided a chance-adjusted statistic at the item level indicating the degree of agreement between the two raters. Similar to the ICC, in order to analyze the agreement of Cohen's kappa with the other rater agreement indices Cohen's kappa was interpreted in two different ways. The first interpretation involved comparing the calculated item level value of Cohen’s kappaa to cut values suggested by the literature. Interpreting the magnitude of the item level kappa incorporated the guidelines suggested by Landis and Koch (1977): 0 ≤ poor agreement; .01 to .20 = slight agreement; .21 to .40 = fair agreement; .41 to .60 = moderate agreement; .61 to .80 = substantial agreement; and .81 to 1.00 = almost perfect agreement. Therefore for this investigation, if an item level kappa value was less than or equal to .20 that item was designated as having poor rater agreement and given a score of “1” for the purposes of comparing to the other item indices; otherwise the kappa value indicated at least a “fair” level of agreement and it was assigned a score of “0”. The 0/1 value based upon the calculated kappa value was used to identify the items designated by the raw kappa value as an item with poor rater agreement, and was used to compare and contrast to the rates of the other rater agreement indices in the simulation portion of the investigation.

The second interpretation of Cohen’s kappa, like the ICC, involved using a test statistic to test the value of Cohen’s kappa against the null hypothesis. In the case of the kappa, the null hypothesis, when kappa = 0, is the amount of rater agreement expected to occur by chance. The statistical test then, at $\alpha = .05$ will test if the observed value of kappa is significantly different than zero. In such a case, a value of kappa significantly different from 0 (zero) would indicate that rater agreement is greater than that expected by chance. Therefore, an item with an observed kappa value associated with $p < .05$ was scored “0”, indicating it was not an item exhibiting poor rater agreement.

In the empirical data the raw value of Cohen’s kappa was used to compare the information provided by kappa to the other rater agreement indices. Significance testing of the observed kappa values was also utilized to interpret the item level kappa values in the empirical data. Again, the details of these comparisons will be described in the following sections.

DIF analyses.

In addition to the two traditional inter-rater reliability measures mentioned in the preceding paragraphs, DIF analyses were also used to describe variability attributable to inter-rater scoring differences in terms of Poly-SIBTEST's β statistic. The Poly-SIBTEST offers a β statistic for each item. The β statistic indicates the degree to which the expected item score, conditioned on ability, differs across groups (Bolt, 2002). Each item has two β statistics associated with it. The reasoning for two β statistics will be explained in detail below.

Two DIF analyses were conducted for each sample of examinees. The first DIF analysis used the data from all examinees to represent both the focal group and the reference group. That is, in terms of the empirical data set, the first DIF analysis considered the scores assigned by Rater 1 to all 177 examinees as the reference group. Then the same 177 examinees were also considered the focal group, however, the score utilized for the focal group was the score the examinees received from Rater 2. This differs from a typical DIF analysis as both the focal group and reference group were comprised of the same examinees; however, in this case the examinees were scored by different raters. This design of utilizing the same examinees for the reference group and the focal group ensured that no impact was present.

Likewise, for the simulated data all of the examinees first comprised the reference group using scores assigned from Rater 1 and then all examinees comprised the focal group using scores assigned from Rater 2. Again, using the same examinees in the focal and reference groups, while not typical, assured that no impact was present. The resultant beta value from this fully-crossed DIF analysis is referred to as β_{FC}, standing for fully-crossed and to discern it from the DIF statistic derived from a nested design DIF analysis which is described next.

The second DIF analysis is more akin to a typical DIF analysis. That is, in the second DIF analysis different individuals comprised the focal group and the reference group. Although the empirical data provided a score from both raters on all items for all participants, for the second DIF analysis, the total group of 177 examinees was randomly divided into two groups of participants. These two mutually exclusive groups comprised the reference group and focal group. For each group the scores from only one rater were used in the analysis. That is, the item scores for all of the examinees randomly selected to comprise the reference group were the item scores provided by Rater 1. Alternatively, the item scores for all of the examinees randomly selected to comprise the focal group were the items scores provided by Rater 2.

For example, referring to the empirical sample of 177 examinees, participants 1 through 88 comprised the reference group, and only the scores assigned from Rater 1 were used in the nested design DIF analysis. Then, for participants 89 through 177 only the scores assigned from Rater 2 were used. These two groups, each of which was nested within one rater for the nested DIF analysis portion, represented what is usually referred to in a DIF analysis as the reference group and the focal group.

Just as was described for the empirical data, the simulated data also generated random focal group and reference group samples from the larger sample that was used for the first fully-crossed DIF analysis. These second and slightly more typical focal and reference groups consisted of different examinees who were assigned item scores from different raters. That is, the examinees in the reference group used only the scores provided by Rater 1 while the examinees in the focal group used only the scores provided by Rater 2. The resultant beta value from this DIF analysis is referred to as β_{Ne}, indicating a DIF analysis design that involved examinees nested within raters.

In the case of the two beta statistics, β_{FC} and β_{Ne}, the item was again identified as an item exhibiting poor rater agreement using two criteria. First, using the guidelines typically applied in DIF analyses of polytomous items, a beta value greater than or equal to 0.1 indicated an item exhibiting poor rater agreement. Specifically, according to guidelines suggested by Gierl (2004), an item in the current study was designated as an item displaying poor rater agreement if $|\beta_{FC}|$ or $|\beta_{\mathrm{Ne}}| \geq 0.10$. Using this criterion each item was assigned either a score of "1" to indicate that beta was greater than or equal to 0.10 and therefore exhibited poor rater agreement, or if the beta value was less than 0.10 the item was scored "0". This criterion, based solely upon the beta value, was applied to the results from both the nested and fully crossed analyses, β_{Ne} and β_{FC}. The second analysis of the beta statistics, like that of the ICC and kappa, involved hypothesis testing to determine if the value of beta differed significantly from zero at $\alpha = 0.05$. Therefore, when the probability associated with the observed beta values was less than 0.05 (i.e., when $p < 0.05$) the item was deemed an item exhibiting poor rater agreement. Similar to the 0/1 scores for the ICC and Cohen's kappa, the beta scores that identified the items with poor rater agreement were then used in the simulation portion of the study to compare the performance of the beta statistics to the other indices.

Ultimately, each item had an associated ICC value, Cohen's kappa statistic value, and two beta statistics. The first beta statistic (the fully-crossed version) described the degree of poor rater agreement present in each item when all examinees were included in both the focal group and the reference group. The second beta statistic (the nested version) represented the degree of poor rater agreement present in each item when examinees were nested within raters and, therefore, different examinees comprised the focal group and the reference group.

The next paragraphs will describe how the different indices were compared and contrasted to one another. Empirical data indices had to be examined differently than the indices that resulted from the simulated data.

Comparison of simulated data indices.

In order to answer the research questions pertaining to the ability of a DIF analysis to detect and quantify items exhibiting poor rater agreement the inter-rater reliability indices were compared based upon two properties. These two properties can be thought of as 1) False Positive Rates or Type I error and 2) True Positive Rates or power. The calculation of the Type I error and the False Positive Rate, or how often each particular index misidentified an item as an item exhibiting poor rater agreement when in fact the item had not been modeled to exhibit poor rater agreement is described in the following section. Subsequent to that, the power rate and the True Positive Rate, or how well each index correctly identified items that had been intentionally modeled to exhibit different degrees of poor rater agreement is described.

Type I error analysis and false positive rate.

Type I error and False Positive Rate both refer to the incorrect designation of having detected poor rater agreement in an item when in fact the item was not modeled to exhibit poor rater agreement. Type I error and False Positive Rate were computed using two different criteria for each of the four studied indices. First, as described above, each index was interpreted according to a cut-score value provided by the literature. In order to avoid redundancy, misidentification of an item as exhibiting poor rater agreement using the cut-score criterion is termed "False Positive Rate" in the reported results. Secondly, each index was interpreted using hypothesis testing at $\alpha = 0.05$ to determine if the index was significantly different than zero. This interpretation of the indices is referred to by the term "Type I Error" in the reported results to differentiate it from the rate determined using the cut-score criterion. In the case of the ICC and kappa values, a zero value is indicative of poor rater agreement (or agreement expected by chance alone for the kappa) and therefore, significantly different than zero would infer rater agreement.

Type I error and False Positive Rate were calculated for all four rater agreement indices across all three sample sizes for five items when the level of simulated rater disagreement was "none." Therefore, in this level of "none" when the five items were not modeled to exhibit poor rater agreement, the Type I error rate and the False Positive Rate are both defined as the percentage of times the item was incorrectly designated by a particular agreement index as being an item that displays poor rater agreement out of the 100 replications within each sample size. For the other levels of rater disagreement (minimal, moderate, and severe), the concepts of Type I error and False Positive Rate didn't apply to the five items, because of course the five items did in fact exhibit some degree of rater disagreement in those levels of rater disagreement.

