6-2016

ITEM RESPONSE THEORY ANALYSIS OF THE TOP LEADERSHIP DIRECTION SCALE

Jung-Jung Lee
California State University - San Bernardino, leej394@coyote.csusb.edu

ITEM RESPONSE THEORY ANALYSIS OF THE TOP LEADERSHIP DIRECTION SCALE

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Psychology:

Industrial/Organizational

---

by

Jung-Jung Lee

June 2016

ITEM RESPONSE THEORY ANALYSIS OF THE TOP LEADERSHIP DIRECTION SCALE

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Jung-Jung Lee

June 2016

Approved by:

Dr. Janet Kottke, Chair, Psychology

Dr. Kenneth Shultz, Committee Member

Dr. Matt Riggs, Committee Member

ABSTRACT

Item response theory (IRT) offers several advantages compared to classical test theory (CTT) in providing additional information on psychometric qualities of the scale. My goal was to demonstrate the superiority of IRT as compared to CTT through two analyses of the Top Leadership Direction scale (TLDS), which was created to measure the effectiveness of top leadership through the followers' perceptions in the context of providing guidance of the organization. Furthermore, the participants (n = 8046) were the employees from various positions at 18 of the 23 California State University campuses. In the graded response model (GRM) analysis, the result showed that IRT provided more information about each item and allowed a useful visual inspection of the items. With the second analysis, I aimed to provide evidence of measurement equivalence across functional groups of employees using differential item functioning (DIF) analysis in IRT. Due to the lack of model fit, the DIF analysis was incomplete. A supplementary multigroup CFA was conducted to investigate the structural difference across the groups for the items of the TLDS. The result of multigroup CFA suggested that item 2 and item 4 did not show measurement equivalence across the groups at the construct level. An alternative model in IRT was discussed due to some limitations of GRM in the present study. Practical and theoretical implications for the use of IRT were also presented and contrasted with CTT.

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Dr. Janet Kottke, for her time and energy to guide my through my thesis from start to finish. I would also like to thank my committee members, Dr. Shultz and Dr. Riggs, for giving me additional direction for my study. I would like to give a special thanks to Dr. Diaz, who willingly served on my committee at my defense meeting and provided feedback to my thesis. My best friends in my cohort: Rachel, Claudia, and Eric for listening to me talk about stats all the time. Finally, I will always remember the support from my family and friends who always told me to continue my efforts and never give up.

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

Item response theory (IRT; Lord & Novick, 1968; Rasch, 1960) has become progressively more popular since the 1970s and 1980s, as more sophisticated computer software became available. In the beginning, IRT was developed primarily to test ability with dichotomous response formats. Specifically, one of the advantages of IRT is to assess more accurately an individual's trait level when compared to Classical Test Theory (CTT). As a result, one popular use of IRT is in computer adaptive testing, which uses a pool of questions that separate by difficulty levels to obtain more accurate results of the test taker's trait level. As IRT became more widely known, it has been applied to fields other than ability testing, which has led to IRT being used with more and varied types of response formats. For example, IRT has been applied to personality and attitude scales with polytomous response formats (Collins, Raju, & Edwards, 2000; Flannery, Reise, & Widaman, 1995; Ryan, Ployhart, Schmitt, & Slade, 2000).

IRT has a strong theoretical background and is model-based measurement, which models are decided first before analyzing the data. The field of IRT is still growing and there are a variety of models that can be applied based on the nature of the data. As already suggested, IRT has many advantages over CTT, but most researchers are better trained in CTT than IRT. This limitation is largely a function of the highly technical language used

in IRT research and the difficulty in obtaining the needed software (Fraley, Waller, & Brennan, 2000). In addition, to understand and select an appropriate model for the data can be difficult for researchers who are new to the field of IRT. IRT researchers have offered evidence of model comparisons using Monte-Carlo studies (Baker, Rounds, & Zevon, 2000; Maydeu-Olivares, Drasgow, & Mead, 1994; Maydeu-Olivares, 2005); however, these studies often show inconsistent results. Furthermore, the fit statistics for model evaluation are problematic (Embretson & Reise, 2000; Ostini & Nering, 2006), in that there is no agreed upon best fit statistic in IRT with many researchers continuing to propose new fit indices for the models. Finally, adding to the complexity in modeling data using IRT methods is that there is no universal software that can cover all the models and fit statistics when using IRT procedures.

I intend to address the advantages and limitations of IRT in this research. The present study is written in non-technical language and examples will be provided to better illustrate difficult concepts of IRT. The focus on model selection is for an attitude scale. I will start by describing the advantages of IRT over CTT in detail, going through the general IRT framework, model selection, and testing measurement equivalence using differential item functioning.

Comparisons between Classical Test Theory (CTT) and Item Response Theory (IRT)

To appreciate the benefits of IRT, Embretson and Reise (2000) proposed ten rules that demonstrated the superiority of IRT over CTT (see Table 1). Here I will offer those rules to lay a foundation for my proposed research.

New Rule 1

The estimate of the standard error of measurement is different in two ways when IRT is compared to CTT. In CTT, one standard error of measurement is based on the reliability and standard deviation of the sample. In IRT, the standard error of measurement can be calculated for each score. In CTT, every observed score has the same standard error of measurement; in contrast, the standard error of measurement is applied differentially to scores in IRT. More specifically in IRT, the standard error of measurement will be smaller when the item is more appropriate for a particular trait level; however, the standard error of measurement increases for scores at the extremes where item information diminishes. Overall, IRT provides more accurate predictions of the trait estimates when compared to CTT.

The second important difference regarding the standard error of measurement is related to generalizability across samples. CTT is sample specific because the standard deviation and the reliability estimate used to calculate the standard error of measurement are from the current sample. When using another sample of respondents, the standard deviation and the

reliability estimate are going to be different and thus lead to a different standard errors of measurement. Unlike CTT, the standard error of measurement in IRT is not based on the characteristics of the sample but rather on independent scores after controlling for difficulty levels. Therefore, the standard error of measurement based on IRT can be generalized across samples rather than a specific sample as with CTT.

New Rule 2

The second “new” rule (Embretson & Reise, 2000) deals primarily with reliability and the length of the test. In CTT, the reliability coefficient increases as the length of the test increases because the number of items is included in the equation to calculate the reliability coefficient are assumed to add more true score variation and proportionately less error variance. Embretson and Reise (2000) gave an example of the differences in reliability coefficients in CTT while comparing an original test to a shortened parallel test. The shortened parallel test maintained two-thirds of the length of the original test and the reliability coefficient of the shortened test was .80 compared to .86 for the original test. This example is consistent with the reliability coefficient calculated based on CTT.

As a comparison, the authors described a test (i.e., a computer adaptive test) developed using IRT to demonstrate that a shorter test had a higher reliability coefficient than a longer test, which conflicts with the general rule provided by CTT. For example, computer adaptive testing (CAT) has

been widely used for many standardized tests (i.e., GRE and licensing exams). Typically, in CAT, test takers receive different items based on their ability to answer the items correctly. Embretson and Reise (2000) showed that a 20-item adaptive test had a lower standard error of measurement across trait levels when compared to a 30-item traditional test (test takers completed all 30 items). As a result, the second rule postulates that a shorter scale can be more reliable than a longer scale using IRT when compared to CTT.

New Rule 3

The third “new” rule involves comparing scores across different test forms (Embretson & Reise, 2000). To compare scores across different test forms, CTT assumes that the test forms have to be parallel to each other. The parallelism of test forms is determined by the equivalence of the means, variances, and reliabilities of the tests. Based on CTT, equating methods should be used before comparing the scores from non-parallel tests. Therefore, when two tests are very different in their means, variances, and reliabilities, more equating errors will occur. Equating errors make the comparison between scores from different tests less accurate. Furthermore, test difficulty can lead to even more problematic predictions when comparing scores using CTT.

To illustrate the issue of equating tests of differing difficulty, Embretson and Reise (2000) conducted a simulation: 3000 examinees were administered two test forms with the same discrimination level but different difficulty levels.

To compare the scores of the examinees on two test forms, the researchers regressed the easy test scores on hard test scores (see figure 1). In their simulation, they demonstrated two issues in equating test scores from an easy test to a hard test using CTT. First, the easy scores regressed on hard scores showed a higher correlation using a cubic regression fit line (i.e., nonlinear regression line) in contrast to linear regression. Second, with nonlinear regression, large variances were observed between the actual score and the nonlinear regression line when looking at low (easy) test scores on the hard test. Since the scores were coming from the simulated data, they further calculated the reliabilities of the test scores based on the trait levels using both CTT and IRT. In CTT, two reliability coefficients were obtained, for the easy test and the hard test. As mentioned previously, both sets of scores showed nonlinear trends along the linear regression line. In IRT, the test scores from a 30-item computer adaptive test were also graphed to the figure along with the regression line based on the trait levels. IRT showed better reliability and prediction of scores compared to CTT.

New Rule 4

The fourth “new” rule illustrated that the item properties are not biased, not dependent on sample representativeness (Embretson & Reise, 2000). In CTT, the probability of passing proportions are used to show the corresponding item difficulty level for the exam. Item-total correlations are similar to item discrimination. To obtain the probability and item-total

correlations based on CTT, the underlying assumption is that the sample has to be representative of the intended population. An unrepresentative sample will give unreliable results. That is, the item-total correlations and the probability will differ from sample to sample.

Embretson and Reise (2000) demonstrated the effect of having an unrepresentative sample in CTT and in IRT using 3,000 scores generated from a simulation. The scores were separated into either a low group or a high group based on the median, and then the scores were shown on a scatterplot with the probability scores of the high group on the x-axis and the probability scores of the low group on the y-axis. In figure 2, the probability scores showed a nonlinear trend, which meant the predictions based on fitting a linear regression line were not consistent across the scores. The correlation between the probability scores was .80. In figure 3, item difficulty level scores were obtained from the Rasch model in IRT. The scores showed an obvious linear trend with a correlation of .99. This demonstration revealed the superior reliability of the IRT analysis in obtaining the test statistics for item properties when compared to the analysis based on CTT when the sample was not representative.

New Rule 5

The fifth “new” rule indicates that IRT gives meaning to the scores by placing them on a continuum of difficulty of items (Embretson & Reise, 2000). In CTT, the meaning of individual scores is obtained through making

references to the norm groups. Individual scores are usually standardized and then positioned, based on the normal distribution of the norm group. To provide a comparison between CTT and IRT for this new rule, Embretson and Reise (2000) provided an example using data from a behavioral scale for older people. A sample item from the scale is “the ability of bladder control.” The individual scores were calculated based on the number of activities that an individual was able to perform in his or her daily life.

To obtain the meaning of an individual score in CTT, a histogram of z-scores was displayed based on the population scores from the data. Each individual score was then placed in the distribution of all the scores. The score interpretation was based on the score’s relevant position compared to others. For example, an individual with a score in the low end of the distribution would be interpreted as being less able to perform the activities compared to individuals who were in the similar age group. However, this score interpretation did not give any information about how many activities a person could actually perform.

To give meaning for an individual score, IRT places the score in reference to the items on the scale instead of placing the score in the context of other scores in the population, as does CTT. In this example, IRT analysis would first rank activities listed on the scale based on their difficulty levels. An easy activity would be placed on the left side of the continuum, and a hard activity would be placed on the right side of the continuum. An individual score

could be reported as the individual's trait level for the scale. The individual's trait level then could be positioned on the item difficulty continuum to compare with other individuals. In IRT, each individual trait level has a correspondent difficulty level. Therefore, IRT analysis not only gives the meaning of the score when comparing to the population, but also gives information regarding what a given individual can do based on his or her correspondent difficulty level. If an individual's trait level was at the moderately difficult activities, this suggests that the individual had a 50 percent chance to perform successfully the moderately difficult activities, a higher than 50 percent chance to perform low difficulty activities, and a less than 50 percent chance to perform high difficulty activities.

New Rule 6

The sixth "new" rule involves the assumption of a normal distribution with an interval-level scale (Embretson & Reise, 2000). One of the assumptions of CTT is that the scores are normally distributed. To achieve a normal distribution in psychological testing, two common methods are usually used. First, items with 50 percent passing rates are selected. Second, nonlinear transformations are applied to the raw scores. The problem associated with nonlinear transformations of the raw scores is that such a transformation changes the relative distance between scores after applying the transformation. For example, two scores that differ by five points on the original scale might be different by less than 1 point on a new scale after the

nonlinear transformation. Interval level of measurement assumes that the scores are in equal intervals. Therefore, the assumption of interval level of measurement is violated. However, this assumption of interval level of measurement seems to be less severe if the scores are normally distributed.

As already noted, the justification of equal intervals is based on not violating the assumption of a normal distribution; however, this assumption raises another problem. The distribution of the scores is derived from a specific population. Furthermore, when applying nonlinear transformations, the equal interval assumption is justified, but this justification is only valid under the condition that the scores are from the same population. To conclude, interval scale properties are population specific when applying a nonlinear transformation to the raw scores to achieve a normal distribution in CTT. In contrast, IRT interval scale properties do not need to meet the assumption of a normal distribution. Depending on the nature of the data, a specific measurement model can be selected in IRT analysis. Therefore, no nonlinear transformation is needed for IRT.

New Rule 7

The seventh “new” rule shows the strength of IRT when encountering mixed item formats within a scale (Embretson & Reise, 2000). CTT provides the basic information of a scale by calculating means and standard deviations. When a scale’s response format is not uniform (e.g., some items on 4-point Likert response format and some items on an 8-point Likert response format),

the mean and standard deviation will become difficult to interpret in CTT. Embretson and Reise (2000) provided an example to demonstrate the change in mean and standard deviation by doubling the range of the response format of a single item in a ten-item scale. The mean and standard deviation increased in the altered scale because one item had more variances compared to the original scale.

Several methods have been proposed in CTT to fix the mixed format issue. One of the methods has been to standardize the scores before performing the calculation. However, standardized scores are sample specific and make generalization problematic. The second approach has been to simply divide the scores with the largest range by two before calculating the mean and standard deviation of the scale. This approach would not produce consistent results if a scale employed different response formats not proportional to each other. For example, some items might be in a 4-point response format and some in a 5-point response format. In IRT, each item has its own item characteristic curve and the overall information contributed from the item will not change based on a different response format.

New Rule 8

The eighth “new” rule deals with the measurement issue when equating score differences (Embretson & Reise, 2000). In CTT, equating score differences is only possible when the scores are in equal intervals; however, it might be hard to achieve equal intervals when a test consists of questions with

different difficulty levels. For example, improving one point from an easy question has a different meaning than improving one point from a hard question. Therefore, this result suggests that using CTT to calculate score differences is inappropriate. Cronbach and Furby (1970) further illustrated this issue of calculating score differences by modifying the existing formulas to provide better estimation; however, they concluded that there was no good way to find the true score difference and the best solution was not to use it at all. This is not an issue in IRT. Embretson and Reise (2000) pointed out that IRT can readily calculate score differences based on the person's trait level; therefore, IRT can provide meaningful results in equating score differences even when the questions are at different difficulty levels.