Power analysis and true positive rate.

In this study power rates and True Positive Rates were determined to be the rate at which an inter-rater agreement index correctly identified an item that was intentionally modeled as an item displaying poor rater agreement. As described in the preceding section regarding Type I error and False Positive Rates, 0/1 scores derived from both the cut-score method as well those derived through hypothesis testing of the observed index values were used to calculate the True Positive Rate and the Power. Specifically, the True Positive Rate used the literature-provided cut value while power was determined using the results of hypothesis testing of the item-level index value. Power and True Positive Rate were calculated and reported for all four inter-rater agreement indices across all nine conditions for each item modeled to exhibit poor rater agreement. The level of "none" in which no rater disagreement was modeled was obviously not helpful in determining how often an item was correctly identified as exhibiting poor rater agreement. For each of the five items intentionally modeled with poor rater agreement, the power and the True Positive Rate were both defined as the percentage of times the item was correctly designated by a particular agreement index as an item with poor rater agreement out of the 100 replications within each specific condition. The only difference between the power calculation and the True Positive Rate calculation is the criterion used to determine if the item was identified as an item exhibiting poor rater agreement.

Comparison of empirical data indices.

Although empirical data provide real data results, unlike simulated data the "truth" isn't known. So while power, True Positive Rate, Type I error, and False Positive Rate analyses are possible ways to examine the indices derived from simulated data, empirical data does not allow the computation of these. Instead, the four indices for each of the 17 empirical items were compared by examining the amount of agreement between the indices. Agreement between the indices was determined by first calculating the four indices as described above. Then the raw value of each index was again considered using two different criteria: 1) in relation to the recommended cut values for that particular index and 2) utilizing hypothesis testing to interpret the value of the index. These two criteria were applied to determine if the value of the index designated the item as an item that displayed poor rater agreement, that is, differences in rater scoring,

The patterns of agreement between the indices as well as the raw item response frequencies were examined in order to draw conclusions. Findings from the simulation portion of the study regarding Type I error rate, Rate of False Positive and power and Rate of True Positives were taken into consideration as the raw values of the indices were interpreted for the empirical data.

Results

This study posed three initial research questions. The first question asked whether DIF analyses can be utilized to detect, analyze and quantify rater variability in constructed-response items across different degrees of rater agreement and sample sizes. For the nested β statistic (β_{Ne}) under investigation in this study the reference group and the focal group were not defined based upon properties of the examinees, as is usually the case in a DIF analysis. Rather the two groups were defined by the rater who scored the constructed response items for that group. As previously described, for the reference group Rater 1 provided all of the scores for that group of examinees and for the focal group Rater 2 provided all of the scores for that set of examinees. The second and third questions then went on to ask how well a DIF analysis could detect differences in rater scoring in comparison to more traditional inter-rater agreement indices often used in applied settings. Specifically, the β statistic from the nested design DIF analysis was compared to Cohen's kappa and an ICC.

Ultimately this investigation examined the ability of the β statistic to detect poor rater agreement and to determine how well it did that when compared to traditional indices under the conditions involving three different sample sizes and four different degrees of rater scoring differences. The simulation portion of this study generated results that provided answers to the research questions posed in Chapter 1. The simulated data allowed a measure of Type I error and False Positive Rate to determine how often the different inter-rater indices designated an item as an item displaying poor rater agreement when in fact the item was not modeled that way. The simulated data also allowed the calculation of the power rate and True Positive Rates for all of the indices to determine how accurately poor rater agreement was detected in the items that were modeled with varying degrees of rater scoring differences. The Type I error, False Positive Rate, power and True Positive Rate information were used to answer the three research questions.

Subsequent to the simulation study, an empirical data study was also conducted to apply the information gleaned from the simulation portion. Insight gained from the simulation about the performance of each of the inter-rater agreement indices was then applied to the interpretation of the indices in the context of an empirical data set. The results from the simulation portion and empirical portion of the study are presented in the following paragraphs.

Simulated Data

A total of twelve conditions were simulated. This was the product of four different levels of rater scoring differences crossed with three different sample sizes. For each unique condition 100 replications were run for a total of 1200 simulated data sets which were used to examine the viability of the β statistic to detect the modeled rater disagreement. The Type I error rate and the False Positive Rate will be presented first to examine the accuracy of the detection of rater disagreement by the different indices. The nested design statistic, β_{Ne}, will be compared to the performance of the other statistics that utilized fully crossed designs. Following that, power and the True Positive Rate will be reported to answer the question as to whether or not a β statistic from a nested DIF analysis design is able to detect poor rater agreement in constructed response items. The reader will recall that there are a total of four indices in this study: 1) Cohen's kappa; 2) an ICC; 3) β from an examinees nested within raters DIF analysis (β_{Ne}*);* and 4) β from a fully-crossed DIF analysis (β_{FC}) that used a fully crossed design, that is, all examinees in the reference group were also in the focal group. This DIF analysis that incorporated a fully crossed design, while not typical, offered a design that assured no impact due to examinee differences.

Type I error and false positive rate results.

Type I error and False Positives both occur when an item is determined by an inter-rater reliability index to be an item modeled with poor rater agreement when in fact the item was not modeled to exhibit rater scoring differences. Information about the rates of incorrect detection of poor rater agreement in the items under the "none" condition, that is when the items were not modeled with rater scoring differences, by the four different indices across all three sample sizes can be found in Table 2 and Table 3. Specifically, Type I error refers to the erroneous detection of rater severity in an item not modeled to display rater scoring differences when the significance level of the inter-rater reliability index was used as the criterion. Type I error rates for the four different inter-rater reliability indices across the five items under the "none" condition are listed in Table 2. The False Positive Rate, that is, the erroneous detection of rater severity when the item was not modeled to display rater scoring differences, based upon a cut value of each index, can be found in Table 3. Because both Type I error and False Positive Rate refer to the erroneous detection of rater scoring differences when "none" were simulated, the percentages reported in Table 2 and Table 3 indicate the percentage of times, within the 100 repetitions of that sample size under the rater variability condition of "none", that an item was incorrectly deemed an item with poor rater agreement. The degree of rater variability in the other three conditions of rater severity is not relevant in the analysis of Type I error and False Positive Rate.

The ICC index of rater agreement will be examined first. The Type I error rate for the ICC, when the significance level was used as the criterion, was consistently 0% for all five items across all three sample sizes. The interpretation of this significance level indicates that all of the ICC values were significantly different from an ICC value of zero. Similarly, when the literature recommended ICC cut score was used (Cichetti, 1994) the False Positive Rate of the ICC index, was again, consistently 0% for all five items across all three sample sizes. Regardless of whether significance level was used as a criterion to interpret the ICC or a cut value was used, the ICC did not erroneously identify any of the five items as exhibiting poor rater agreement across all sample sizes in this condition of “none” in which rater severity was not modeled.

It can also be seen from Table 2, using the significance testing criterion to determine Type I error rates, the kappa index had a Type I error rate of 0% for all items across all sample sizes. This mimics the Type I error rate of the ICC. In the case of the kappa statistic, this means all kappa values were significantly different than a kappa value expected due to chance. The False Positive Rates, which used the cut score to interpret kappa, are listed in Table 3. According to the cut score criterion, the False Positive Rates for the kappa index ranged from a minimum of 0% to a maximum of 4% across the five items and the three sample sizes. The single item for which kappa had any False Positive Rates above 0%, the nominal rate of False Positives associated with kappa decreased as the sample size increased. Similar to the ICC, it appears that the kappa interpreted using the significance criterion performs much like the kappa interpreted with the cut score criterion in that almost all items were correctly identified as not exhibiting poor rater agreement in this rater severity condition of “none”.

The fully-crossed β index (β_{FC}), in which the same examinees comprised the reference group and the focal group, had a range of Type I error rate from 2% to 12% (Table 2). In contrast to the pattern observed with the kappa and ICC, different results were found when the cut value was utilized to interpret the β_{FC}. The False Positive Rate, or the rate of misidentification by β_{FC} listed in Table 3, ranged from 0% to 25% across the five items and three sample sizes. The False Positive Rate by β_{FC} was greatly attenuated by increasing sample size, such that when the sample size was equal to 1000 for both the focal group and the reference group the False Positive Rate by β_{FC} was uniformly 0% across all five items within this large sample size. This attenuating effect of sample size was not as pronounced when the β_{FC} index was interpreted using significance level. Of note is that when the small sample size was used the maximum Type I error of β_{FC} was 12% compared to the maximum False Positive Rate by β_{FC} was equal to 25%.