New Rule 9

The ninth "new" rule describes how IRT overcomes the problem of factor analyzing binary items in CTT (Embretson & Reise, 2000). Carroll (1945) provided an example to explain how Pearson's correlations between items can have different meanings when items vary in their difficulty levels. The magnitude of the correlation coefficient among the items should be related to the underlying constructs that the items represent. For example, highly correlated items suggest that those items come from similar underlying construct and low correlations among the items mean that they are coming from rather different constructs. However, Carroll (1945) demonstrated that even if all the items were from the same construct, Pearson's correlation

coefficients would be dependent on their difficulty level. He concluded that Pearson's correlation is not appropriate when items have different difficulty levels and he further proposed that tetrachoric correlation should be used instead, which has been a relatively common practice since.

Embretson and Reise (2000) suggested several limitations using tetrachoric correlation. For example, applying adjustments to item correlations might result in singularity and the assumptions of linearity and normality would be violated. However, IRT correlation methods can overcome the limitations related to CTT correlation methods. For example, full information factor analysis (Bock, Gibbons, & Muraki, 1988) uses all the information from the data. Furthermore, adjustment of the model was not needed because the most appropriate IRT model was applied to the data based on the nature of the data.

New Rule 10

The tenth "new" rule indicates that item stimulus features can provide additional information on item properties in IRT but are often ignored in CTT (Embretson & Reise, 2000). Embretson and Reise (2000) indicated that item features are often only considered with content validity and test bias in the beginning of the test development process; however, item features rarely play a part in the later item selection process. An example can be in developing a certain ability scale. When writing items for a scale, test developers follow some specific definitions for items on each dimension. The item features can

be general or specific, within a certain context, or targeting a certain psychological perspective (i.e., affect, behavioral, or cognitive). Then, the next process is to collect data and select the items based on their psychometric properties (e.g., item loadings). However, the psychometric properties of the scale hardly inform the researchers about the specific item features related to the test. In contrast, certain models in IRT (e.g., multicomponent latent trait model; Whitely, 1980) can associate specific item features, psychometric properties, and even the trait level of the individual.

In sum, the ten “new” rules have clearly demonstrated the benefits of IRT over CTT, but a number of these rules are primarily associated with ability testing. In my thesis, I am using IRT to evaluate an attitude scale and compare the quality of this scale evaluation to the information from CTT evaluation results. Specifically, I will provide information on standard error of measurement and the meaning of the responses. Furthermore, I believe the information will be sample independent.

Fundamentals of Item Response Theory (IRT) Modeling

Because in this thesis I intend to examine a Likert-typed scale, I will describe relevant IRT models for that type of scale after a brief overview of general IRT modeling. In addition to the ‘new’ rules of IRT that can be used to make comparisons with CTT, IRT and CTT are different in their parameter estimation methods. It is very important to explain the parameter estimation method for IRT analysis because IRT is based on modeling. Generally, the

most common method of CTT parameter estimation is least-squares estimation (LSE), and IRT is strongly driven by maximum likelihood estimation (MLE). Myung (2003) suggested that the two methods are different in nature because LSE can only summarize the data while MLE is used to test hypotheses or build confidence intervals. He further provides a step-by-step tutorial to demonstrate how MLE works.

Myung (2003) demonstrated two different methods to estimate the underlying population distribution if the data were given. The first method was to calculate a probability density function from a fixed parameter of a given data set. The purpose was to focus on the function of the data. The second method was to calculate the likelihood function of the parameters from the given data. In this case, the focus was on the function of the parameters instead of the data. The likelihood function was more useful to estimate the underlying distribution of the population. The maximum likelihood estimation was a likelihood function that maximized the parameter values to conform to the underlying distribution most likely to emerge from the observed data.

Item Response Theory (IRT) Basics

The purpose of IRT is to estimate the latent trait based on individuals' responses to a set of items and to estimate the item properties using IRT models (Embretson & Reise, 2000). To better illustrate how IRT models operate, I will start by describing models with binary responses (e.g., true/false, yes/no, and agree/disagree). These concepts from binary models

can be extended to models with polytomous responses (e.g., Likert-scale response).

The first feature of IRT models relates to the parameters used for the estimation. Two types of parameters are displayed in all IRT models: person parameter and item parameter(s). A person parameter is the latent trait, usually represented by theta (θ), which indicates the probability of endorsing an item at a given trait level. Theta is displayed in a continuum similar to z units with a mean of zero and a standard deviation of one. In practice, the range of the theta is usually from -3 to 3. There are three possible item parameters: difficulty, discrimination, and/or pseudoguessing. Different IRT models vary in their item parameters and all IRT models have at least one item parameter (i.e., difficulty). Which item parameter is selected depends on the nature of the data. For example, if the researcher is only interested in knowing the difficulty of a given exam and has no reason to believe other parameters would differ, then it is best to choose the model that includes the difficulty parameter but not the other two parameters. The target scale to be used in this thesis, top leadership direction, is an attitude scale, so the two item parameters, difficulty and discrimination, will be used to evaluate the scale items.

In the binary case, the difficulty parameter indicates the latent trait level required to have a 50% probability of endorsing the item and is usually denoted by b, which is often shown on the same scale as theta. An item with a

high difficulty level means that this item requires a high trait level to have a 50% probability to endorse the item; if a person does not have that level of ability, the item will not be endorsed. A lower trait level will have less than a 50% probability to endorse the item. For example, a difficulty level of 2.00 means that this item requires the trait level of 2.00 to have a 50% probability to endorse the item and individuals with a trait level less than 2.00 are not likely to endorse the item. A difficult item is endorsed by individuals with high trait levels and is better at discriminating individuals with high trait levels because the individuals with low trait levels are not likely to answer the items correctly. An easy item is endorsed by individuals with moderate and high trait levels and is better at discriminating individuals with low trait levels because the individuals with high trait levels are much more likely to answer the items correctly.

The second parameter, discrimination, indicates the change of the individuals' probability of endorsing an item between different trait levels and is often denoted by a, which can range from 0.00 to positive infinity, theoretically. However, it most often ranges from 0.75 to 2.50 in practice (Flannery et al., 1995). An item with a larger a value suggests that the probability of endorsing an item changes dramatically at a certain trait level. High discrimination items are better at differentiating individuals at different trait levels and are said to have better quality than the items with low discriminability.

Based on the parameters of interest, an appropriate dichotomous IRT model can be used to calculate the probability of the item endorsement associated with a specific trait level, which is defined as an item response function (IRF). Since we are interested in difficulty and discrimination parameters, a two-parameter logistic (2PL) model (Birnbaum, 1968) is most appropriate. See figure 4 for a sample IRF. An IRF equation for the 2PL model is shown below.

$$P(\theta) = 1/[1 + \exp(-a(\theta - b))],$$

Where, $P(\theta)$ = the probability of endorsement at the trait level θ, a = item discrimination parameter, b = item difficulty parameter.

Furthermore, an important function that is unique to IRT is the ability to calculate an information index from the IRF, which is called an item information curve (IIC; see figure 5). In the graph of IIC, the x-axis indicates the individual's trait level and the y-axis indicates the amount of the information. The amount of information given by an item is contingent upon item discrimination and difficulty parameters. An item with high discriminability yields a higher amount of information and also shows a higher peak in its item information curve. An item with low discriminability yields a lower amount of information and also has a less peaked item information curve. For an item, the location where the difficulty value equals the trait level is the location that provides the most information. That is, the highest peak of the IIC is the point of most information. The equation to calculate information is shown below:

$$I(\theta) = a^2 \times P(\theta) \times [1 - P(\theta)],$$

Where $I(\theta)$ = the amount of information for an item at a given level of θ, a^2 = squared item discrimination parameter, $P(\theta)$ = the probability of endorsement at the trait level θ.

Once the IIC of each item has been estimated, a test information curve (TIC) can be obtained. TIC is the sum of IICs from all the items in the scale and can be used to indicate the amount of information available on a test. TIC also has an individual's trait level on the x-axis and the amount of information on the y-axis similar to IIC. Furthermore, IIC and TIC are not the same across the trait continuum. That is, an item or a scale might be more informative at certain trait levels compared to other trait levels. This characteristic of IIC and TIC represents relative precision across the trait continuum. Therefore, the standard error of measurement (SEM) associated with the trait continuum can be estimated from the value of IIC and TIC at a specific trait level. The higher the values of an IIC or a TIC, the lower the values of SEM. The lower the values of an IIC or a TIC, the higher the values of SEM.

Differences Between Dichotomous Polytomous Models

Ostini and Nering (2006) explained the difference between dichotomous and ordered polytomous models. One way to extend a dichotomous model into a polytomous model for polytomous items is to dichotomize polytomous items. Different polytomous models can be produced based on various ways of dichotomizing polytomous items and combining the

results of the dichotomization. Comparing to dichotomous items, ordered polytomous items have more than two response categories. The categories are separated by boundaries or thresholds. For example, dichotomous items have two categories and one boundary or threshold. A five-point Likert-scaled item has five categories and four boundaries (thresholds). The number of the boundaries or thresholds is one less than the categories of the item. See figure 6 for a sample polytomous model.

Another difference between dichotomous and polytomous models is the probability calculated from the IRF (Ostini & Nering, 2006). In dichotomous models, the estimated probability means the probability of choosing to respond to the positive category and the responses are in a positive direction at the boundary. However, this is not true with polytomous items with more than one boundary. Therefore, a single probability in dichotomous models becomes two types of probabilities in polytomous models. The two types of probabilities are being calculated from two different functions: category boundary response function (CBRF) and item category response function (ICRF). CBRF represents the probability of having the responses in a positive manner at each boundary. ICRF represents the probability of choosing to respond in a specific category.

Graded Response Model (GRM)

Samejima (1969; 1972) proposed a graded response model with the purpose of applying IRT to ordered polytomous items. Theoretically,

Samejima (1969) classified GRM into two broad cases: homogeneous and heterogeneous cases, which can be distinguished based on how responses are derived. In homogeneous cases, the cognitive process of coming to an answer is direct. In contrast, the heterogeneous case requires a more complicated cognitive process of choosing a response. Since an attitude scale is used in the present analysis, the model description is going to be based on the homogeneous case framework.

As mentioned previously, polytomous models developed from various ways of dichotomizing polytomous items. In the GRM (Samejima, 1969; 1972) framework, the number of dichotomies is one less than the number of categories. For example, a five-point Likert-type item has five response categories and four response dichotomies. Thus, the dichotomies are being compared in a graded manner: Category 1 compared to Categories 2, 3, and 4, Categories 1 and 2 compared to Categories 3 and 4, and Categories 1, 2, and 3 compared to Category 4. Finally, CBRFs of GRM can be calculated for each dichotomy using the generalized 2PL IRF. In this example of a five-point Likert-typed item, four CBRFs are generated.

The next step in a GRM framework is to calculate the probability of choosing to respond in a specific category: ICRF, which can be calculated through resulting CBRFs. The assumption behind ICRF is that when an individual chooses a response option, he or she must go through all previous categories (Samejima, 1972). Therefore, the ICRF of a given category is the

probability of responding in a positive manner of the given category minus the probability of responding in a positive manner of the next category. The ICRF equation is shown below:

$$P^*_x(\theta) = P_x(\theta) - P_{x+1}(\theta),$$

Where $P_x(\theta)$ is the probability of responding in category x. $P_{x+1}(\theta)$ is the probability of responding in the next highest category or higher.

Samejima (1969) further pointed out several conditions with ICRFs. First, $P_x^*(\theta)$ has the same discrimination parameter values but different difficulty parameter values across categories within an item. Second, $P_x^*(\theta)$ of the lowest category of an item is 1. Third, $P_x^*(\theta)$ of the highest category is 0. Therefore, the ICRFs of a five-point Likert-type item can be written as:

$$P_1^*(\theta) = 1 - P_2(\theta), \text{ monotonically decreasing}$$

$$P_2^*(\theta) = P_2(\theta) - P_3(\theta),$$

$$P_3^*(\theta) = P_3(\theta) - P_4(\theta),$$

$$P_4^*(\theta) = P_4(\theta) - P_5(\theta),$$

$$P_5^*(\theta) = P_5(\theta) – 0, \text{ monotonically increasing}$$

Finally, the information and SEM of each item and scale can be estimated in GRM, based on similar concepts of calculating information of IIC and TIC illustrated for 2PL model.

Comparing Graded Response Model (GRM) with Other Polytomous Models

There are other polytomous models that might be possible to apply to Likert-typed data. To better understand other polytomous models, it is efficient to describe the polytomous models by Thissen and Steinberg's (1986) taxonomy of item response models. The two relevant classes of polytomous models for Likert-scale data are difference models and divide-by-total models (see figure 7 for graphical representation of these two type of models). Samejima's (1969) GRM is the best representative of a difference model. The divide-by-total models that can be used for Likert-typed data are the partial credit model (PCM; Masters, 1982), Thissen and Steinberg's (1986) extension of the PCM model, an equivalent to the generalized partial credit model (GPCM; Muraki, 1992), and the nominal response model (NRM; Bock, 1972).

The different ways of categorizing the models are primarily due to two general types of model expressions. As mentioned previously, polytomous models are derived based on different methods of dichotomizing data to obtain the probability of theta (Ostini & Nering, 2006). GRM typed models take into consideration all the categories when making comparisons. The categories are separated by a boundary. The boundary function is estimated from comparing each dichotomy above and below the boundary. A second way to dichotomize data to obtain the probability of theta takes into account only the category immediately above and below the boundary. For example, to estimate the first boundary function, the probability is estimated from the first

category and the second category (i.e., the ‘1’ and ‘2’ on a 5-point scale) and not all the remaining categories.

As already noted, two general types of model equations to calculate CBRF, based on these two ways dichotomous data, can be expressed. In Thissen and Steinberg’s nomenclature (1986), the models are classified as difference and divide-by-total models. GRM-typed models are classified as difference models because the CBRF of a given category is the probability of responding in a positive manner of the given category minus the probability of responding in a positive manner of the next category. The CBRF equation is shown below:

$$P_x^*(\theta) = P_x(\theta) - P_{x+1}(\theta),$$

Where $P_x(\theta)$ is the probability of responding in category x. $P_{x+1}(\theta)$ is the probability of responding in the next highest category or higher.

However, the divide-by-total model is defined because it has the general model expression of “an exponential divided by a sum of the total of all the exponentials” (Thissen & Steinberg, 1986, p. 569). The general IRF for the divide-by-total models is shown below:

$$P_x^*(\theta) = P_x(\theta)/[\, P_{x-1}(\theta) + P_x(\theta)],$$

Where $P_x(\theta)$ is the probability of responding in category x. $P_{x1}(\theta)$ is the probability of responding to the category x-1.

Furthermore, in another categorization approach Embretson and Reise (2000) also called difference models and divide-by-total models, indirect

models and direct models, respectively. GRM-typed models are defined as indirect because two equations are needed to estimate the conditional probability of an individual trait level when responding in a specific category. Other models are considered as direct models because one equation is needed to obtain the estimate.

Among the divide-by-total models, PCM, GPCM, and NRM can be applied to Likert-typed data. The difference between PCM, GPCM, and NRM lies in their parameter constraints. PCM has the most constraints and assumes all items have the same discrimination parameter. GPCM has moderate constraints and assumes a discrimination parameter for each item, which is the same as GRM. NRM has only minimal constraints compared to the other two models, which estimates one discrimination parameter for each category of each item.

Several studies have made comparisons between difference (indirect) and divide-by-total (direct) models on Likert-typed data using different methods (Baker et al., 2000; Maydeu-Olivares et al., 1994; Maydeu-Olivares, 2005). Maydeu-Olivares et al.'s (1994) study compared GRM and GPCM using a personality scale. Their results suggested both GRM and GPCM are appropriate for Likert-typed personality scales. Similarly, Baker et al. (2000) compared GRM and PCM using a mood scale. However, their results suggested that the data fitted GRM better than PCM. Furthermore, Maydeu-Olivares (2005) provided a comprehensive comparison between

GRM and all three models from the divide-by-total model: NRM, GPCM, and PCM, using a personality scale. They all found evidence on these key findings that supported the appropriateness of applying GRM for Likert-typed data. To further elaborate, each study is summarized in the following paragraphs.

In Maydeu-Olivares et al.'s (1994) study, simulated data were generated with different scale lengths (5-, 15-, and 25-items) and different sample sizes (N = 250, 500, 1000, and 3000) from a social problem solving scale. A sample item is "when my first attempt to solve a problem fails, I believe if I don't give up, I will eventually succeed." To make comparisons between the two models, a two-step process was required. The first step was to generate the data from each model. The second step used both models to estimate data that were generated from each model. For example, GRM was used to generate data and both GRM and GPCM were used to reproduce the generated data, and vice versa for GPCM. The comparison then could be made from the data estimated from GRM and GPCM. It was assumed that the data fit better when generated and estimated with the same model. Therefore, a model was said to be superior when the different model made a better estimate of the item parameters. For example, if GRM is used to generate data, but GPCM shows better estimates compared to GRM, then GPCM is the superior model. Their results indicated that both models fit better when the data were generated from the same model and both models were appropriate for Likert-typed data (i.e., fit the data).

In Baker et al.'s (2000) study, the survey data were collected from 713 undergraduate psychology students. The survey included 80 Likert-typed items that related to the psychological latent trait of subjective well-being on mood. The model-data fit was estimated with both GRM and PCM. Furthermore, the assumption of item parameter invariance was analyzed and the model accuracy was tested by comparing the conditional probability of theta between low and high trait items for each model. Based on model-data fit, the result suggested that GRM had a better fit than PCM because it was necessary to allow the discrimination parameter to vary between items; GRM also showed better item parameter invariance than PCM.

In Maydeu-Olivares's (2005) study, the survey data were collected from undergraduate psychology students in two different time periods and both samples had about 1000 participants. The second sample was used to cross-validate the results of the first time period. The survey included 52 Likert-typed personality items that intended to measure how individuals solve problems. The author made comparisons between GRM and three divide-by-total models (i.e., PCM, GPCM, and NRM) based on goodness-of-fit. It was expected that NRM would perform better compared to PCM and GPCM because NRM has the least constrained parameters among the three. The results suggested that NRM did perform better for the first sample compared to PCM and GPCM, but not always in the cross-validation sample. PCM showed the worse fit among the three divide-by-total models. Furthermore, the

results suggested that GRM had the smallest chi-square to degrees of freedom ratio over all three divide-by-total models in both samples. Maydeu-Olivares (2005) further concluded that GRM is most appropriate for personality data.

Baker et al. (2000) and Maydeu-Olivares (2005) found similar results, suggesting that GRM is the most appropriate model for Likert-typed data. However, the results were inconsistent between Maydeu-Olivares et al.'s (1994) study and Maydeu-Olivares's (2005) study; the reason might be due to different study designs and different estimation methods between these two studies. That is, Maydeu-Olivares et al. (1994) used a simulated sample and a different estimation method (i.e., ideal person index), and Maydeu-Olivares (2005) used actual participants and a residual goodness-of-fit. Even though Maydeu-Olivares et al. (1994) did not indicate that GRM was superior to GPCM, they did suggest that both models could provide good fits to the data. Taken together, these three studies strengthen the conclusion that GRM is the most appropriate model to use for an attitude scale.

General Assumptions, Regardless of Model

The general assumptions of most IRT models are unidimensionality and local independence of the responses. Unidimensionality suggests that there is one common factor of all items on the scale. Unidimensionality is usually assessed with factor analysis (EFA/CFA). Local independence suggests that the items are not related when holding theta level constant.

Furthermore, the last assumption is that the probability of endorsement increases monotonically as the trait level increases.

Fit Statistics

IRT is a model-based measurement. To gain an idea about how good the model is, fit statistics are thus often utilized. In general, model fit statistics can be categorized into four classes: residual-based, multinomial distribution-based, response function-based, and Guttman error-based (Ostini & Nering, 2006). The general way to calculate the fit for each class is summarized in the following paragraphs. Furthermore, the problems related to fit statistics are discussed at the end of this section.

The simplest way to calculate the fit statistics for residual-based measures is the difference between observed and expected respondent scores. The difference then can be standardized by dividing the standard deviation of the observed score. The fit for each item can be obtained by summing the residuals for all the participants of that item. The formula to obtain the sum is through calculating the mean square, which is the squared standardized residuals divided by the total number of the participants. This calculation of residual-based fit statistics is often called an unweighted mean square. A weighted mean square can be calculated by multiplying standardized residuals by the variance of observed responding scores, and then dividing by the sum of the variances. Unweighted and weighted mean

squares can be transformed to unweighted and weighted t-statistics to obtain normal distributions.

In multinomial distribution-based fit statistics, the models are compared based on the distribution of response patterns. The possible response patterns can be calculated from the number of the items and the item categories. For example, five items with seven categories have 5^7 possible response patterns. The model fit is obtained by comparing the observed and expected response patterns. The commonly used test statistics are different types of chi-squares (e.g., log-likelihood ratio and Pearson's chi-square).

Response function-based fit statistics measure person fit. The fit statistic is calculated from the difference between expected and observed log-likelihoods of each item's responses instead of response patterns in multinomial distribution-based fit statistics. Other than response function-based fit statistics, Ostini and Nering (2006) suggested that Guttman error-based fit statistics used a nonparametric method to calculate the fit by counting the number of errors across pairs in Guttman responses. An example of a Guttman error-based statistic is the Q_i statistic.

Regardless of fit index, Ostini and Nering (2006) suggest that one common problem of the fit indices is the sensitivity to a large sample size because the fit indices are derived from inferential statistics. Embretson and Reise (2000) further suggest fit indices should be selected based on the motive of the researcher. For example, person fit-statistics might interest a

researcher who wants to select the persons who give different response patterns. Furthermore, several authors have suggested that the best way to evaluate model fit is to provide different fit indices for comparison (Embretson & Reise, 2000; Tay, Meade, & Cao, 2014).

Differential Item Functioning (DIF)

Other than looking at the overall quality of a scale, researchers are also concerned with how a scale may function across subgroups within a sample. Drasgow (1982; 1987) proposed two different types of equivalence that are important for psychological measures: measurement equivalence and relational equivalence. Measurement equivalence is defined, as when individuals who are at equal trait levels should have the same expected scores, independent of the subpopulation that they come from. Using an ability example, a female and a male with high ability on math should both have the same expected scores on a math test. Following the same example, when a female and a male have the same math ability but have different expected scores, it is said that the test failed to provide measurement equivalence and was biased.

Relational equivalence takes into consideration the association between a target measure and another variable of interest (Drasgow, 1982, 1987). The relational equivalence is achieved when the association between a target measurement and another variable of interest is the same across different groups within a sample. For example, researchers might find the

association between math test scores and GPA is low. The measure would achieve relational equivalence when the subgroups (e.g., gender) show the same association between math test scores and GPA. When the association between math test scores and GPA is different for men and women in a sample, then the measure has failed to achieve relational equivalence, which is also labeled differential validity. Furthermore, Embretson and Reise (2000) suggested that differential validity might not be a concern for research if it is incorporated as the primary research question. Therefore, the major focus of this study is to test measurement equivalence instead of relational equivalence.

Testing measurement equivalence is an important topic in both CTT and IRT, and there are many analyses that have been developed to test for measurement equivalence. However, the analyses usually come from the different conceptual reasoning behind CTT versus IRT. The major difference between CTT and IRT analyses is level of focus (Embretson & Reise, 2000; Tay et al., 2014). In CTT, the focus of testing measurement equivalence is the latent construct of the test and the most common method is confirmatory factor analysis (CFA). In IRT, the focus is on individual items and their options. Furthermore, IRT-based measurement equivalence methods have been suggested by researchers rather than CTT-based measurement equivalence methods because the advantage of providing more information by focusing on the item functioning level. Many analyses have been proposed

to test for measurement equivalence in IRT and the overview of the IRT measurement equivalence methods will be provided in the next section.

In IRT, differential item functioning (DIF) is often used to test measurement equivalence across the items and there are many methods proposed. To detect DIF, IRF is used for comparison between groups for each item. When IRFs of an item are the same across the groups, measurement equivalence has been established. However, when IRFs are different across groups on an item, the result suggests that this item shows DIF. For example, researchers might be interested in DIF between men and women for an attitude item. DIF of the item exists when a given level of attitude has different probabilities of choosing a specific response. Different methods are proposed below based on this conceptual framework of testing DIF.

In Tay et al.'s (2014) review article of DIF, they described different types of DIF tests and their general procedures, which I will outline here. The most common types of DIF tests include Lord's chi-square (Lord, 1980), differential item functioning of items and tests (DFIT; Raju, van der Linden, & Fleer, 1995), likelihood ratio (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1993), Mental Haenszel (MH; Holland & Thayer, 1988), simultaneous item bias (SIBTEST; Shealy & Stout, 1993), bivariate residual (BVR), and Wald chi-square (Vermunt & Magidson, 2005). Of these, the three major tests for DIF in IRT are Lord's chi-square, the DFIT, and the likelihood ratio. Specifically, Lord's chi-square has been applied to dichotomous items

instead of polytomous items. The MH and SIBTEST are nonparametric methods. BVR and the Wald chi-square are newer methods that use latent variable modeling to test for DIF. In this study, I am going to discuss DFIT and the likelihood ratio in detail since they are more commonly used in polytomous psychological scales (Collins et al., 2000; Edelen, Mccaffrey, Marshall, & Jaycox, 2009; Flannery et al., 1995; Ryan et al., 2000).

As noted earlier, DIF analysis usually involves the comparison of two groups. The groups are often labeled as the reference group and the focal group. The reference group is the majority group in the sample that has the most participants when compared to the other group in the sample. The focal group is the minority subgroup in the sample that has fewer participants when compared to the majority group. Alternatively, the majority group may have been established on the basis of a key population parameter (e.g., race and gender). When comparing multiple groups using DIF analysis, the majority group is still the reference group, but other groups are labeled as focal groups. In DIF, the reference group is coded as 0 and the focal group is coded as 1.

Once the groups are identified, the next step is to compare if the groups have different IRFs. To detect DIF in general, item parameters are estimated from each group and an item exhibits DIF if the item parameters differ significantly. To make comparisons between IRFs is not simple and this is where different DIF frameworks are implemented. One difficulty with DIF is

that the comparison can only be made if the measurement metric is the same. Likelihood ratios (Thissen et al., 1986; 1993) can be used to determine anchor items to solve the metric problem. In the likelihood ratio method, steps are applied to identify anchor items, which are free of DIF, from the items of interest. Then, anchor items are used to put parameters on a common metric.

Furthermore, a likelihood ratio model approach can be used to detect potential DIF items. Two types of the models are being compared: a compact model and an augmented model. The compact model constrains all item parameters and assumes no differences between the groups. The augmented model, which frees the constraint between the groups of a possible DIF item, is being compared to the compact model. Then, the metric of the anchor item can be applied to both the compact model and the augmented model. Once the compact model and the augmented model are on the same metric of the anchor item, the ratio can be calculated to make the comparison.

DFIT (Raju et al., 1995) solves the metric problem by performing linking. After item parameters are estimated separately for each group, linking (i.e., a linear transformation) is used to equate the item parameters on the same metric. For example, with a group of high ability individuals, the item parameters might show low difficulty and low discriminability. In contrast, with a group of individuals with various levels of ability, the estimated item parameters will show moderate difficulty and high discriminability. When the item has DIF, the accuracy of the result from linking will be affected.

Therefore, iterative linking is proposed to address this problem and makes sure only non-DIF items are linked. Furthermore, DFIT can provide item level and test level DF (i.e., DTF), but the likelihood ratio provides only item level DF.

Several statistics are calculated under the DFIT framework to assess DIF and DTF (Raju et al., 1995). At the item level, the DFIT framework estimates the compensatory DIF (CDIF) and the noncompensatory DIF (NCDIF) statistics. The CDIF estimates the DIF of each item while considering DIF of the rest of the items in a scale. The values of the CDIF can be either positive or negative because CDIF is the difference between two groups' IRFs. In contrast, NCDIF is estimated under the assumption that the rest of the items in a scale have no DIF. The value of NCDIF is the average squared difference between individuals' true scores estimated from the reference group and the focal group. The sum of the values of CDIF for all the items is the value of DTF of the scale. The values of CDIF can be either positive or negative for each item; therefore, it might be possible that the scale with a lot of CDIF items did not show DTF. I note DTF because Collins et al. (2000) suggested that it is important to evaluate differential functioning on a scale level instead of just the item level.

Top Leadership Direction Scale (TLDS)

The construct of interest for this study is top leadership. Although much research has focused on an individual leader, there is precedent to consider

the top echelon as a first hand entity, i.e., top leadership can be considered as a collective body that directs the business of the organization. Furthermore, to better understand leadership, it is important to consider it in the context of follower perceptions (Bennis, 1999; Chaleff, 1995; Meindl, 1990). For most followers, leaders are often a distance away and the perceptions of leaders through the eyes of followers are rather indirect. Furthermore, leadership can be considered as a function of an organization instead of a person or a position (Denis, Lamothe, & Langley, 2001; Denis, Langley, & Rouleau, 2010). Working together, a top management team often performs the essential functions of an organization, and the leadership may be considered a set of shared responsibilities (cf., Michalisin, Karau, & Tangpong, 2004; Pearce, 2004; Wielkiewicz & Stelzner, 2005).

To assess this collective function, Kottke, Pelletier, and Agars (2013) created the top leadership direction scale (TLDS) which will serve as the scale of interest for this study.