Finally, the other β index, (β_{Ne}), in which the examinees in reference group and focal group were nested within the rater, had a range of Type I error rate for all five of the items across all three sample sizes from a minimum of 2% to a maximum of 10%. In large contrast to this the range of the False Positive Rate for the β_{Ne} index, that is, when the cut score was utilized, had a large range from 0% to a maximum of 60%. As was seen with the index from the fully crossed design β_{FC} , the index β_{Ne} showed large attenuation when the sample size was increased and the β_{Ne} index was interpreted using a cut score rather than a significance level. This attenuation was so pronounced that the maximum False Positive Rate of β_{Ne} was 4% when the sample size was equal to 1000. Therefore, it appears that at the smaller sample sizes, that is, 200 and 500 examinees, the β_{Ne} index performed best when it was interpreted using the significance level. However, with a large sample size, that is, 1000 examinees, the False Positive Rate that utilized the cut value of β_{Ne} index, could be counted on to exhibit attenuation of the misidentified items. Again, the reader is reminded that the sample size of 1000 refers to 500 examinees nested within each rater when referring to the β_{Ne} index compared to the fully crossed design required for the other indices. This sample size is different from all of the other examined indices. In the other three inter-rater agreement indices the sample size indicated the number of people scored twice, once by each rater.

Table 2.

Type I Error Rate Using the P-level Criterion for all Four Indices across Three Sample Size Conditions and No Rater Variability

ITEM Reference Group Parameters	Agreement Index	Sample Size		
		N = 200	N=500	N=1000
Item 1	ICC	0%	0%	0%
a = 2.80	kappa	0%	0%	0%
b1=-0.56	β_{FC}	12.00%	7.00%	4.00%
b2=0.33	β_{Ne}	9.00%	5.00%	4.00%
b3= 0.98				
b4= 1.74				
Item 3	ICC	0%	0%	0%
a = 2.51	kappa	0%	0%	0%
b1=-0.46	β_{FC}	6.00%	5.00%	7.00%
b2=0.31	β_{Ne}	9.00%	4.00%	10.00%
b3= 0.98				
b4= 1.72				

Table 2. (continued)

ITEM Reference Group Parameters	Agreement Index	Sample Size		
		N = 200	N=500	N=1000
Item 9	ICC	0%	0%	0%
a = 1.75	kappa	0%	0%	0%
b1=-0.41	β_{FC}	7.00%	6.00%	2.00%
b2=0.85	β_{Ne}	9.00%	9.00%	2.00%
b3= 1.04				
b4= 1.81				
Item 21	ICC	0%	0%	0%
a = 1.57	kappa	0%	0%	0%
b1=-1.52	β_{FC}	4.00%	7.00%	6.00%
b2=-0.35	β_{Ne}	10.00%	10.00%	2.00%
b3= 0.66				
b4= 1.92				
Item 27	ICC	0%	0%	0%
a = 1.91	kappa	0%	0%	0%
b1=-0.14	β_{FC}	9.00%	6.00%	5.00%
b2=0.68	β_{Ne}	9.00%	6.00%	7.00%
b3= 1.32				
b4= 2.07				

To summarize the Type I error investigation, although the ICC and the kappa exhibited uniformly 0% erroneous detection of items, the rate at which the β_{Ne} statistic misidentified any item was never greater than 10%, even for the small sample size of 200. Summarizing the False Positive Rate information, the ICC and the kappa had negligible rates while both β statistics had fairly high False Positive Rates at the smaller sample size. Both β statistics showed large attenuation with the larger sample sizes making the interpretation of the β statistics at the largest sample size using the cut score (i.e., False Positive Rate) slightly more favorable than the interpretation of the β statistics at the largest sample size using the significance level (i.e., Type I error).

Table 3.

False Positive Rate for all Four Indices Using Cut-Score Criterion across Three Sample Size Conditions and No Rater Variability

ITEM Reference Group Parameters	Agreement Index	Sample Size		
		N = 200	N=500	N=1000
Item 1	ICC	0%	0%	0%
a = 2.80	kappa	0%	0%	0%
b1=-0.56	β_{FC}	20.00%	1.00%	0%
b2=0.33	β_{Ne}	50.00%	6.00%	0%
b3= 0.98				
b4= 1.74				
Item 3	ICC	0%	0%	0%
a = 2.51	kappa	0%	0%	0%
b1=-0.46	β_{FC}	15.00%	3.00%	0%
b2=0.31	β_{Ne}	48.00%	13.00%	4.00%
b3= 0.98				
b4= 1.72				
Item 9	ICC	0%	0%	0%
a = 1.75	kappa	0%	0%	0%
b1=-0.41	β_{FC}	23.00%	0%	0%
b2=0.85	β_{Ne}	47.00%	14.00%	2.00%
b3= 1.04				
b4= 1.81				
Item 21	ICC	0%	0%	0%
a = 1.57	kappa	4.00%	1.00%	0%
b1=-1.52	β_{FC}	25.00%	3.00%	0%
b2=-0.35	β_{Ne}	60.00%	23.00%	1.00%
b3= 0.66				
b4= 1.92				
Item 27	ICC	0%	0%	0%
a = 1.91	kappa	0%	0%	0%
b1=-0.14	β_{FC}	21.00%	2.00%	0%
b2=0.68	β_{Ne}	44.00%	12.00%	2.00%
b3= 1.32				
b4= 2.07				

Examination of the performance of the β_{Ne} index in terms of power and True Positive Rate in the subsequent sections will provide additional information to help guide interpretation of the β_{Ne} index.

Power and true positive rate results.

Table 4 lists the power of the four different indices to detect poor rater agreement in a polytomous item across the nine different conditions in which rater severity was modeled. The power was determined using the significance level of the index to determine if an item was displaying rater severity. Table 5 lists the True Positive Rate of the four different indices across the nine different conditions. The True Positive Rate was determined using the cut values for each individual index as suggested in the literature.

Both the power and the True Positive Rate are reported for the five items that were modeled to simulate some degree of rater scoring disagreement depending upon the level of the rater differences condition. The rates reported in Table 4 and Table 5 are the percentage of times across the 100 replications within each unique condition (sample size X degree of rater differences) that the index correctly identified the item as exhibiting poor rater agreement. Again, the power refers to the rate when the significance level of the index was used as the criterion and the True Positive Rate refers to the rate when the cut value was used as the criterion.

Considering the values in Table 4, in which significance level was the criterion utilized, it is apparent that the traditional rater agreement statistics had powers of 0% across all five items modeled to display poor rater agreement across all nine conditions in which rater disagreement was modeled. Moreover the values in Table 5, in which cut values were utilized, indicate that the ICC again had a uniform 0% True Positive Rate using this criterion. The ICC did not detect poor rater agreement in any of the items in which rater scoring differences were modeled, regardless of the sample size, regardless of the degree of modeled rater disagreement and regardless of whether significance level or cut value was used to interpret the ICC.

When kappa was used to detect rater variability in the simulated data using the cut score criterion it also, similar to the ICC, generally had a 0% True Positive Rate for the five items. The exceptions to the rule were the detection of poor rater agreement by kappa for two of the five items modeled to exhibit poor rater agreement plus one True Positive Rate of 1% for Item 27, in the single condition in which severe rater variability was modeled with a sample size of 200. Even within these items in which kappa detected scoring differences, the detection rates by kappa, as measured by True Positive Rates were fairly low. Other than the general rule of 0% True Positive Rates for kappa for the majority of the items and conditions, the True Positive Rates for kappa ranged from 1% to a maximum power of 43%. The largest True Positive Rate for kappa of 43% occurred when the modeled rater scoring differences were severe and the sample size was equal to 1000 examinees. The item for which the True Positive Rates were highest according to kappa was the item with the lowest discrimination value of the five items modeled with poor rater agreement.

As can be seen in Table 4 and Table 5, the power and the True Positive Rates found for the β statistics were much higher than those found for the traditional inter-rater agreement statistics, that is, kappa and the ICC. The fully-crossed, β_{FC} statistic had the highest power as well as the highest True Positive Rates of all the agreement indices for all of the items across all nine conditions. The β_{FC} statistic detected the items modeled with severe amount of rater disagreement 100% of the time for all of the items across all of the sample sizes when both the cut score and the significance level were used as the criterion. When moderate differences in rater scoring the β_{FC} statistic performed better when the cut score was used as the criterion (Table 5) compared to when the significance level was used (Table 4). The range of the power for the β_{FC} statistic to detect poor rater agreement when the modeled rater scoring differences were minimal was from a minimum power of 19% to a maximum power of 99% across all items and sample sizes as shown in Table 4. The larger power by β_{FC} occurred with larger sample sizes while the lower power values were found with the smaller sample sizes. At moderate levels of rater scoring differences however, the minimum power of β_{FC} jumped up to 89% in the small sample size and was uniformly 100% power at moderate scoring differences between the raters in the medium and large sample size conditions. Examining the True Positive Rate by the β_{FC} statistic in Table 5, again it is found that the larger True Positive Rates occurred within the larger sample sizes and the True Positive Rate made a fairly large leap from the small sample size of 200 examinees to the next larger sample size of 500.

Table 4.