Present Study

As already noted, IRT shows several advantages compared to CTT in providing additional psychometric qualities of the scale (Embretson & Reise, 2000). I aim to demonstrate the superiority of IRT compared to CTT based on two analyses in the present study. The first analysis is to evaluate the TLDS using IRT. I believe IRT will provide additional item information and can be used for item reduction for future use of the scale. IRT is model-based

measurement and model selection is based on the nature of data. GRM is the most appropriate model for an attitude scale and will be used in the present study. I will provide several fit indices for model evaluation. The second analysis is to conduct DIF analysis to provide evidence of measurement equivalence across functional groups of employees.

CHAPTER TWO

METHOD

Participants

The data were collected through a larger organizational survey (see Pelletier, Kottke, & Reza, 2015) across the California State University (CSU) system. The survey was conducted to investigate employees' reactions to furloughs that were implemented in the CSU system in 2009. The employees (n = 8046) from various positions at 18 of the 23 CSU campuses responded to the survey. Based on employee positions in the CSU system, employees could be grouped into three broad categories: managerial personnel, support staff, and faculty. Each category can better be described from the nature of its job responsibilities. Managerial personnel provide supervisory duties and monitor the operational departments of the university. Support staff handles mostly clerical duties. In contrast to managerial personnel and support staff, the faculty's focus is in academic-related work with the three major areas being teaching, service, and research. Taken together, different responsibilities related to the position might result in different attitudes toward the top management team. Due to a relatively small sample size of managerial personnel, I investigated the potential DIF between support staff and faculty. Participants' gender, age, and year of employment were also collected as the demographic variables.

Procedure

As part of the larger survey, participants were asked to rate their attitudes on organizational commitment, organizational identity, need for furloughs, turnover intentions, and confidence in the top leadership. Specifically, confidence in the leadership was directed toward the system and campuses. Here, I evaluated the TLDS as applied to the Chancellor's Office (CO) of the CSU because it purports to measure the employees' attitude towards the system leadership. In the CSU system, the CO represents the top leadership team. The furloughs were implemented to address California's public funding shortfalls following the housing crisis. Since the CO administered the furlough process, the perspective of the employees of the CSU system on their evaluation of the CO was especially relevant.

Measure

Top Leadership Direction Scale (TLDS)

The measure that was evaluated with Item Response Theory (IRT) was the TLDS. The 4-item TLDS (Kottke et al., 2013) was created to measure the effectiveness of top leadership through the followers' perceptions in the context of providing guidance of the organization. Specifically, they aimed to measure an employee's confidence in the leadership team to lead its organization to future success. A sample item from the adapted scale is: The mission of the CSU has clearly been spelled out by the Chancellor's Office to CSU employees. This scale consists of four items on a 7-point Likert scale

ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). Higher ratings indicate greater confidence in the direction provided by top leadership. The psychometric properties of the scale were tested through traditional test theory methods in two studies (Kottke et al., 2013). Coefficient alpha for the two studies was found to be .81 and .87, with corrected item-total correlations ranging from .59 to .69. TLDS also showed good convergent validity with organizational coordination, extrinsic satisfaction, and trust in innovation, correlated moderately with centralization, intrinsic satisfaction, organizational commitment, and turnover intentions, and divergent validity with formalization, the Big Five personality factors, organizational citizenship behaviors, and communication from top leadership. Further, I selected this measure because it has not been analyzed previously with IRT. The full scale, with item 3 and item 4 reverse coded, can be found in Appendix A

Design and Analysis

I used IRT to examine the psychometric qualities of TLDS, which had established construct validity and reliability from two previous studies (Kottke et al., 2013), using Classical Test Theory (CTT).

Assumptions

Before conducting the IRT analyses, I conducted a Confirmatory Factor Analysis (CFA) to evaluate the unidimensionality of the scale. M-plus 7 (Muthén & Muthén, 2012) with maximum-likelihood estimation was used for this analysis. I also used EQS 6.1 (Bentler & Wu, 2005) for comparison.

Chi-square, comparative fit index (CFI), and root mean square error of approximation (RMSEA) indices were used to assess the model fit. Different criteria are used to indicate a good fit for each fit index. The ratio of chi-square value to degrees of freedom when less than 2:1 indicates a good fit (Tabachnick & Fidell, 2007). A CFI value of greater than .90 indicates an acceptable fit, and greater than .95 indicates a good fit. RMSEA less than .06 indicates a good fit (Hooper, Coughlan, & Mullen, 2008).

Graded Response Model (GRM)

My IRT analysis followed Flannery et al.'s (1995) procedure for evaluating Likert type scales. All procedural steps were conducted through IRTPRO. The first step was to use the graded response model (GRM; Samejima, 1969) to estimate the difficulty and discrimination estimates of each item. Difficulty parameters are identified by number of response options. For example, a 4-point Likert scale has three difficulty ($b_1 \ldots b_3$) parameters. In a Likert scale, the difficulty parameters represent the 50% probability of endorsing a higher option relative to the prior (ordered) response option. Thus, b_1 represents the endorsement of option 2 relative to option 1, b_2 represents the endorsement differentiating options 2 and 3, and b_3 represents the endorsement differentiating options 3 and 4. In the present study, we used a 7-point Likert scale, so there were six difficulty ($b_1 \ldots b_6$) parameters for each item. The difficulty parameter in IRT usually ranges from -3.00 to +3.00.

Other than the difficulty parameter, GRM also identifies one discrimination parameter (*a*) for each item. The discrimination parameter indicates the change of the individuals' probability of endorsing an item between different trait levels and most often ranges from 0.75 to 2.50 in practice (Flannery et al., 1995). An item with a larger value suggests that the probability of endorsing an item changes dramatically at a certain trait level. High discrimination items are better at differentiating individuals at different trait levels and are said to have better quality than the items with low discriminability.

The second step in the IRT analysis was to graph ICRFs for each item, based on the estimated parameter values (Flannery et al., 1995). The height of the ICRF depends on the value of the item's discrimination parameter. An item with a high ICRF provides more information than an item with a lower information curve. The location of the ICRF depends on the value of the difficulty parameter. An item with a low difficulty value has the ICRF located toward the negative side of the x-axis and vice versa. We can use the location of ICRF to match the item with the most appropriate trait level. ICRFs of each item were evaluated based on the location and the shape.

The third step was to estimate the amount of information provided by each item and could be presented as an IIC (Flannery et al., 1995). As mentioned previously, an IIC is displayed with individual trait levels on the x-axis and the amount of information on the y-axis (see figure 5). Furthermore,

the shape of an IIC is dependent on item discrimination and difficulty parameters. The peak of an IIC shows the highest amount of information associated with the specific trait level. For example, if an item's IIC peaked at the theta value of 2.00, this item is most accurate when the individual has the higher trait level. The amount of information of each item can be added up to provide the total information of the measure, which can be presented as a TIC.

Finally, I evaluated the fit of the data to the GRM. As mentioned previously, all the fit statistics are sensitive to a large sample size and there is no agreement as to the best fit index. The recommended practice is to assess the fit of the model with several fit indices for comparison. IRTPRO provides several fit indices including the M_2 statistic (Maydeu-Olivares, & Joe, 2005, 2006), standardized local dependence (LD) χ^2 (Chen & Thissen, 1997), S- χ^2 statistic (Orlando & Thissen, 2000, 2003), and information criteria. The M_2 statistic assesses the overall model-data fit by comparing differences between the observed responses and modeled responses based on marginal tables. To evaluate the M_2 statistic, Tay et al. (2014) suggested the criteria for a good model-fit are non-significant p-value and a small RMSEA, close to zero. Standardized local dependence (LD) χ^2 examines the local dependence between a pair of items on the scale by calculating the χ^2 between the difference in observed and estimated response frequencies for each pair of items. To evaluate Standardized local dependence (LD) χ^2, Tay et al. (2014)

suggested the value of (LD) χ^2 less than 3 indicates items not locally independent. Standardized local dependence (LD) χ^2 was not reported in the present study because it was assumed for an attitude scale. S- χ^2 statistic assesses the fit for each item from computing the χ^2 between observed and expected responses for the item across categories. An item has a good fit when its calculated p-value associated with S- χ^2 statistic is not significant (Tay et al., 2014). The sample size with this analysis was deemed adequate to recover item response parameters.

Differential Functioning of Items and Tests (DFIT)

My plan was to use a DFIT analysis using the general DFIT analysis procedure from Collins et al., (2000) for an attitude scale. However, DFIT was not feasible because the data had a poor fit to the GRM and no items on the scale could be used to link other items on the same metric. The original steps for conducting DFIT were as follows. In the first step, MULTILOG estimates the item parameters separately for each group using GRM. In the present study, the faculty was the reference group and the support staff was the focal group.

Furthermore, it is necessary to place the item parameters estimated from each group on the same metric, which is called linking or equating, in the second step (Collins et al., 2000). EQUATE is used to perform linking between the reference group and the focal group for each item. Specifically, linking applies a linear transformation for the scale of the second sample to

match the distribution of the first sample. After linking, the item parameters estimated from each sample can be used interchangeably. However, Collins et al. (2000) suggested that linking is inappropriate if an item shows DIF across the groups. Therefore, an iterative linking procedure is proposed to overcome this problem (Candell & Drasgow, 1988).

The general steps of iterative linking are outlined here. First, DIF analysis is used to test the linked item parameters of the item for DIF. Second, items that do not differ significantly will be used to relink the parameters. The process continues for each item until the significant DIF items have been identified.

Finally, DFIT8 produces three statistical indices to test for item DIF. Noncompensatory DIF (NCDIF) tests item level DIF for the target item without taking consideration of possibility of DIF for other items in a scale. Compensatory DIF (CDIF) estimates DIF for the target item while estimating the DIF for the rest of the items in a scale at the same time. The last statistic, differential test function (DTF), tests differential functioning on the scale level. DTF is simply the sum of all item CDIFs.

Multigroup Confirmatory Factor Analysis (CFA) Analysis

Due to the difficulties encountered with DFIT in IRT because of lacking model-data fit, I also analyzed the scales for structural invariance between the groups in the present study. A series of steps were conducted to compare the tested models in multigroup CFA analysis. The first step was to obtain

good-fitting models for each group separately. The two groups were faculty and support staff in this study. The model fit and path coefficients for each group were tested individually. The second step was to test for differences across groups by obtaining the multigroup model, which is the baseline model. In the baseline model, the path coefficients were free to vary across the groups. In the following steps, the constraints were applied to the models across the groups. The third step was to constrain all path coefficients for all the items across the groups. The model fit for the fully constrained model would be compared to the baseline model in order to determine if it was necessary to release any constraint among the paths. Finally, if removing constraints were necessary, then the constraint would be removed item by item. The model fit between the current and the subsequent reduced model estimates would be compared to determine the final model.

CHAPTER THREE

RESULTS

Data Screening

The data set consisted of the responses from a total of 8046 participants. A missing value analysis was conducted to examine the pattern of the missing data (e.g., mismatched responses and missing responses). The results showed that the missing pattern is missing completely at random based on Little's MCAR test (Chi-square = 30.90, p = .157). The missing values were removed listwise from the four TLDS items. After removal of the missing cases listwise, data from 6968 participants were usable. Two of the reverse coded items on the TLDS were recoded within the data, so that the high score means high confidence towards the chancellor's office.

Using the remaining sample, the data were screened for normality, univariate outliers, and multivariate outliers. Based on visual inspection of the histogram, all the variables were positively skewed. However, no transformation was applied because it was expected for an attitude scale. No univariate outliers were detected using the standard of $z > 3.5$ or < -3.5. Mahalanobis distance of each case was calculated to screen for multivariate outliers. Based on p value less then .001 and discontinuity from the distribution of the scores, a total of 17 cases were removed from the sample. The data from 6951 participants were used in this study.

Confirmatory Factor Analysis (CFA)

The dimensionality of the TLDS was tested through a one-factor model of CFA with EQS 6.1 (Bentler & Wu, 2005). Mardia's test of normality was used to test for multivariate normality. Mardia's coefficient was 64.72 with a p value of less than .05, indicated a violation of multivariate normality; therefore, the maximum likelihood estimation with robust standard errors and chi-square are reported for this measurement model. Satorra-Bentler scaled chi-square was used to evaluate the overall model fit, and two other fit indices (i.e., CFI and RMSEA) were also evaluated to provide further evidence of goodness of fit. The initial model was partially supported [Satorra-Bentler χ^2 (N = 6951, 2) = 304.72, $p < .01$, Robust CFI = .96, Robust RMSEA = .15]. The chi-square was significant and the ratio of degrees of freedom to chi-square was greater than 2:1; however, this result was expected with a large sample size. Furthermore, RMSEA has shown to be influenced by small degrees of freedom and sample size; therefore, the result of RMSEA was also expected (Kenny, Kaniskan, & McCoach, 2014). CFI was the most appropriate to this model because it was least sensitive to sample size among all the fit indices (Hooper et al., 2008). One modification was added for the error variance between item 3 and item 4. The final model indicated a good fit, Satorra-Bentler χ^2 (N = 6951, 1) = 9.07, $p = .003$, Robust CFI = 1.00, Robust RMSEA = .03. The results supported the unidimensionality of the TLDS.

Mplus showed comparable results with EQS, so only one set of statistics are reported here.

Classical Test Theory (CTT)

Descriptive item statistics and inter-item correlations for the four TLDS items are presented in Table 2 and Table 3, respectively. The overall mean score of the four items was 3.00 (*SD* = 1.67), with the means ranging between 2.67 to 3.44. The mean scores of each item were slightly lower on a 7-point Likert-typed scale. The mean item-total correlation was 0.62, with the item-total correlations ranging from 0.54 to 0.69. The Cronbach's alpha for the TLDS was 0.80. With the exception of the item means, which are lower in this study, these results were also comparable to the first and second study in Kottke et al. (2013). Specifically, the item *SD*s and item-total correlations were nearly equivalent to the prior studies of the scale.

Graded Response Model (GRM)

Item Response Theory (IRT) Parameter Estimates

Baker (2001) suggested the following magnitudes to examine the value of a_i within the context of ability testing: $a_i > 1.7$ is considered as very high, 1.35-1.69 is high, 0.65-1.34 is moderate, 0.25-0.63 is low, and 0.01-0.24 is very low. Since there are no real standards yet in the context of attitude measurement, the similarities of the a_i parameter in IRT make it possible to adapt Baker's (2001) suggestions to evaluate a_i in the present study. The

parameter estimates of each item are presented in Table 4. The item discrimination parameters (a_i) ranged from 1.71 to 4.01, with a mean SE of 0.07, suggesting that all items had very high discrimination. In comparing the items, item 2 has the highest discriminability, and item 4 has the lowest discrimination values with 2.3 units different in their *a* values.