Power for Four Agreement Indices across Nine Conditions Using Significance Value as the Criterion

		N = 200			N=500			N=1000		
ITEM Reference Group Parameters		Degree of Rater Differences			Degree of Rater Differences			Degree of Rater Differences		
		Min	Mod	Severe	Min	Mod	Severe	Min	Mod	Severe
Item 1	ICC	0	0	0	0	0	0	0	0	0
a = 2.80	kappa	0	0	0	0	0	0	0	0	0
b1=-0.56	β_{FC}	38%	99%	100%	85%	100%	100%	99%	100%	100%
b2=0.33	β_{Ne}	21%	67%	96%	47%	100%	100%	88%	100%	100%
b3= 0.98										
b4= 1.74										
Item 3	ICC	0	0	0	0	0	0	0	0	0
a = 2.51	kappa	0	0	0	0	0	0	0	0	0
b1=-0.46	β_{FC}	26%	99%	100%	78%	100%	100%	98%	100%	100%
b2=0.31	β_{Ne}	16%	67%	96%	43%	100%	100%	72%	100%	100%
b3= 0.98										
b4= 1.72										
Item 9	ICC	0	0	0	0	0	0	0	0	0
a = 1.75	kappa	0	0	0	0	0	0	0	0	0
b1=-0.41	β_{FC}	19%	91%	100%	46%	100%	100%	85%	100%	100%
b2=0.85	β_{Ne}	12%	44%	85%	26%	99%	100%	54%	100%	100%
b3= 1.04										
b4= 1.81										
Item 21	ICC	0	0	0	0	0	0	0	0	0
a = 1.57	kappa	0	0	0	0	0	0	0	0	0
b1=-1.52	β_{FC}	22%	89%	100%	45%	100%	100%	85%	100%	100%
b2=-0.35	β_{Ne}	14%	48%	89%	26%	97%	100%	53%	100%	100%
b3= 0.66										
b4= 1.92										
Item 27	ICC	0	0	0	0	0	0	0	0	0
a = 1.91	kappa	0	0	0	0	0	0	0	0	0
b1=-0.14	β_{FC}	19%	89%	100%	67%	100%	100%	85%	100%	100%
b2=0.68	β_{Ne}	7%	46%	80%	28%	95%	100%	53%	100%	100%
b3= 1.32										
b4= 2.07										

The β_{Ne} statistic, which allows the nested design rather than the fully crossed design, followed the same general pattern as the β_{FC} statistic, however the power values (Table 4) weren't quite as high for the β_{Ne} statistic as for the β_{FC} statistic and generally speaking the True Positive Rates (Table 5) for the β_{Ne} statistic were smaller than those for the β_{FC} statistic. Values in Table 5 show that when the β_{Ne} statistic was interpreted using the cut score value the True Positive Rate, similar to the True Positive Rate by $\beta_{FC,}$ generally made a large leap from the condition of minimal rater disagreement to moderate rater disagreement. Even in the small sample size of N = 200 (or 100 examinees per rater for the case of β_{Ne}), moderate rater variability was detected a minimum of 80% of the time when the cut score criterion was used to interpret β_{Ne}. The β_{Ne} statistic had a True Positive Rate of either 99% or 100% in all cases of moderate and severe rater differences at the medium and large sample sizes (i.e., N = 500 and N = 1000).

Table 5.

True Positive Rates for Four Agreement Indices across Nine Conditions Using Cut-Value as the Criterion

ITEM Reference Group Parameters		N = 200			N=500			N=1000		
		Degree of Rater Differences			Degree of Rater Differences			Degree of Rater Differences		
		Min	Mod	Severe	Min	Mod	Severe	Min	Mod	Severe
Item 1	ICC	0	0	0	0	0	0	0	0	0
a = 2.80	kappa	0	0	0	0	0	0	0	0	0
b1=-0.56	β_{FC}	50%	100%	100%	58%	100%	100%	54%	100%	100%
b2=0.33	β_{Ne}	63%	92%	99%	52%	100%	100%	54%	100%	100%
b3= 0.98										
b4= 1.74										
Item 3	ICC	0	0	0	0	0	0	0	0	0
a = 2.51	kappa	0	0	0	0	0	0	0	0	0
b1=-0.46	β_{FC}	49%	100%	100%	56%	100%	100%	59%	100%	100%
b2=0.31	β_{Ne}	57%	92%	99%	57%	100%	100%	55%	100%	100%
b3= 0.98										
b4= 1.72										
Item 9	ICC	0	0	0	0	0	0	0	0	0
a = 1.75	kappa	0	1%	7%	0	0	4%	0	0	1%
b1=-0.41	β_{FC}	40%	98%	100%	40%	100%	100%	48%	100%	100%
b2=0.85	β_{Ne}	56%	82%	98%	49%	99%	100%	50%	100%	100%
b3= 1.04										
b4= 1.81										

Table 5. (continued)

ITEM Reference Group Parameters		N = 200			N=500			N=1000		
		Degree of Rater Differences			Degree of Rater Differences			Degree of Rater Differences		
		Min	Mod	Severe	Min	Mod	Severe	Min	Mod	Severe
Item 21	ICC	0	0	0	0	0	0	0	0	0
a = 1.57	kappa	2%	9%	27%	0	2%	39%	0	0	43%
b1=-1.52	β_{FC}	42%	98%	100%	38%	100%	100%	35%	100%	100%
b2=-0.35	β_{Ne}	62%	87%	100%	40%	99%	100%	46%	100%	100%
b3= 0.66										
b4= 1.92										
Item 27	ICC	0	0	0	0	0	0	0	0	0
a = 1.91	kappa	0	0	1%	0	0	0	0	0	0
b1=-0.14	β_{FC}	38%	98%	100%	56%	100%	100%	39%	100%	100%
b2=0.68	β_{Ne}	50%	80%	98%	43%	99%	100%	37%	100%	100%
b3= 1.32										
b4= 2.07										

Keeping in mind the relatively high erroneous False Positive Rates (Table 3) by the β_{Ne} statistic in the smaller sample sizes when the cut value was used it makes sense to examine the power values listed for β_{Ne} statistic in Table 4. The β_{Ne} statistic had power values ranging from 80% to 100% when rater disagreement was severe, across all sample sizes. The minimum power of 80% occurred when the sample size was smallest, that is, 200 examinees, and the power to detect severe rater disagreement increased to 100% across all items when the sample size increased to 500 examinees (again, 250 examinees per rater for the β_{Ne} index). Similarly, when rater scoring differences were moderate the range of power by the β_{Ne} index in the smallest sample size was from 44% to 67% but increased with the increase of the sample size to 500 examinees. Specifically, the power of the β_{Ne} index to detect moderate rater disagreement with a sample size of 500 examinees ranged from 95% to 100%. The three items for which the power of the β_{Ne} index was slightly less than 100% when rater scoring differences were moderate and sample size equaled 500 occurred in the three items with discrimination values less than

2.0. Finally, when the modeled rater scoring differences were severe and the sample size was small (N = 200) the range of power for the β_{Ne} was from a minimum of 80% to a maximum of 96%. The maximal power values of 96% for β_{Ne} when the sample size was 200 occurred with the items that had higher discrimination values, that is, > 2.0.

In summary, Table 4 shows the power rates for the four indices of interest. The power rates, which utilized the significance level as a criterion, correspond to the Type I error rates listed in Table 2 which also used the significance level as the criterion to interpret the inter-rater reliability indices. The strong, apparent pattern revealed that the β_{FC} and the β_{Ne} indices had much larger power values than did the kappa and the ICC. This is notable because, as mentioned, the β_{Ne} utilized a nested design to detect the items with poor rater agreement compared to the fully crossed design required for the other three indices examined in this study. Therefore, to address the first research question it appears from the information in Table 4 that DIF analyses can be used to detect rater variability in scoring constructed response items, even when the DIF analysis involves a design of nesting examinees within raters, as in the β_{Ne} index.

Empirical Data

Because the empirical data do not allow the knowledge of which items are displaying poor rater agreement it wasn’t possible to determine the proportion correct or the rate of false positives for the different item level rater agreement statistics. Instead, agreement between the four indices was examined, bearing in mind the patterns and capabilities of the indices found in the simulation portion of this study.

Table 6.

Descriptives for Rater 1 (Reference Group) and Rater 2 (Focal Group) Scores When N = 177

Rater	N	Minimum	Maximum	Mean	Standard Deviation
Rater 1 Total Score	177	3.00	46.00	24.68	9.09
Rater 2 Total Score	177	3.00	45.00	24.37	8.96

The descriptive statistics for the real data for all 177 examinees can be found in Table 6 which compares the descriptive statistics of the two raters. The reader will recall that the DIF analysis that generated the β_{Ne} statistic, the examinees were randomly split in half and each half was nested within a single rater. The descriptive statistics for those groups can be found in Table 7. In both cases, that is where N = 177 for both raters and a fully crossed design was utilized and for the case in which half of the examinees were randomly nested within Rater 1 or Rater 2 (N = 88, N = 89 respectively), the scores provided for Rater 1 and Rater 2 resulted in similar descriptive statistic values. In the case of the fully crossed design this is most likely evidence of the rater training that was mentioned in the Methods section. The intent of the rater training when the test was initially delivered was to maximize rater agreement. However, in the case of the descriptive statistics for the nested groups, that is, the reference group scored by Rater 1 and the focal group scored by Rater 2, the similar descriptive statistics also provide evidence that the examinees were randomly assigned to the two raters and that no impact was present between the groups.

Table 7.

Descriptives for Rater 1 (Reference Group) and Rater 2 (Focal Group) Scores, Nested Design

Rater	N	Minimum	Maximum	Mean	Standard Deviation
Rater 1 Total Score	89	4.00	46.00	25.07	8.83
Rater 2 Total Score	88	3.00	45.00	24.26	9.26

The values of all four indices of rater agreement at the item level can be found in Table 8 for all 17 items. The values of the indices found in Table 8 are not 0/1 values, but instead these are the raw values calculated for each rater agreement index given the empirical item scores. The β statistics have associated *p* values listed in the next column. The *p* values of the kappa statistic and the ICC are not listed in Table 8 as these *p* values were uniformly less than 0.05, that is, $p < 0.05$ for all kappa and ICC values. This is noted in Table 8 and will be examined further in subsequent paragraphs. The indices will be discussed first using the significance criterion to determine if an item exhibits poor rater agreement. Using the significance criterion to interpret the inter-rater reliability index corresponds to the power and Type I error analyses in the simulation study. Second, the indices from the empirical data will be examined using the cut-score criterion. This interpretation corresponds to the True Positive and False Positive Rates found in the simulation portion both of which also used the cut values to interpret the index values.

It was already noted that the significance values for both the kappa and the ICC were not included in Table 8. As stated in the *Note* at the bottom of Table 8, all ICC values and all kappa values were significantly different than zero ($p < .05$); indicating poor rater agreement was not exhibited in any of these items according to the significance level criterion.

A statistically significant value of kappa means that the amount of agreement present was greater than the amount of agreement due entirely to chance. Therefore, a significant kappa value indicates that an item does not exhibit poor rater agreement. Similarly, a significant value of an ICC indicates that the ICC value is significantly different from zero. A value of the ICC significantly different from zero would provide evidence of some rater agreement. Theses interpretations of significance end up being the inverse of the interpretation that is applied to the significance of the beta statistics.

Therefore, the finding that all of the kappa values and all of the ICC values for all of the empirical items were found to be significant at $\alpha = .05$ indicates, as explained above, that none of the empirical items were found to have poor rater agreement when the ICC and kappa values were interpreted using significance level. These findings are not entirely surprising given the findings regarding the Type I error (Table 2) and the power (Table 4) in the simulation portion of the study. In the simulation portion the ICC and kappa also exhibited uniform Type I error of 0% as well as uniform power of 0%.

Using the significance level to interpret the beta statistics provided more information and also mirrored the findings of the simulation portion of the study. Specifically, the β_{FC} statistic, when interpreted using significance level, indicated that Item 5 and Item 9 both exhibited poor rater agreement. The β_{Ne} statistic agreed that Item 5 had poor rater agreement but did not agree with the β_{FC} statistic that Item 9 exhibited poor rater agreement. Instead, the β_{Ne} statistic indicated that Item 8 had poor rater agreement. This finding is not surprising, given the results of the power analysis of the simulation study. The power of both β statistics were high in the simulated data, especially when the rater variability was modeled as moderate or severe. Additionally, the Type I error for both β statistics, even in the smallest sample size, did not exceed 12%.

Next, the inter-rater indices will be interpreted according to their cut-values. The raw inter-rater values listed in Table 8 that indicate the item is displaying poor rater agreement, according to the cut score criterion, are marked with a single asterisk. The reader will recall that in the case of both β statistics a value greater than or equal to 1.0 accompanied by a p value < 0.5 indicates that an item exhibits poor rater agreement (Gierl, 2007). Therefore items with β greater than or equal to 1.0 were interpreted as items with poor rater agreement. However, for both the kappa and the ICC, a lower value of the index indicates more rater variability (i.e., poor rater agreement). Therefore, the kappa and the ICC values are expected to vary inversely with the value of the β statistics. As indicated earlier, a kappa value less than or equal to 0.2 is associated with poor rater agreement (Landis & Koch, 1977) and an ICC value less than or equal to 0.4 offers evidence of poor rater agreement (Cicchetti, 1994).

Using these cut points and examining the information in Table 8 it is evident that neither the kappa nor the ICC indicated any of the items as items with poor rater agreement. In this empirical study the ICC value ranged from a low value of 0.525 to a maximum value of 0.883. None of these ICC values falls below the proposed 0.4 cut point proposed to indicate poor rater agreement (Cicchetti, 1994). The kappa coefficients ranged from a low value of 0.290, indicating "fair agreement" between the raters to a high kappa value of 0.749 indicating "substantial agreement" between the raters (Landis

& Koch, 1977). This finding of no items with poor rater agreement when using the cut score criterion for the ICC and the kappa might be concerning but for the findings in the portion of the simulation study that examined the True Positive Rates. Referring back to the simulation study results, both the ICC and the kappa had very low True Positive Rates to detect items that had been modeled to exhibit poor rater agreement (Table 4). Moreover, the low True Positive Rate that was illustrated by kappa in the simulation portion was attenuated in the smaller sample sizes. The empirical data set, like most empirical data, would be considered a small sample size in this context and therefore one would expect a low True Positive Rate by the ICC and the kappa statistics given the results from the simulation study.

Unlike the kappa and ICC, the β statistics identified quite a few items as items with poor rater agreement in the empirical data, according to the established cut score values. In the empirical data the β statistic from the fully crossed design, β_{FC}, identified eight items as items with poor rater agreement. The β statistic from a nested design, β_{Ne}, agreed with the finding of those eight items plus an additional seven more items for a total of fifteen items designated by β_{Ne} as items with poor rater agreement. This large number of items identified by β_{Ne} as items with poor rater agreement could be somewhat anticipated considering the large False Positive Rate of β_{Ne}, ranging to a maximum 60%, observed in the smallest sample size in the simulation study.

Table 8.

Inter Rater Reliability Indices for 17 Items of the Empirical Data

	Inter Rater Agreement Indices					
	Fully-crossed β		Nested β		Kappa	ICC
Item	β_{FC}	p of β_{FC}	B_{Ne}	p of β_{Ne}		
1	-0.003	0.980	-0.166*	0.345	0.356	0.679
2	0.036	0.806	0.131*	0.540	0.494	0.837
3	0.036	0.553	0.065	0.490	0.739	0.688
4	0.126*	0.093	0.163*	0.133	0.409	0.585
5	0.316*	0.000**	0.330*	0.009**	0.290	0.525
6	0.124*	0.105	0.248*	0.062	0.529	0.685
7	-0.085	0.380	-0.149*	0.303	0.740	0.880
8	-0.123*	0.229	-0.400*	0.031**	0.584	0.814
9	0.186*	0.037**	0.107*	0.423	0.562	0.754
10	0.082	0.453	0.129*	0.429	0.716	0.846
11	0.078	0.444	0.249*	0.143	0.749	0.883
12	-0.104*	0.176	-0.177*	0.135	0.439	0.579
13	-0.076	0.282	-0.038	0.740	0.554	0.696
14	-0.124*	0.374	-0.147*	0.507	0.382	0.771
15	0.057	0.412	0.261*	0.025**	0.735	0.863
16	-0.185*	0.139	-0.362*	0.049**	0.475	0.788
17	-0.054	0.484	-0.108*	0.381	0.468	0.638

Note. * Indicates values which deem that item as an item with poor rater agreement according to the cut values for that index as discussed in the Methods section. **Indicates significance values which deem that item as an item with poor rater agreement

To assist in further investigating the agreement on one item by the β statistics when they were interpreted using significance level, score frequencies of Rater 1 and Rater 2 are listed for the 17 items in Table 7. Item 5, the single item that β_{Ne} and β_{FC}

agreed upon as an item exhibiting poor rater agreement when significance level was the criterion, showed the largest disparity of scoring frequency of all 17 items. One score category within Item 5 differed by a frequency of 37 examinees, or 21% of total examinees. It is noteworthy that this item also earned the lowest value of the ICC and the kappa statistics, but not low enough to be deemed an item with poor rater agreement according to the cut values. This is another juncture at which to remind the reader that while all the indices seem to be pointing to Item 5 as an item with poor rater agreement the β_{Ne} statistic is the only index that was able to make this designation with a nested rather than fully crossed study design.