Furthermore, a_i should be interpreted in the context of the latent trait parameters of the item. The threshold parameters (b_i) ranged from -1.16 to 2.71, with a mean *SE* = 0.03. Based on the range of the b_i parameter of the four items, the TLDS seemed to measure more accurately with participants at the higher ranges of trait levels. TLDS might be more problematic to measure the participants at the lower ranges of trait levels. The SEs were small across the trait level, which suggested that the parameter estimates were accurate. Item 1's b_i parameters range from -1.16 to 2.28, with a mean *SE* = 0.03. For item 2, the b_i parameters range from -0.43 to 2.20, with a mean *SE* = 0.03. Item 3 had the b_i parameters range from -0.96 to 2.19, with a mean *SE* = 0.03. Item 4's b_i parameters range from -0.86 to 2.71, with a mean *SE* = 0.04.

Item Category Response Functions (ICRFs)

The next step was to examine each item visually through their ICRFs (see figure 8). Based upon the visual inspection of ICRFs for four items, all the items seemed to be centered slightly to the positive side of the trait level. Item 2 showed the least amount of overlap between each category compared to the other three items. Item 1 and item 3 showed similar overlap between their

adjacent ICRFs. Item 4 showed the most overlap between its adjacent categories. To conclude, the graphs of each item's ICRF indicated that item 2 appeared to be the best item among the four items in the TLDS in terms of the lack of overlap with the adjacent categories.

Item Information Curves (IICs)

Once the parameters for each item were estimated, the amount of information provided by each item could be calculated as IIC. A high value of IIC indicates greater information. The SEs can then be calculated by inverse square root of the information value. See Figure 9 for IIC graphs for the four TLDS items and Table 5 for information values and corresponding SEs across the trait levels for each item. Based on visual inspection of the IIC graphs, item 2 contributed the most information to the TLDS with the highest amount of information that ranged from the trait level of -.4 to 2.4. The peak of IIC for item 2 had an information value of 4.86 with the corresponding SE of .45 at the trait level of .4. Similarly, both item 1 and item 3 contributed moderate amount of information. The IIC of item 1 had the highest amount of information with the peak of the information value of 1.75 with the corresponding SE of .76 at the trait level of .0 ranged from the trait level of -1.2 to 2.4. The IIC of item 3 had the highest amount of information with the peak of the information value of 1.48 with corresponding SE of .82 at the trait level of .0 ranged from the trait level of -.8 to 2.0. Item 4 contributed little information for the TLDS across the trait continuum as indicated by the flat

shape of its IIC. SEs were not constant across the trait level and it was the lowest at the highest range of information provided for each item.

Test Information Curve (TIC)

The test information curves for the four items on the TLDS are presented in Figure 10. The TIC showed that the TLDS provided the most information between the trait ranges of -.4 to 2.0 with the peak of the information value of 9.99 at the trait level of .4. The result suggested that TLDS as a scale provides the most accurate information with individuals who have a mildly positive attitude toward their top leaders.

Fit Statistics

Finally, IRTPro provides χ^2 and M_2 statistics (Maydeu-Olivares & Joe, 2005, 2006) for the overall model fit of GRM. Tay et al. (2015) suggests that not significant p-values (> .05) and RMSEAs closest to zero indicate a good fit. χ^2 was significant: χ^2 (N = 6951, 2372) = 1822075.33, p < .001, RMSEA = .33. M_2 statistics were also significant: M_2 (N = 6951, 212) = 24205.16, p < .001, RMSEA = .13. Furthermore, IRTPro provides S- χ^2 (Orlando & Thissen, 2000, 2003) to assess the model fit at the item level. See Table 6 for S- χ^2 for each item. Tay et al. (2015) suggests that nonsignificant S- χ^2 values (p > .05) indicate a good fit. As with all other fit indices, there is some debate as to the value of the χ^2 statistic to assess fit (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Reise, Widaman, & Pugh, 1993).

However, the results from all the fit indices indicated that the data are not a good fit to the GRM.

Differential Item Functioning (DIF) Analysis

IRTPro utilizes Wald tests, which were modified from Lord (1977), to detect DIF for items. I conducted DIF analysis in IRTPro due to software difficulty for DFIT. Furthermore, several studies had recommended the advantages of using IRTPro to detect DIF and provided step-by-step illustrations for the procedures (Meade, 2010; Tay et al., 2014). For example, anchor items need to be identified prior to DIF analysis which often uses different software. However, the anchor item can easily be identified using IRTPro by following Tay et al.'s (2014) procedure. Parameters were estimated using GRM for each group (see Table 7). The fit-indices were evaluated based on the same criteria described in the GRM section. The overall model fit was not supported for either group: faculty, M_2 (N = 2068, 212) = 4040.66, $p < .001$, RMSEA = .09; support staff, M_2 (N = 2481, 212) = 10361.74, $p < .001$, RMSEA = .14. See Table 8 for S- χ^2 (Orlando & Thissen, 2000, 2003) for each item separating by groups. All the items in the faculty group and support staff group showed significant differences in their response frequencies when examining their S- χ^2 statistics.

The first model was estimated with the parameters free to vary across the groups and all items were tested for DIF as on the same metric. This step was used to identify an initial set of items with significant DIF. The model did

not fit the data well, M_2 (N = 4549, 424) = 13467.13, p < .001, RMSEA = .09. Furthermore, DIF diagnostic tests suggested that all four items exhibited DIF and none of the items were identified as a potential anchor item. Figure 11 presents each item's ICRF graphics by each group. Based on a visual inspection of each item's ICRFs across the groups, it appeared that item 2 had the most similar ICRFs across the two groups.

Tay et al. (2015) suggests a two-step iterative procedure to identify all the anchor items from the scale. The initial anchor items identified from the DIF diagnostic tests would be used to identify more anchor items in the next model. Once the next model was estimated, the items that did not show significant DIF would be added to the initial anchor items. This procedure ends when all the items remained show significant DIF. In my analysis, all the items were identified as significant DIF items, so the two-step iterative procedure was not appropriate. Since there were only four items on my scale, I tried anchoring one item at a time to see if the model fit increased. The resulting model fit did not change for any of the items used as an anchor item: item 1, M_2 (N = 4549, 424) = 13494.88, p < .001, RMSEA = .08; item 2, M_2 (N = 4549, 424) = 14144.00, p < .001, RMSEA = .08; item 3, M_2 (N = 4549, 424) = 13361.98, p < .001, RMSEA = .08; item 4, M_2 (N = 4549, 424) = 13713.76, p < .001, RMSEA = .08. To conclude, the DIF analysis suggests that all four items exhibit significant DIF across the two

groups; however, the DIF procedure might not be appropriate due to the problem with the model fit of GRM.

Data Screening for Multigroup Analysis

Because the data did not provide a good fit to the GRM, I analyzed the scales for structural invariance between the groups. Before conducting multiple group analysis, the data were examined with the grouping variable: positions. The data set consists of the responses from a total number of 6951 participants. A missing value analysis was conducted to examine the pattern of the missing data (e.g., mismatched responses and missing responses) for the grouping variable. The results showed that the missing pattern was missing completely at random based on Little's MCAR test ($\chi^2 = .48$, $p = .976$). The missing values were removed listwise from the grouping items. After removal of the missing cases listwise, data from 5973 participants were usable. Due to a small sample size of the other positions in the sample, the usable data were further reduced to two major position groups (i.e., a total number of 2068 faculty members and 2481 support staff) for the following multigroup analysis. See Table 9 for descriptive statistics for the faculty and support staff samples.

Multigroup Confirmatory Factor Analysis (CFA) Analysis

Because the DIF procedure was inappropriate for these data, a multigroup CFA was conducted to evaluate the scale for structural invariance.

Multigroup CFA analysis with maximum likelihood estimation was conducted with EQS 6.1 (Bentler & Wu, 2005). The models were evaluated using the same criteria as mentioned previously in the CFA session: Satorra-Bentler scaled chi-square due to nonnormality of data, CFI, and RMSEA.

The assumption of multivariate normality was tested for each group. Mardia's coefficient was 44.45 for the faculty group and 37.33 for the support staff group ($p < .05$), indicated a violation of multivariate normality; therefore, the maximum likelihood estimation with robust standard errors and Satorra-Bentler scaled chi-square was reported. The measurement model was estimated for each group separately. The models indicated a good fit: faculty group, Satorra-Bentler χ^2 ($N = 2068$, 1) = 3.55, $p = .06$, Robust CFI = 1.00, Robust RMSEA = .04; support staff group, Satorra-Bentler χ^2 ($N = 2481$, 1) = 2.22, $p = .14$, Robust CFI = 1.00, Robust RMSEA = .02. See Table 10 for the path coefficients of variables in each model and Table 10 for the residuals of variables in each model.

The baseline model was estimated with the path coefficients free to vary across the two groups. The model fit was good, Satorra-Bentler χ^2 ($N = 4549$, 2) = 5.76, $p = .06$, Robust CFI = 1.00, Robust RMSEA = .03. Next, the fully constrained model (i.e., all path coefficients were constrained to be equal across the groups) was estimated to see if the measurement structure of the latent construct was the same across the groups. The model was partially supported, Satorra-Bentler χ^2 ($N = 4549$, 6) = 62.10, $p < .001$,

Robust CFI = .99, Robust RMSEA = .06. To compare the models, a chi-square difference test was computed with a scaling correction. The chi-square difference test indicated that the fully constrained model was degraded greatly compared to the baseline model, Satorra-Bentler $\chi^2_{difference}(4) = 58.29$, $p < .001$. This result suggests that some paths were different across the two positions and further investigation was necessary to discover the different paths. Once the different paths were identified, the path was removed one at a time and the model fit between the previous model and the current model compared.

The Lagrange multiplier multivariate tests were used to identify the paths that differed across two position groups. The path coefficients for item 2 were weaker for the faculty than for the support staff (standardized coefficient faculty = .83, standardized coefficient support staff = .88). Therefore, the paths were released for item 2 in the next model: fit for this reduced model was Satorra-Bentler χ^2 (N = 4549, 5) = 17.10, p = .004, Robust CFI = 1.00, Robust RMSEA = .03. The model (i.e., current model) then was tested against the fully constrained model (i.e., previous model). The chi-square difference test indicated that the current model was improved compared to the fully constrained model, Satorra-Bentler $\chi^2_{difference}(1) = 49.58$, $p < .001$. Furthermore, the constraint for item 4 was removed for the next model.

This model was also supported by fit indices: Satorra-Bentler χ^2 (N = 4549, 4) = 7.98, p = .092, Robust CFI = 1.00, Robust RMSEA = .02.

The path coefficients for item 4 were weaker for the faculty than for the support staff (standardized coefficient faculty = .35, standardized coefficient support staff = .52). The model was tested against the previous model with the item 2 path coefficients released. The chi-square difference test indicated that the current model was improved compared to the fully constrained model, Satorra-Bentler $\chi^2_{difference}(1) = 10.94$, $p < .001$. Finally, removing the constraint for the next item (i.e., item 3) did not result in significant model improvement, Satorra-Bentler $\chi^2_{difference}(1) = .40$, $p > .05$. The model was partially supported: Satorra-Bentler χ^2 ($N = 4549$, 3) = 6.38, $p = .095$, Robust CFI = 1.00, Robust RMSEA = .02. The final model constrained item 1 and item 3 and left item 2 and item 4 free to vary between the groups. To conclude, item 1 and item 3 showed measurement invariance across faculty and support staff. Item 2 and item 4 seemed to show different estimation depending on which group the participants belonged.

CHAPTER FOUR

DISCUSSION

In Classical Test Theory (CTT), the psychometric properties of a scale are being evaluated through the reliability coefficients of the scale. To obtain the psychometric properties in the item level of a scale, researchers usually examine the mean, standard deviation, and item-total correlations of each item. In Item Response Theory (IRT), the mean is most closely represented by the difficulty parameters and item-total correlations are most closely presented by discrimination parameters. To make decisions about keeping the items on a scale, CTT researchers usually inspect the improvement of reliability after deleting the item. However, reliability of the scale is not a good indicator to eliminate items in a relatively short scale due to the fact that the equation of reliability takes into account the number of items on the scale. Thus, in CTT, the longer the scale, the higher the reliability when the items are relevant to the construct of interest. Except for the limitation on reliability, Embretson and Reise (2000) had listed many advantages that IRT had over CTT. Several advantages were demonstrated in the present study with comparisons between CTT statistics and IRT statistics in Top Leadership Direction Scale (TLDS). Furthermore, I will compare the conclusions about the quality of each item as derived from CTT and IRT.

Based on CTT statistics, the TLDS has good reliability and deleting any item from the scale would result in lower reliability of the scale. On the item

level, the mean scores of each item showed that participants endorsed an attitude toward the chancellor's office that was less than the midpoint of the scale. The results indicated that the scale had slightly lower difficulty levels and it might be hard to discriminate among participants who were at the lower level of the trait continuum. Based on item-total correlations, item 2 had the highest discriminability and item 4 had the lowest discriminability. However, both items were highly discriminating items. The standard deviations were small and equivalent across the items. To conclude from the CTT analysis, all items seemed to be of good quality and all items should be retained.

In IRT statistics, the psychometric qualities of each item were provided from discrimination parameter, difficulty parameter, standard error, and item information indices at each trait level, as well as graphically. Considering the possible range of the difficulty parameter was from -3 to 3, the overall difficulty parameter of each item seemed to span higher in the positive end compared to the negative end. Based on the results from the difficulty parameters, the scale might be more problematic to distinguish the participants who had lower trait levels, specifically, the trait level lower than -1.50. Furthermore, discriminability parameters suggested that all items had good discriminability with item 2 as the most discriminating item and item 4 as the least discriminating item within the scale. To further investigate the differences between item 2 and item 4, it was necessary to evaluate the nature the items. In the context of measuring followers' confidence toward their top

management team, item 2 is more specific in wording when compared to item 4. For example, item 2 uses “inspire confidence” compared to item 4, which uses “direction.” Furthermore, item 2 indicates the outcome of good management by using the phrase, “future success,” but there was no specific outcome indicated in item 4. Therefore, the level of specificity and clarification in the wording might have led to differences in psychometric qualities between these two items. Thus, IRT seemed to provide similar results in comparison to CTT; however, it is important to note that IRT provided more specific information regarding the items by providing the location at the trait level. Except for discrimination and difficulty parameters, IRT provided additional information for each item graphically from each item’s Item Category Response Functions (ICRFs) and Item Information Curve (IIC), relative to CTT.

The shape of the ICRFs can be compared to the standard Graded Response Model (GRM) ICRFs’ shapes to provide the quality of the items based on shape and location. These results suggested that item 2 resembles most closely the standard shape of a GRM ICRF (see Figure 6) and the shape of the ICRF for item 4 seemed to be truncated at the positive end of the trait continuum. To further examine each item, I inspected IICs for each item. IICs indicated that item 2 contributed the most information to the scale and item 4 did not contribute much to the scale at all. Items 1 and 3 each contributed a moderate amount of information to the scale. The actual difference between

the amount of information provided by item 2 and item 4 were compared using their peak information values on the trait continuum. Item 2 had the highest information value, close to 5, and item 4's highest information was close to 1. Furthermore, an item's information value should be examined in combination with its standard error. Item 2 seemed to measure accurately at the trait level with the highest amount of information values; however, examining the standard error of the lowest trait level of item 2 suggested inaccuracy when measuring participants at extremely low trait levels. When investigating the nature of the items, item 2 is the only item that specifies an outcome from the leader's actions when compared to the other items in TLDS. The level of specificity can lead to a narrower range of measurement accuracy across the trait continuum compared to other three items. This result suggested that item 2 should best be used with item 1 or item 3 to ensure the measurement accuracy at the low level of trait because item 1 and item 3 covered a broader range of trait levels, when compared to item 2. That is, a shorter, two-item Top Leadership Direction Scale (TLDS),that can conveniently be included in the survey would be possible based on the recommendation from IRT. To conclude from the IRT analysis and the nature of the item, it is best to remove or reword item 4 from the scale because of the overall low quality of this item.