In summary, the empirical data results should be interpreted keeping in mind the findings of the simulation study. Specifically, for all four rater agreement indices it was found that small sample size had the effect of increasing Type I error as well as increasing the False Positive Rate. Moreover, in the simulation portion of the study the power and the True Positive Rate of both β statistics had their lowest values in the small sample size. The empirical portion of this study involved a sample size somewhat smaller than the small sample size used in the simulation.

Table 9.

Rater Scoring Frequencies for Items 1 – 17 of the Empirical Data

	Score Assigned				
	1	2	3	4	5
Q1 R1	25	55	59	22	16
R2	26	40	81	22	8
Q2 R1	37	19	40	43	38
R2	39	24	38	38	38
Q3 R1	74	103	0	0	0
R2	89	94	0	0	1

Table 9 (Continued)

	Score Assigned				
	1	2	3	4	5
Q4 R1	19	69	89	0	0
R2	26	73	78	0	0
Q5 R1	29	46	101	1	0
R2	46	67	64	0	0
Q6 R1	37	77	63	0	0
R2	44	74	59	0	0
Q7 R1	43	59	62	10	3
R2	37	65	65	9	1
Q8 R1	120	27	17	6	7
R2	104	36	19	11	7
Q9 R1	9	14	75	60	19
R2	9	22	82	52	12
Q10 R1	31	32	61	41	12
R2	20	48	67	32	10
Q11 R1	28	41	64	36	8
R2	28	43	67	32	7
Q12 R1	49	63	65	0	0
R2	34	76	67	0	0
Q13 R1	22	79	76	0	0
R2	18	71	88	0	0
Q14 R1	36	35	29	35	42
R2	24	30	52	41	30
Q15 R1	17	32	128	0	0
R2	18	42	117	0	0
Q16 R1	67	44	37	17	12
R2	53	39	54	21	10
Q17 R1	39	75	63	0	0
R2	40	70	67	0	0

Interpreting the findings from the empirical portion of this study with the insight offered by the simulation portion of the study, it appears that the DIF statistic derived from a nested design, β_{Ne}, when it was interpreted using significance level as the criterion, identified an item with poor rater agreement that was not detected by other more common means of measuring rater agreement.

Discussion

According to McClellan (2010), " CR (constructed response) items are popular choices in standardized educational assessment, so it is important to assure that the human scoring aspect of the tests, just as all others, is as standardized as possible" (p. 7). Identifying items with poor rater agreement was at the basis of the current investigation. This study utilized both simulated and empirical data to investigate the research questions which pertain to the ability of a DIF analysis to detect constructed response items with poor rater agreement. The advantage to using the investigated DIF analysis over current inter-rater agreement indices, is the more achievable design of examinees nested within raters versus the fully crossed design required by the other studied inter-rater agreement indices.

The first research question addressed if DIF analyses can be utilized to detect, analyze and quantify rater variability in constructed-response items across different degrees of rater variability and sample sizes. Based upon the simulation study findings, when the β_{Ne} statistic was interpreted using the cut value criterion the resulting False Positive Rates by the β_{Ne} statistic (Table 3) were fairly high and less than desirable. However, when the significance level criterion was utilized to interpret β_{Ne} the findings revealed acceptable Type I error by the β_{Ne} statistic (Table 2) as well as acceptable power values (Table 4). The power calculation also utilized the significance level criterion.

The findings from the power study of the simulation portion of this investigation indicated that the DIF analysis that utilized a nested design was able to detect rater scoring differences even in the small sample size of 200 when the scoring differences between raters was severe. This is of interest because the sample size of 200 in the context of the nested design DIF analysis requires 100 examinees per rater for a total of

200 examinees. This design is much more attainable in the applied setting than the requirement of a fully crossed design which is necessary for the calculation of the other inter-rater reliability indices that were investigated in this study. In the context of a clinical setting this implies that rater agreement data may be collected within the course of a clinician's typical practice rather than having two raters rate the same client. The practice of two clinicians rating a single client is seldom done in the clinic due to time and fiscal constraints. Instead, according to results from the power and Type I error studies of the current investigation, a clinician may be able to gather ample data for analysis by performing his or her usual evaluations and aggregating the data for analysis in relation to another clinician. The acceptable power and fairly low Type I error rates from the β_{Ne} statistic in the simulation study seem to indicate that with 100-150 examinees per rater one should feel fairly confident about detecting items that have moderate to severe rater variability.

The second research question asked if the β_{Ne} statistic was able to offer more precise information about rater agreement at the item level compared to the information gleaned from an ICC. Based upon the rationale discussed above, the significance level criterion was utilized to interpret the β_{Ne} statistic in order to answer this second question comparing the performance of the β_{Ne} statistic to the ICC. To determine which criterion to use to interpret the performance of the ICC, comparison of the ICC results that utilized the significance level to the ICC results that utilized the cut value reveals equivalent results in both interpretations, that is, the False Positive Rate (Table 3) and the Type I error (Table 2). Additionally, equivalent findings for the ICC were found for power calculation (Table 4) which utilized the significance level criterion and for the True

Positive Rate calculation (Table 5), which utilized the cut value criterion. Based upon these simulation results interpretation of the ICC was the same regardless of whether the cut score value was used or the significance value was used.

Therefore, using the Type I error and power values to answer the second question it is apparent that while the β_{Ne} statistic incorrectly identified more items as exhibiting poor rater agreement than did the ICC, however, the Type I error value by the β_{Ne} never exceeded 10% in any of the sample sizes. This observed Type I error value is acceptable and is accompanied by the much larger power afforded by the β_{Ne} over the ICC. The ICC had zero power for all of the items modeled with poor rater agreement across all levels of rater scoring differences and all sample sizes while the β_{Ne} did a respectable job detecting moderate and severe rater scoring differences even in the small and medium sample size.

Finally, the third research question asked if the β_{Ne} offered more precise item-level information for constructed response items, when compared to the information offered by a kappa coefficient. Again, the first step in answering this involved comparing the interpretations of the kappa when the cut value was used as the criterion to the interpretations of the kappa when the significance value was used as the criterion. Similar to the simulation results of the ICC, when the Type I error of the kappa (Table 2), using the significance level criterion, was compared to the False Positive Rate of the kappa (Table 3), using the cut value criterion, no difference was present in the results. Turning to the comparison of the power of the kappa (Table 4) to the True Positive Rate of the kappa (Table 5) to determine the best criterion for kappa interpretation, the kappa performed only slightly better when the cut value was utilized to calculate the True

Positive Rate. Therefore, it appears that the cut score criterion offered slightly improved findings when interpreting the kappa over the significance level criterion.

As mentioned previously, when the β_{Ne} statistic was interpreted using a cut value the False Positive Rate by β_{Ne} made it a fairly unreliable index to determine which items were exhibiting poor rater agreement especially in the small sample size. However, despite the high False Positive Rate by β_{Ne} it still offered more information about items with poor rate agreement when compared to the information gleaned from kappa. This statement in favor of β_{Ne} can be made in large part due to the extremely low True Positive Rate by kappa. Furthermore, these conclusions are made without the consideration that the burden of achieving a fully crossed design does not apply to the β_{Ne} statistic as it does to the kappa.

Limitations and Further Research

Although this investigation appears to support the idea that a DIF analysis can be used to identify constructed response items with poor rater agreement, the limitations of the current study should be respected. One factor that should be considered is the manner in which poor rater agreement was simulated in the current study. Namely, the rater scoring differences that were modeled in the simulation portion of this study followed a constant pervasive pattern (Penfield, Alvarez, & Lee, 2009). Specifically this form of rater scoring difference is modeled by adding a constant to each *b* parameter of the reference group in each item that was selected to be modeled with poor rater agreement. Arguably, other types of rater scoring differences should be investigated to determine if similar power and Type I error rates are found for the β_{Ne} when different patterns of rater severity are present.

Another limitation pertains to the percentage of items in the simulation that were modeled to exhibit poor rater agreement, and related to that, is the remaining percentage of items that were not modeled to exhibit poor rater agreement. The items that were not modeled with rater differences provide the subtest scores which are utilized by Poly-SIBTEST to represent the ability level of the latent trait of interest. The current simulation study utilized 27 total items, five of which were modeled to have poor rater agreement leaving the remaining 22 items to provide the subtest scores. Assessments that consist entirely of constructed response items may be comprised of fewer items overall compared to assessment tools consisting solely of multiple choice items. The reason for the shorter test may be due to the longer time required for an examinee to complete a constructed response item and/or the added time required for a rater to assign a score to the item compared to an assessment made up of multiple choice items. In either case it would be worthwhile for future studies to determine if the DIF analysis performs as well with shorter tests in general and with shorter subtests in specific. Again, the number of items exhibiting poor rater agreement as well as the number of items that comprise the subtest would both be variables of interest to further investigate the use of the β_{Ne} statistic as a means to identify items with poor rater reliability in constructed response items.

Finally, with the information gleaned from the current investigation, it was found that the Type I error rate and the power rater of the β_{Ne} statistic made large leaps in their values between the small sample size of 200 and the medium sample size of 500. Future studies could investigate sample sizes between these values to ascertain if there is some minimum sample size that maximizes the information gleaned from the β_{Ne} regarding items with poor rater agreement. This optimal sample size would be of interest for the applicability of this statistic.

Conclusion

This investigation examined the ability of a DIF analysis to detect items with poor rater agreement. The simulation used in this investigation involved a constant pervasive pattern of rater scoring differences. The simulation involved modeling the rater scoring differences in items with differing discrimination values and it examined detection of poor rater agreement across three different degrees of rater variability crossed with three different sample sizes. The findings indicated that the studied DIF analysis had very good power and fairly low Type I error rate when detecting items with large rater scoring variability, that is, items with poor rater agreement. This finding was most pronounced with larger degrees of rater variability and larger sample sizes. That being said, sample sizes of 500 examinees and moderate amounts of rater variability were amply detected by the DIF analysis that utilized the nested design. The implication of these findings is that because the studied inter-rater reliability analysis does not require a fully crossed design of both raters scoring all the examinees, it is a more attainable means to gather inter-rater agreement information by individuals in classroom and clinic settings. The traditional inter-rater agreement indices that were examined in this study did not exhibit as high of power raters as did the DIF statistic derived from a nested design, β_{Ne}. The DIF statistic derived from a fully crossed design, $\beta_{\mathrm{FC},}$ did rival the β_{Ne} in terms of power and Type I error, however, the β_{FC} and the traditional indices in the current investigation, all required a fully crossed design.

Further investigation of the DIF analysis using a nested design as a means to identify and quantify items with poor inter-rater agreement is needed to fully understand the limitations and appropriate generalizations that can be gleaned from the current findings. That being said, the findings from this investigation seem to indicate that it is feasible to use this more easily achievable index of inter-rater reliability for such applications as rater training, item selection and item development in assessments that utilize constructed response items.

References

Barr, M. A., & Raju, N. S. (2003). IRT-Based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods, 6,* 15-43. doi:10.1177/1094428110362107

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*(2), 113-141. doi:10.1207/S15324818AME1502_01

Brennan, R.L. (1992). Elements of generalizability theory (Revised edition). Iowa City, IA: The American College Testing Program.

Buysse, D. J., Yu, L., Moul, D. E., Germain, A., Stover, A., Dodds, N. E.,…Pilkonis, P. A. (2010). Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. *Sleep*, *33,* 781–792.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, *33,* 333-353. doi:10.1111/j.1745-3984.1996.tb00496.x

Cicchetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology, *Psychological Assessment, 6*, 284-290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46. doi:10.1177/001316446002000104

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Davis, L. L. (2003, April). *Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

Dodd, B.G., Koch, W.R., & De Ayala, R.J. (1989). Operational characteristics of adaptive testing procedures using the Graded Response Model. *Applied Psychological Measurement, 13,*129-143.

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

Gierl, M. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practices, Spring,* 3-14.

Graham, M. (2011). What is inter-rater agreement and how can designers of teacher evaluation systems maximize it? Washington DC: Center for Educator Compensation Reform, U.S. Department of Education. http://cecr.ed.gov/pdfs/Inter_Rater.pdf

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12,* 38-47. doi:10.1111/j.1745-3992.1993.tb00543.x

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64,* 903-915. doi:10.1177/0013164403261769

Howell, D.C. (2010). *Statistical methods for psychology* (8th ed.). Belmont, CA: Wadsworth.

Junker, B. W. & Patz, R. J. (1998, June). *The hierarchical rater model for rated test items*. Presented at the Annual North American Meeting of the Psychometric Society, Champaign-Urbana, IL.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953. doi:10.1177/0013164405275668

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

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.) (pp. 387-431). Westport, CT: Rowman & Littlefield.

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

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

McGraw, K.O. & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods 1*(1), 30-46.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes, *Educational Measurement: Issues and Practice 2*(1), 3-9. doi:10.1111/j.1745-3992.1992.tb00220.x

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6(13).* Retrieved from http://olam.ed.asu.edu/epaa

Muraki, E., Hombo, C. M., & Lee, Y-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24,* 325-337. doi:10.1177/01466210022031787

Myers, N. D., Wolfe, E. W., Feltz, D. L., & Penfield, R. D. (2006). Identifying differential item functioning of rating scale items with the Rasch Model: An introduction and an application. *Measurement in Physical Education and Exercise Science, 10,* 215-240. doi:10.1207/s15327841mpee1004_1

Myford, C. M. & Wolfe E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement, 4*, 386-422.

Patz, R. J. (1996*). Markov Chain Monte Carlo methods for item response theory models with applications for the National Assessment of Educational Progress* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA.

Patz, R. J., Junker, B. W., Johnson, M. S., Mariano, M., & Louis, T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341-384. doi: 10.3102/10769986027004341

Penfield, R. D., Alvarez, K., & Lee, O. (2009). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 22*, 61-78. doi:10.1080/ 08957340802558367

Penfield, R. D., Gattamora, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF on polytomous items. *Educational Measurement: Issues and Practice, 28*(1), 38-49. doi: 10.1111/ j.1745-3992.2009.01135.x

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, *19*(3), 5–15. doi:10.1111/j.1745-3992.2000.tb00033.x

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press. (Original work published 1960)

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428. doi: 10.1037/0033-2909.86.2.420.

Sim, J. & Wright, C.C. (2005). The kappa statistic in reliability studies: Use, interpretation and sample size requirements. *Physical Therapy* 85:257-268.

Taylor, C., Kurtz Smith, K., Tudor, M., Ferguson, L., & Bartosh, O. (2003). *Evidence for the validity and reliability of environment based classroom assessments as measures of the Washington State essential academic learning requirements.* Technical Report 7. Pacific Education Institute, Olympia, WA.

Uebersax, J. (2010, March 18). Kappa Coefficient: A critical appraisal. Retrieved from http://www.john-uebersax.com/stat/kappa.htm

Walker, C. M. (2011). What's the DIF? Why differential analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment, 29,* 364-376. doi:10.1177/0734282911406666

Wang, Z.G. (2012). *On the use of covariates in a latent class signal detection model, with applications to constructed response scoring* (Doctoral dissertation). Available from ProQuest Dissertations and Theses Database. (UMI No.3506363)

Welch, C. J. (2006). Item and prompt development in performance testing. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 303-327). Mahwah, NJ: Lawrence Erlbaum Associates. doi:10.1080/15305050701813433

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement, 32,* 511-526. doi:10.1177/0146621607310402

Woods, C. M. (2011). DIF testing for ordinal items with poly-SIBTEST, the Mantel and GMH tests, and IRT-LR-DIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement, 35,* 145-164. doi:10.1177/0146621610377450

APPENDIX

Item Difficulty and Discrimination Parameters for Reference Group and Focal Groups under None, Minimal, Moderate and Severe Rater Variability Conditions

	Reference Group (Rater 1) ALL Focal Group (Rater 2) :None					Focal Group (Rater 2): Minimal				Focal Group (Rater 2): Moderate				Focal Group (Rater 2): Severe			
Items	a	b1	b2	b3	b4	b1	b2	b3	b4	b1	b2	b3	b4	b1	b2	b3	b4
1	2.80	-0.56	0.33	0.98	1.74	-0.46	0.43	1.08	1.84	-0.26	0.66	1.28	2.04	0.04	0.93	1.58	2.34
2	2.09	-1.10	0.05	0.94	1.84												
3	2.51	-0.46	0.31	0.98	1.72	-0.36	0.41	1.08	1.82	-0.16	0.61	1.28	2.02	0.14	0.91	1.58	2.02
4	2.18	0.03	0.85	1.55	2.38												
5	1.19	-0.98	0.33	1.76	3.30												
6	1.64	0.21	1.13	2.02	2.96												
7	2.37	0.28	1.02	1.62	2.37												
8	1.77	0.22	1.21	1.95	2.73												

	Reference Group (Rater 1) ALL Focal Group (Rater 2) :None					Focal Group (Rater 2): Minimal				Focal Group (Rater 2): Moderate				Focal Group (Rater 2): Severe			
Items	a	b1	b2	b3	b4	b1	b2	b3	b4	b1	b2	b3	b4	b1	b2	b3	b4
9	1.75	-0.57	0.41	1.04	1.81	-0.47	0.51	1.14	1.91	-0.27	0.71	1.34	2.11	0.03	1.01	1.64	2.41
10	1.40	0.67	1.56	2.26	3.13												
11	1.52	-0.19	0.95	1.76	2.72												
12	2.47	0.03	0.66	1.19	1.88												
13	1.99	-0.02	0.89	1.61	2.22												
14	1.85	-0.58	0.66	1.37	2.30												
15	2.19	-0.90	0.10	1.00	1.78												
16	3.66	-0.61	0.16	0.96	1.62												
17	2.17	-0.55	0.36	1.31	2.24												
18	1.97	0.22	1.03	1.65	2.47												