Embretson and Reise (2000) proposed many advantages of IRT over CTT. I further demonstrated the advantages of IRT over CTT in the present GRM analysis. First, IRT provides more information about each item when

taking into account the specific trait level associated with the item parameters. Second, IRT produces the item parameters graphically, which allows researchers to examine the items visually. Third, IRT provides information regarding each item contribution to the overall scale across the trait continuum. Finally, standard errors can be calculated from the information values, which in turn provide the accuracy of the item quality at the specific trait level. In conclusion, IRT provides much more flexibility and information at the item level when compared to CTT.

Possible Alternative Item Response Theory (IRT) Model Based on Likert versus Thurstone Scaling Approaches

IRTPro provides several fit indices to evaluate the overall model fit and the goodness of fit at the item level. However, these data did not fit the GRM (Samejima, 1969) for either fit index. Prior research has made it evident that GRM is the most appropriate IRT model for Likert-typed data and many studies had made comparisons between GRM and other models (e.g., Baker et al., 2000; Maydeu-Olivares et al., 1994; Maydeu-Olivares, 2005). The results from those studies consistently revealed that GRM provided a good fit to Likert-type data. To further investigate the possible alternative models that might explain my result of model misfit, I discovered that there was an alternative model that might also provide model fit to attitude scales. This alternative model is called the Generalized Graded Unfolding Model (GGUM;

Roberts, Donoghue, & Laughlin, 2000), which was developed based on different assumptions as compared to GRM.

The difference in the assumptions of these two different types of models comes from the two traditional methodological approaches to measure attitude. The assumption of GRM (Samejima, 1969) is derived from Likert (1932) scaling, which suggests the probability of endorsement increases monotonically as the trait level increases. The model assumptions based on Likert scaling are called dominance models (e.g., 2PL, GRM, and NRM). However, the assumption of GGUM (Roberts et al., 2000) is derived from Thurstone's (1928) approach to measurement and labeled an ideal point model. The model has the key assumption that the probability of endorsement is highest when the trait matches the difficulty level of the item. The difference between the Likert (1932) and Thurstone (1928) approaches to measurement should be explained with regard to how scales are developed from these two approaches.

In the Likert (1932) approach to construct an attitude scale, a large number of items are developed to measure the underlying construct with either positively or negatively worded items. The participants then rate the items on a Likert-type scale with disagree on one end and agree on the other end. The items are examined with item-total correlations and other methods (e.g., alpha if item deleted) to decide which items should be eliminated from the scale. The final scale usually contains items with high item-total

correlations and high overall reliability (Cronbach's alpha) of the scale. In Thurstone's (1928) approach, the initial items are written with the intention to cover a wide range of attitude options toward the specific attitude (e.g., negative, moderately negative, neutral, moderately positive, and positive). The participants then rate each item on how much they agree or disagree with the degree of attitude the item represents. The final set of items are believed to cover a wide range of the attitude continuum of the target construct. Therefore, Likert's (1932) approach has best been described as a dominance process (Coombs, 1964), which suggests the endorsement increases monotonically as individual's trait increases. In contrast, Thurstone's (1928) approach is best described as an ideal point process (Coombs, 1964), which suggests that endorsement increases as a degree of how much the person believes that the item best reflects his or her own attitude. See Figure 12 for example IRFs for the dominance and ideal point models.

Roberts, Laughlin, and Wedell (1999) provided the evidence for attitude scales that showed characteristics that were consistent with the ideal person process. They invited a group of participants to rate a 10-item scale that measured attitudes toward abortion. The degree of attitude toward each item was obtained through the Thurstone (1928) procedure. The participants' attitudes towards abortion were determined from their median score of agreement on the items of the scale. The researchers then sorted the participants who had similar attitudes into the same group. The ICRF of each

item was presented visually. They found that ICRFs for extreme positive and negative attitude items increased and decreased monotonically as suggested by Likert (1932). However, moderate or neutral-level attitude items showed the highest endorsement when the trait best matched the participant's own attitude towards abortion. The participants who were at high or low levels of attitude towards abortion did not endorse the item in similar fashion to moderate or neutral items. Specifically, as the participants' attitude moved further away from the degree of attitude expressed within an item, the endorsement decreased monotonically. This example showed that monotonically increasing or decreasing ICRFs only resulted for extremely positive or negative attitude items. That is, the prediction of probability of endorsement of extreme attitude items will be the same from a dominance model and an ideal point model.

Understanding the similarities and differences in model predictions between a dominance model and an ideal point model is important because many studies showed a good fit when applying the dominance models to attitude scales (Baker et al., 2000; Maydeu-Olivares et al., 1994; Maydeu-Olivares, 2005). This difference in prediction becomes greater as the item's attitude becomes more moderate or more individuals have extreme attitudes. The greatest difference in prediction between the dominance model and ideal point model happens for items that represent a neutral attitude. Therefore, a scale with a lot of moderate or neutral attitude items is going to

have a better fit when using an ideal point model than when using a dominance model. However, an ideal point model or a dominance model will both perform well for a scale with mostly extreme attitude items or if most individuals hold moderate attitudes. This will result in the popularity of recommending the dominance model to attitude scales.

Based on the theoretical rationale of the ideal point model, it is clear that the ideal point model might show advantages over the dominance model when applied to attitude scales. However, there are no absolute criteria to evaluate which attitude scale would be most likely to benefit from an ideal point model instead of a dominance model since the dominance model will probably perform equally well under most of the situations in which researchers find themselves. To argue that the TLDS will fit better using an ideal point model instead of a dominance model (e.g., GRM), it would be important to examine the existence of extreme attitude participants. The TLDS data used here were for an attitude scale that intended to measure the employees' confidence towards the Chancellor's office after implementation of furloughs, which was one of the financial strategies to deal with the state budget cuts. Specifically, furloughs gave CSU employees two options: either accept the pay cut indirectly by taking 2 workday off or layoffs. Many employees suggested negotiating the implementation of furloughs, but furloughs were implemented as planned by the Chancellor's office without delay. In Pelletier, Kottke, and Reza's (2015) study, they found evidence that

employees showed differences in attachment to their organizations after the institution of furloughs, including a sense of belonging and needs not being fulfilled. Furthermore, the Chancellor's office played an important role in this process. Therefore, I argue that there might be a lot of extreme attitude participants in my current sample, thus making an ideal point model a possibility.

The second criterion to assess with the TLDS is the level of attitude contained under items on the scale. Roberts et al. (1999) demonstrated that the Likert (1932) approach would lead to the moderately extreme attitude items being identified as the best items on the scale, based on a simulation study. In their study, the degree of attitude of each item was first identified using Thurstone's (1928) approach. Then, factor analysis was applied to a set of items, and the final items were selected based on the largest factor loadings. The final scale had an alpha coefficient of .96, and all items had over .70 item-total correlations. The final set of items on the scale did not have extreme attitude items. They further concluded that identifying extreme attitude items might not be feasible when using item analysis and the extreme attitude items might set limitations on measuring individuals across the trait continuum. Since the TLDS was developed using the Likert (1932) approach, the items' trait attitude could not be determined for this study. However, after a visual inspection of TLDS scale items, the items did not seem to be very extreme. For example, one of the items on the TLDS is "there is very little

leadership from the Chancellor's Office," and for Thurstone scaling, it might be possible to change this item to a more extreme item by rewording it to state, "there is no leadership from the Chancellor's Office."

Finally, Roberts et al. (1999) provided an example to explain the difference between applying an ideal point model (i.e., GGUM) and a dominance model (i.e., GRM) to a moderate attitude item graphically. They superimposed the theoretical ICRFs from a dominance model and an ideal point model of a slightly positive item on the same graph for demonstration (see Figure 13). The ICRFs from the dominance model and an ideal point model showed a lot of overlap on the monotonically increasing region, and the divergent point on the two ICRFs was on the most extreme region. That is, the two models make similar predictions on the probability of endorsement for individuals who had extremely negative attitudes to slightly positive attitudes. However, the prediction differs for individuals who had extremely positive attitudes. The dominance models would predict that the individuals who had extremely positive attitudes would show the highest endorsement of this item, but the ideal point models would predict the level of endorsement decreases because a slightly positive attitude item would not match the actual attitude of the individual. The ideal point model proposes that the probability of endorsement starts decreasing monotonically as the difference between the individual's attitude and item's attitude gets greater.

In sum, Thurstone's (1928) approach takes into account the item's own trait level in addition to a person's trait level. Based on different approaches to scale development (i.e., Likert or Thurstone), the models can be categorized into either an ideal point model or a dominance model. Ideal point models suggest that the highest endorsement of the item option happens when a person's trait matches the item's trait. As the person's trait deviates from the item's trait, the endorsement of the item decreases monotonically. Dominance models suggest that as long as a person's trait increases, the response increases monotonically. The ideal point models and the dominance models showed similar predictions for extreme attitude items because it accommodates the person who has the most extreme attitude. As the item's attitude gets less extreme or more individuals have extreme attitudes than the items' trait attitude level, the difference in prediction between these two models becomes more apparent. The trait attitude for the items is usually unknown since most of our scales have been developed using the Likert approach. Therefore, there is no definitive way to suggest under which situation an ideal point model might be a better model than a dominance model for the data. In the situation of TLDS, it is evident that individuals with strong attitudes exist in this dataset. Therefore, an ideal point model (i.e., GGUM) might provide a better fit to the data compared to a dominance model (i.e., GRM).

Differential Item Functioning (DIF) Analysis

The first step of the DIF analysis is to evaluate the model fit for the overall scale and the model fit of the groups. Tay et al. (2014) recommended evaluating the model fit before proceeding to the next step of the DIF analysis. The data did not fit the GRM well, which was evident from my GRM analysis. Furthermore, the essential step to conduct DIF analysis is to identify the items that do not function differently across the groups, and using those non-DIF items to either link or anchor other items on the same metric. This step was not successful because none of the items from the TLDS were identified as possible anchor items. The statistics from the DIF items can only be obtained after successfully placing the items on the same metric for comparison. Therefore, the results from the DIF analysis in IRT were not interpretable.

Structural and Measurement Invariance

To further investigate the group differences in item level, I conducted a multigroup CFA analysis to investigate the structural difference for the four items of the TLDS. The initial model fit was good for the overall and across the groups. Then, the fully constrained model was estimated to test for possible measurement structure differences across the groups. The constraint was released one at the time based on the recommendation of the Lagrange multiplier multivariate tests and the models were compared to see if the fit was significantly worse than the previously constrained model. The model showed that item 1 and item 3 were invariant across faculty and support staff. Items 2

and 4 showed structural differences depending on which group to which they belonged. That is, item 2 and item 4 might have different meaning for faculty and support staff in the context of their attitudes toward leadership, specifically to the Chancellor's office. Furthermore, faculty members are highly educated and because of the nature of their work (i.e., class schedules are their primary fixed schedule), they have more freedom in deciding their own schedule compared to most workers. Support staff members are more likely to have their daily tasks assigned to them with someone supervising their work. Therefore, the different nature of faculty and support staff work might lead to a different understanding of leadership. Specifically, item 4 speaks to the uncertainty of the Chancellor's office in giving directions, which may not be considered as relevant for faculty because they are not under supervision in the way of taking orders from others.

Measurement equivalence can be tested through CFA or IRT procedures, but the choice of CFA or IRT procedures is often dependent on the purpose of the research (Tay et al., 2014). Researchers interested in group differences at the *construct* level often chose CFA. The CFA measurement equivalence results can also provide an understanding of observed differences between the groups. Furthermore, the results from the CFA analysis can provide meaning to the construct by testing each step separately (e.g., configural, metric, scalar, and residual invariances). In contrast, researchers who test measurement equivalence using IRT are

interested in how *items* might function differently across groups. The results from the IRT measurement equivalence are emphasized if the items functions differently across groups. If the items show measurement inequivalent across the groups, it is suggested that the item wording might be biased against a certain group. This difference in usage between CFA and IRT measurement comes from the fact that IRT is most often used in the ability testing context. Therefore, it was not appropriate to compare the results from my DIF analysis to the multigroup CFA analysis in this context.

In sum, I would argue that IRT is superior over CTT because IRT provides more information on psychometric properties of each item and allows researchers to investigate the nature of the item based on the information. I provided the evidence that IRT allows researchers to examine the items graphically and take into account the person as well. Furthermore, the similarities and differences for the items on a scale are easier to identify based on IRT instead of CTT. For example, based on the result of information provided from each item, item 2 was identified to contribute a lot more information to the scale compared to the rest of the items. When examining the item content, I found that item 2 is the most specific compared to the other items. These results were contradicted with the results from CTT because it suggested that all the items should remain as they are. It was evident that IRT should be utilized more to evaluate psychological scales in the future compared to CTT.

Limitations

It was unexpected that the data would show a poor fit to the GRM and DFIT was not feasible because no items could be used to link other items on the same metric. Furthermore, good model fit was recommended before proceeding to a group analysis (Tay et al., 2015). This suggested that there might be some limitations when applying GRM to attitude scales. In general, researchers suggested that the model fit might be problematic if the sample size is too small; however, sample size was not the problem in this dataset. One possible reason might be that the sample size to number of item ratio caused the misfit of GRM. That is, the sample size of the furloughs data may have been too large for a four-item scale.

Practical and Theoretical Implications

As noted earlier, the old rules of measurement (CTT) have meant that a longer scale was more reliable than a shorter scale; however, IRT brings a different perspective. It was possible to evaluate a short and reliable scale based on CTT, using an IRT approach. Theoretically, IRT analysis would provide better psychometric information compared to CTT. No study had investigated the TLDS using IRT, nor DIF. Furthermore, the current study suggested that GRM has some limitations of its usage for attitude scales, the alternative model (GGUM) might be considered for IRT researchers specifically regarding to the attitude scale. Practically, the current study provided the IRT analysis of the TLDS at the item level and the result of the

IRT analysis showed the possible combinations when using only partial items (i.e., item 2 with either item 1 or item 3) on the TLDS in the future.