	Reference Group (Rater 1) ALL Focal Group (Rater 2) :None					Focal Group (Rater 2): Minimal				Focal Group (Rater 2): Moderate				Focal Group (Rater 2): Severe			
Items	a	b1	b2	b3	b4	b1	b2	b3	b4	b1	b2	b3	b4	b1	b2	b3	b4
19	2.45	-1.20	-0.15	0.72	1.59												
20	1.51	-0.65	0.44	1.59	2.61												
21	1.57	-1.52	-0.35	0.66	1.92	-1.42	-0.25	0.76	2.02	-1.22	-0.05	0.96	2.22	-0.92	0.25	1.26	2.52
22	2.30	-0.29	0.69	1.45	2.33												
23	3.39	-1.22	0.00	1.08	1.90												
24	2.17	-1.56	-0.16	0.77	1.81												
25	2.77	-1.25	-0.34	0.43	1.09												
26	2.58	-1.35	-0.34	0.49	1.28												
27	1.91	-0.14	0.68	1.32	2.07	-0.04	0.78	1.42	2.17	0.16	0.98	1.62	2.37	0.46	1.28	1.92	2.67

CURRICULUM VITAE

Tamara B. Miller

EDUCATION

Graduate Student, currently Dissertator 8/04 to present
University of Wisconsin – Milwaukee, Milwaukee, WI
Department of Educational Psychology
Area of Concentration: Statistics and Measurement

Master of Science in Kinesiology 5/99
University of Wisconsin-Madison, Madison, WI
Emphasis: Measurement

Bachelor of Science in Physical Therapy 12/87
University of Wisconsin-Madison, Madison, WI

AWARDS

Margaret Kohli Award, School of Physical Therapy, University Wisconsin-Madison 12/87

GRADUATE ASSISTANTSHIPS DURING DOCTORAL STUDIES

University of Wisconsin – Milwaukee
Department of Educational Psychology
Research Assistant 8/11 to 8/14
Teaching Assistant 8/09 to 5/11
Research Assistant 8/05 to 5/09

GRADUATE ASSISTANTSHIPS DURING MASTER OF SCIENCE STUDIES

University of Wisconsin-Madison
Department of Kinesiology
Teaching Assistant 1/93 to 1/95

FACULTY POSITION

Carroll University
Department of Physical Therapy
Waukesha, WI
Adjunct Assistant Professor in Entry Level Masters of Physical TherapyProgram 8/99 to 8/04

COURSES TAUGHT

As a teaching assistant at University of Wisconsin-Milwaukee

Workshop: Computerized Analysis of Educational Data
Teaching Assistant for Educational Statistical Methods II

As a teaching assistant at University of Wisconsin-Madison

Teaching Assistant for Measurement of Motor Behavior
Teaching Assistant for Introduction to Biomechanics
Teaching Assistant for Electrotherapy for Physical Therapists

At Carroll University

Human Anatomy, Course associate
Basic Patient Management, Course coordinator
Teaching Laboratory Practium I, II and III , Course associate
Research Methods in Physical Therapy, Course coordinator
Independent Study Advisor

PROFESSIONAL PHYSICAL THERAPY EXPERIENCE

Physical Therapist Aurora at Hartford Hospital, Hartford, WI	5/99 to 8/99
Physical Therapist Cedar Haven Rehabilitation Agency, West Bend, WI	1/95 to 5/99
Physical Therapist University of Wisconsin Hospitals and Clinics, Sports Medicine Center, Madison, WI	2/92 to 8/93
Physical Therapist Independent Contracting in Southeast WI and NYC, NY	6/89 to 2/92
Physical Therapist Milwaukee Industrial Clinic, Milwaukee, WI	1/89 to 6/89
Physical Therapist Curative Rehabilitation Center, Milwaukee, WI	10/87 to 1/89

PUBLICATIONS

Schapira, M.M., Walker, C.M., Miller, T B., Fletcher, K.E., Ganschow, P.S., Jacobs, E.A., Imbert,D., O'Connell, M., Neuner, J. (in press), Development and validation of the Numeracy Understanding in Medicine Instrument (NUMi) Short Form. *Journal of Health Communication*.

Brosky, J.A., Hopp, J.F., Miller, T.B., & Deprey, S.M. (2001). Integrating theory and practice on wellness and prevention with older adults in self-contained clinical education experiences. *Journal of Physical Therapy Education*.

Miller, T.B., & Kane, M.T.(2000). The precision of change scores under absolute and relative interpretations. *Applied Measurement in Education.*

Kane, M.T., Miller, T.B., Trine, M., Becker, C., & Carson, K. (1995). The precision of practice analysis results in the professions. *Evaluation in the Health Professions.*

PRESENTATIONS

Miller, T.B., & Walker, C.M. (April, 2014). *Measuring Inter-rater Reliability in Performance Assessment Items Using DIF*, Poster presented at the Annual Meeting of the National Council for Measurement in Education, Phildelphia, PA.

Schapira, M., Okamato, M., Hokkaido, O., Kyutoku, Y., Sugimoto, Y., Dan, I., Miller, T.B., Walker, C.M. (January, 2014). *Translation and Application of the Numeracy Understanding in Medicine Instrument in Japan*., Paper presented at the Inaugural Meeting of the Asia-Pacific Society of Medical Decision Making. Singapore.

Okamato, M., Kyutoku, Y., Walker, C.M., Miller, T.B., Sugimoto, Y., Clowney, L., Dan, I., Schapira, M. (October, 2013). *Translation and Application of the Numeracy Understanding in Medicine Instrument in Japan*., Poster presented at the Annual Meeting of the Society of Medical Decision Making, Baltimore, MD.

Schapira, M., Walker, C.M., Ganschow, P.S., Jacobs, E.A., Fletcher, K.E., Neuner, J., Miller, T.B., Imbert,D. (October, 2013). *Development of a Computer Adaptive Test of Health Numeracy: The CAT Numeracy Understanding in Medicine Instrument (CAT-NUMi).* Paper presented at the Annual Meeting of the Society of Medical Decision Making, Baltimore, MD.

Walker, C. M., Schmitt, T. A., & Miller, T. B. (October, 2006). *How Valid are Self-report Data Obtained from School District Personnel.* Paper presented at the MSP Evaluation Summit II, Minneapolis, MN.

Truskowski, C., Deprey, S.M., Hopp, J.F., Miller, T.B., & Gosch, J.(February, 2005). *Physical Therapist Student Self- Assessment Ratings and Clinical Instructor Ratings Utilizing the Clinical Performance Instrument During Service Learning Practicums and Clinical Internship Experiences.* Poster presented at APTA Combined Sections Meeting, New Orleans, LA.

Miller, T.B., & Deprey S.M.(June, 2004). *Analyzing Blood Pressure Responses in Clients Receiving Soft Tissue Massage from Student Physical Therapists.* Poster presented at National Physical Therapy Conference Chicago, IL.

Brosky, J.A., Deprey, S.M., Dornemann, T., Hopp, J.F., & Miller, T.B., (February, 2002). *Integrating Service Learning into an Entry-Level Physical Therapist Curriculum*. Paper presentation at Section of Education, Combined Sections Meeting, Boston, MA.

Kane, M.T., & Miller, T.B. (April, 2001). *The precision of change scores under absolute and relative interpretations.* Paper presentation at National Council on Measurent in Education Annual Conference Seattle, WA.

GRANTS

Deprey, S.M**.,** & Miller, T.B. Analyzing Blood Pressure Responses in Clients Receiving Soft Tissue Massage from Student Physical Therapists. Carroll College Received July 2003 Faculty Development Grant $260.

Hopp, J.F., Brosky, J.A., Deprey, S.M. , & Miller, T.B. Older Adult Caregiver Safe Mobility Educational Video Initiative, Alvin and Marion Birnschein Foundation, Incorporated. Funding period: January 1, 2003 – December 30, 2003. Total funding $5,000.

Hopp, J.F., Brosky, J.A., Deprey, S.M., & Miller, T.B. Filling the Gap. Improving the Quality of Life for Youth and Adults with Special Needs" Stackner Family Foundations. Funding period: July 2002 -July 2004. $30,000.00.

SERVICE

Proposal Reviewer: NCME Graduate Student Research Proposals (October 2013)
Member of the University of Wisconsin-Milwaukee Graduate Student Advisory Council (2012 to 2013)
Member of Editorial Board for Carroll University Publication, *The Faculty Forum* (2002 to 2004)
Assessment Coordinator for Carroll University Physical Therapy Program (2001 to 2004)
Alumni Coordinator for Carroll University Physical Therapy Program (2000 to 2004)
Admissions Coordinator for Carroll University Physical Therapy Program (2000 to 2004)