Obtaining the psychometric properties for the items is the first step to evaluate items. To interpret the psychometric properties of each item, it is important to examine the nature of the items. Item 4 consistently performed poorly compared to other items in the sense of making a contribution to the overall scale and showing measurement inequivalent across the groups. When examining the content of item 4, I found that item 4 is worded differently in terms of specificity and clarity compared to other items on the TLDS. I recommended a revision for item 4 by making it more specific in wording or removing item 4 from the scale. A possible rewording to item 4 could focus on making the “direction” in item 4 more specific or changing “employees are unsure” to “employees do not know” to make item 4 more clear in wording.

As I suggested that there might be a limitation for GRM when the ratio between item length and the sample size is too great, a possible solution is to write more items for the TLDS. For example, to measure followers’ confidence toward the top management team, it might be important to have items that include culture and climate concepts. To further expand TLDS, the researchers might want to write items regarding to different level of the top management team (e.g., direct supervisors vs. middle management, and top management).

Conclusion

IRT provides more information of each item on the scale, when compared to CTT. Based on the information being evaluated, researchers might draw a different conclusion regarding the items on the scale. In the current study, the conclusion was a deletion or revision of one item and a possible shorter two item scale. Based on the conclusion from IRT, the item contents were examined further based on the psychometric properties of the item. That is, the psychometric properties of the items obtained from IRT can be used to understand and refine the items. This point was clearly demonstrated in the present study. Even though IRT is a flexible theory, there are some difficulties when using IRT compared to CTT. First, it takes expert judgment to decide which IRT model to use. Second, the sample size requirement of IRT may require more time for data collection. Finally, as noted earlier, the software for IRT is highly specialized and does not uniformly provide all the models and fit statistics for comparison. In the future, IRT researchers should try to make IRT more accessible for the researcher who is not familiar with IRT methods by sharing information or procedures to conduct the IRT analysis.

APPENDIX A

TOP LEADERSHIP DIRECTION SCALE

Top Leadership Direction Scale

1. The mission of the CSU has clearly been spelled out by the Chancellor's Office to CSU employees.
2. The Chancellor's Office inspires confidence in the future success of the CSU.
3. There is very little leadership from the Chancellor's Office.
4. Employees are unsure about the direction the Chancellor's Office is going.

Kottke, J. L., Pelletier, K. L., & Agars, M. D. (2013). Measuring follower confidence in top leadership direction. *Leadership & Organization Development Journal*, *34*(4), 292-307. doi:10.1108/LODJ-07-2011-0062

APPENDIX B

TABLES

Table 1. Ten “New” Rules (Embretson & Reise, 2000)

1.	Standard error of measurement can be applied to each score and can be generalized across the test.
2.	The length of the test is no longer a concern because the reliability can be better for short tests.
3.	Make comparisons of test scores across multiple forms with different difficulty levels are ideal.
4.	Unrepresentative samples can produce unbiased results.
5.	Test scores have meaning by placing the scores on a continuum of difficulty of items.
6.	Normal distribution is no longer needed because specific model is applied based on the nature of the data.
7.	Mixed response formats can produce ideal results.
8.	Equating score difference can provide meaningful results when the questions are at different difficulty levels.
9.	Take into account full information of the data without adjustments when factor analysis binary items.
10.	Item stimulus features can be the components to evaluate the psychometric properties of the items.

Table 2. Descriptive Statistics of Mean, Standard Deviation, and Item-Total Correlation

	Item	Mean	SD	Item-total r
1.	The mission of the CSU has clearly been spelled out by the Chancellor's Office to CSU employees.	3.44	1.73	.61
2.	The Chancellor's Office inspires confidence in the future success of the CSU.	2.70	1.65	.69
3.	There is very little leadership from the Chancellor's Office. (R)	3.20	1.75	.63
4.	Employees are unsure about the direction the Chancellor's Office is going. (R)	2.67	1.55	.54

Table 3. Inter-Item Correlations

	Item 1	Item 2	Item 3	Item 4
Item 1	--	.66	.45	.38
Item 2.	.66	--	.55	.43
Item 3	.45	.55	--	.55
Item 4	.38	.43	.55	--

Table 4. Graded Response Model Item Parameter Estimates

Item	***a***	b_1	b_2	b_3	b_4	b_5	b_6
1	2.36	-1.16	-0.38	0.10	0.70	1.34	2.28
(SEs) *item1*	(0.05)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.05)
2	4.01	-0.43	0.13	0.53	1.06	1.61	2.20
(SEs) *item2*	(0.13)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
3	2.17	-0.96	-0.22	0.33	1.03	1.50	2.19
(SEs) *item3*	(0.05)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
4	1.71	-0.86	0.22	0.97	1.72	2.10	2.71
(SEs) *item4*	(0.04)	(0.03)	(0.02)	(0.03)	(0.04)	(0.05)	(0.06)

Note. *a* is the discrimination parameter. *b* is the difficulty parameter.

Table 5. Information Values of each Item

	θ:														
Item	-2.8	-2.4	-2.0	-1.6	-1.2	-0.8	-0.4	0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8
1	0.11	0.27	0.60	1.09	1.51	1.65	1.72	1.75	1.74	1.73	1.69	1.61	1.56	1.43	0.98
SEs	*3.02*	*1.92*	*1.29*	*0.96*	*0.81*	*0.78*	*0.76*	*0.76*	*0.76*	*0.76*	*077*	*0.79*	*0.80*	*0.84*	*1.01*
2	0.00	0.01	0.03	0.15	0.68	2.46	4.40	4.66	4.86	4.63	4.63	4.64	4.45	3.49	1.22
SEs	*n/a*	*10.0*	*5.77*	*2.58*	*1.21*	*0.63*	*0.48*	*0.46*	*0.45*	*0.46*	*0.46*	*0.46*	*0.47*	*0.53*	*0.91*
3	0.08	0.19	0.40	0.76	1.15	1.38	1.45	1.48	1.48	1.47	1.49	1.47	1.40	1.18	0.79
SEs	*3.54*	*2.29*	*1.58*	*1.15*	*0.93*	*0.85*	*0.83*	*0.82*	*0.82*	*0.82*	*0.82*	*0.82*	*0.85*	*0.92*	*1.13*
4	0.10	0.18	0.32	0.50	0.69	0.82	0.86	0.89	0.91	0.92	0.92	0.93	0.93	0.90	0.81
SEs	*3.16*	*2.36*	*1.77*	*1.41*	*1.20*	*1.10*	*1.08*	*1.06*	*1.05*	*1.04*	*1.04*	*1.04*	*1.04*	*1.05*	*1.04*

Table 6. S- χ^2 Item Level Fit Statistics for each Item (n = 6951)

Item	**χ^2**	***d.f.***	***p***
1	3547.50	90	< .001
2	4723.38	83	< .001
3	4368.79	96	< .001
4	2129.95	102	< .001

Table 7. Separate Group Item Parameter Estimates

Groups	**Item**	***a***	***b*$_1$**	***b*$_2$**	***b*$_3$**	***b*$_4$**	***b*$_5$**	***b*$_6$**
Faculty								
	1	2.05	-0.71	0.17	0.61	1.14	1.79	2.55
	2	4.01	0.20	0.85	1.22	1.62	2.10	2.34
	3	1.95	-0.35	0.46	0.94	1.40	1.81	2.17
	4	1.43	-0.47	0.72	1.44	1.95	2.27	2.78
Support Staff								
	1	2.38	-1.53	-0.73	-0.17	0.54	1.27	2.33
	2	3.41	-0.93	-0.27	0.23	0.86	1.52	2.19
	3	2.02	-1.46	-0.60	0.06	0.98	1.53	2.47
	4	1.73	-1.08	0.06	0.88	1.78	2.28	2.96

Note. *a* is the discrimination parameter. *b* is the difficulty parameter.

Table 8. S- χ^2 for each item separated by groups

Groups	**Item**	χ^2	***d.f.***	***p***
Faculty (n = 2068)				
	1	1085.98	83	< .001
	2	832.60	71	< .001
	3	709.05	83	< .001
	4	508.45	93	< .001
Support Staff (n = 2481)				
	1	622.97	76	< .001
	2	921.89	72	< .001
	3	831.25	84	< .001
	4	724.50	89	< .001

Table 9. Descriptive Statistics of Mean and Standard Deviation Separate by Groups

Item	Faculty (n = 2068) Mean	Faculty (n = 2068) SD	Support Staff (n = 2481) Mean	Support Staff (n = 2481) SD
1. The mission of the CSU has clearly been spelled out by the Chancellor's Office to CSU employees.	2.84	1.70	3.73	1.61
2. The Chancellor's Office inspires confidence in the future success of the CSU.	1.89	1.32	3.16	1.63
3. There is very little leadership from the Chancellor's Office. (R)	2.60	1.79	3.45	1.59
4. Employees are unsure about the direction the Chancellor's Office is going. (R)	2.40	1.62	2.72	1.44

Table 10. Path Coefficients across the Groups

	Faculty (n = 2068)		Support Staff (n = 2481)	
Item	**Standardized Coefficients**	**Unstandardized Coefficients**	**Standardized Coefficients**	**Unstandardized Coefficients**
1	.71	28.74*	.76	39.58*
2	.83	27.14*	.88	49.66*
3	.51	19.40*	.60	25.22*
4	.35	12.58*	.52	20.79*

Note. *$p < .05$

Table 11. Residuals across the Groups

	Faculty (n = 2068)		Support Staff (n = 2481)	
Item	**Standardized Residuals**	**Unstandardized Residuals**	**Standardized Residuals**	**Unstandardized Residuals**
1	.71	15.40*	.65	18.43*
2	.56	7.85*	.47	9.19*
3	.86	17.82*	.80	19.82*
4	.94	19.52*	.86	19.43*

Note. *$p < .05$

APPENDIX C

FIGURES

Figure 1.

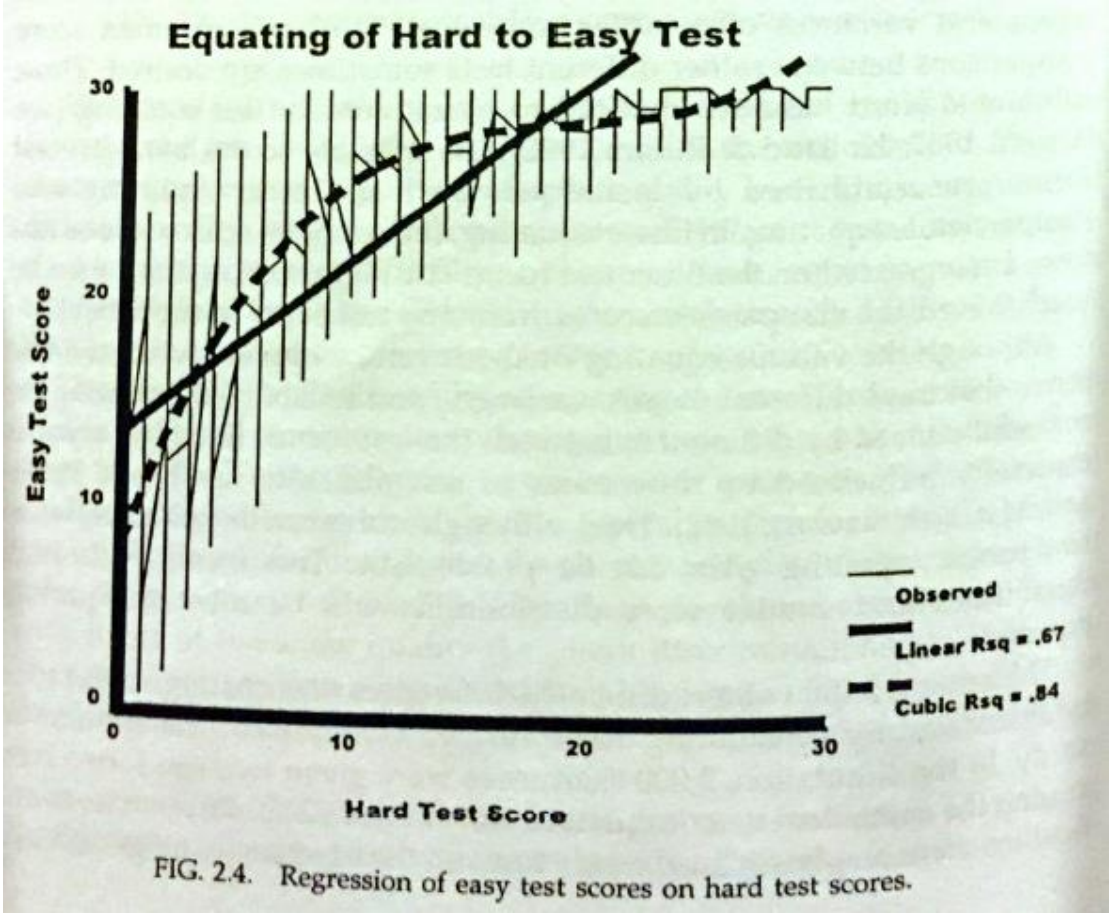
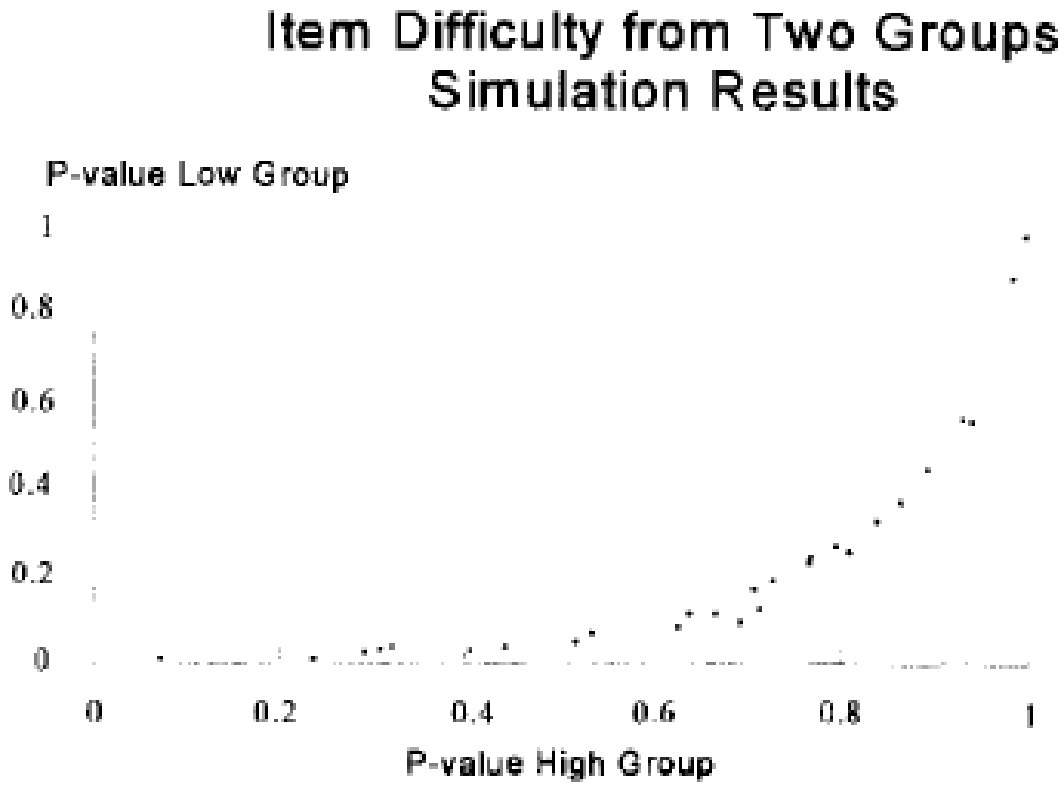


FIG. 2.4. Regression of easy test scores on hard test scores.

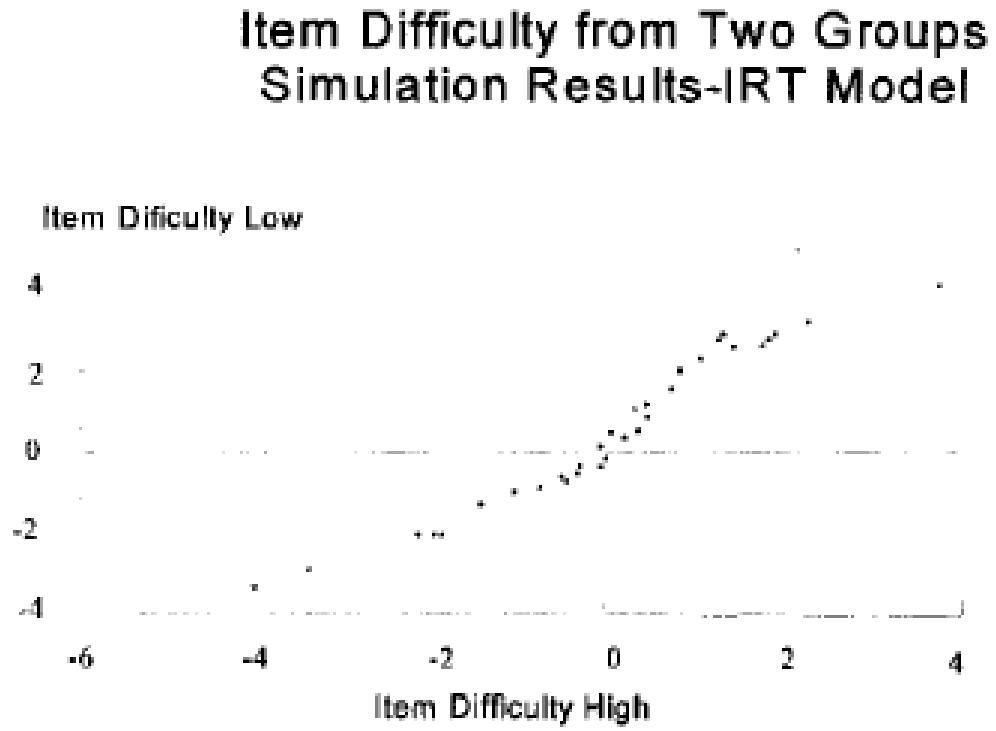
Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Psychology Press. (pp. 22).

Figure 2.



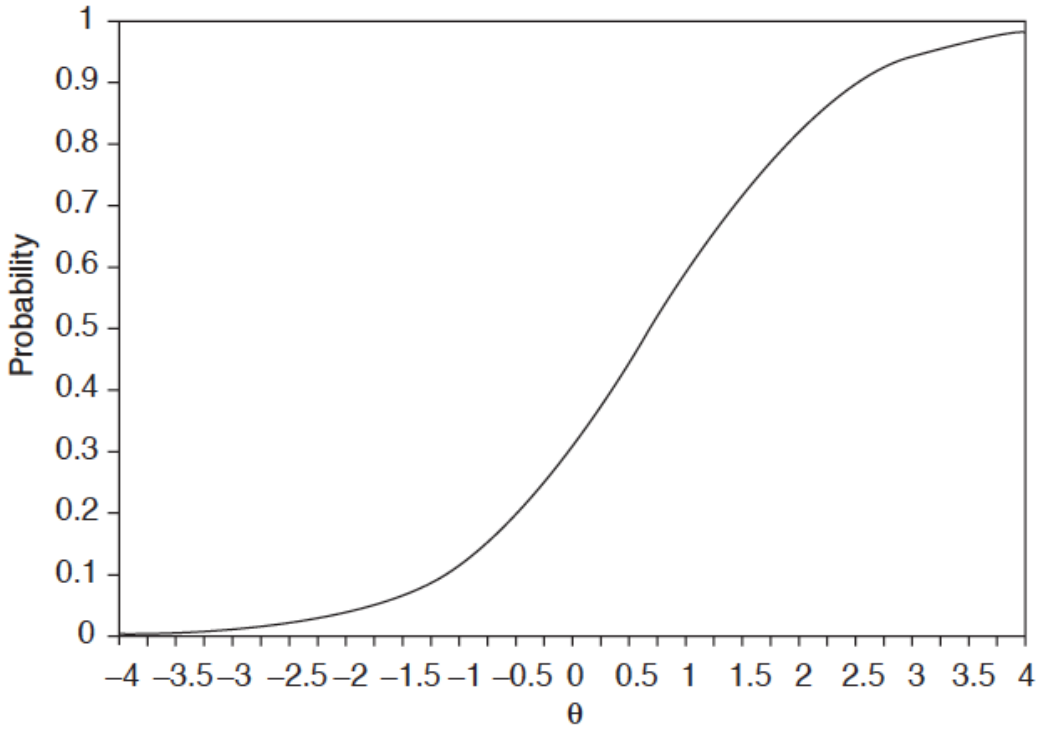
Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341. (pp. 345).

Figure 3.



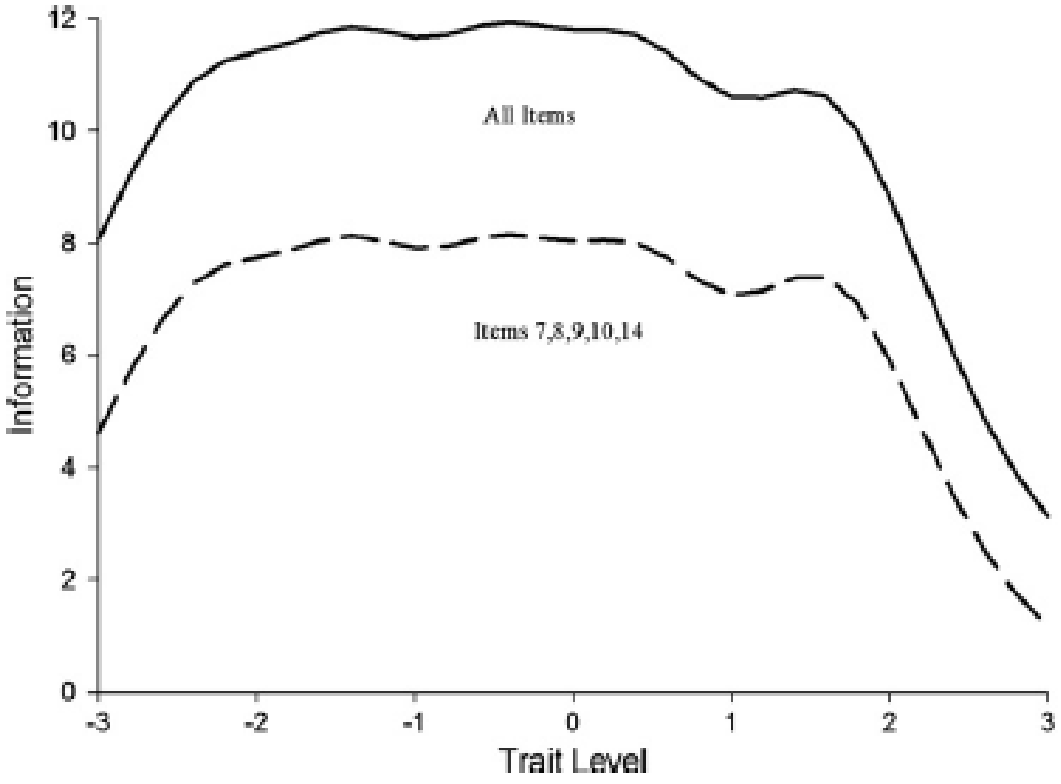
Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341. (pp. 345).

Figure 4.



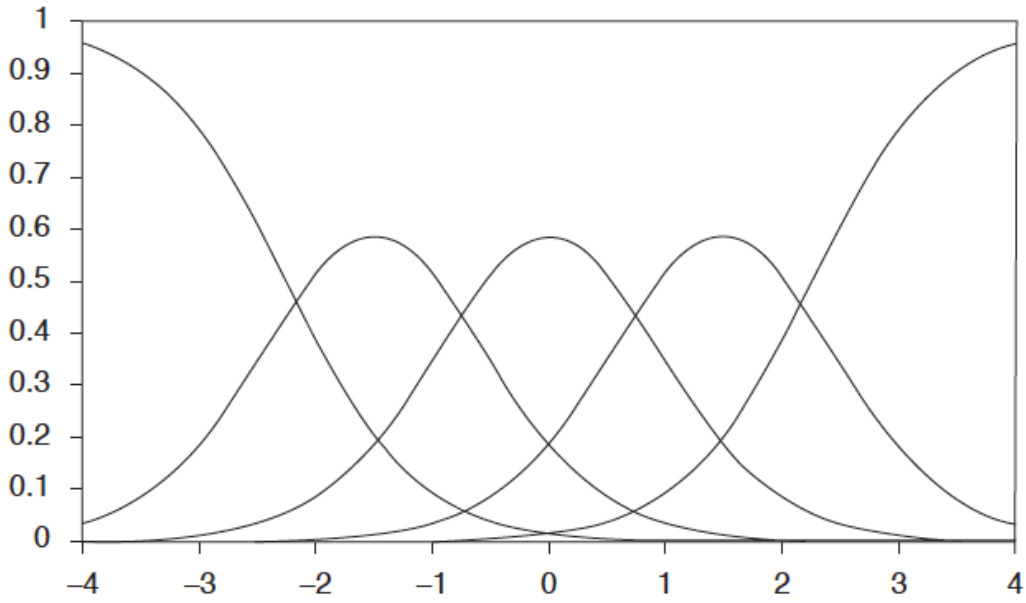
Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage Publications, Inc. (pp. 4).

Figure 5.



Van Dam, N. T., Earleywine, M., & Borders, A. (2010). Measuring mindfulness? An item response theory analysis of the Mindful Attention Awareness Scale. *Personality and Individual Differences*, *49*(7), 805-810. (pp. 809).

Figure 6.



Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage Publications, Inc. (pp. 6).

Figure 7.

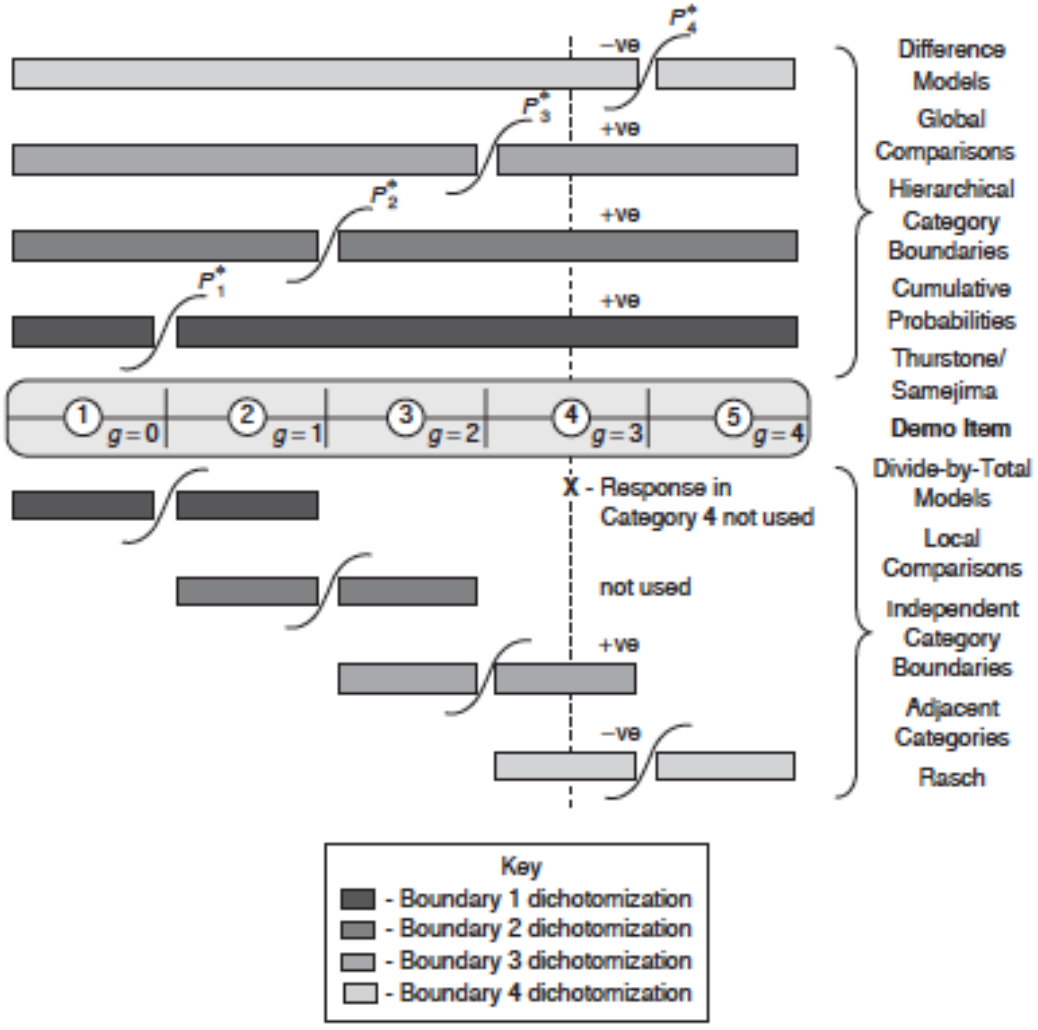
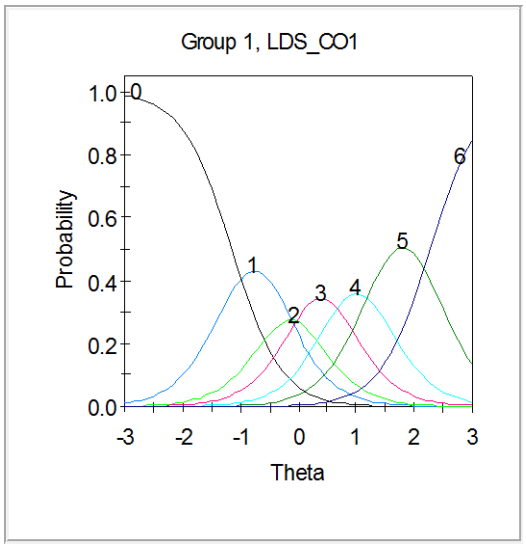
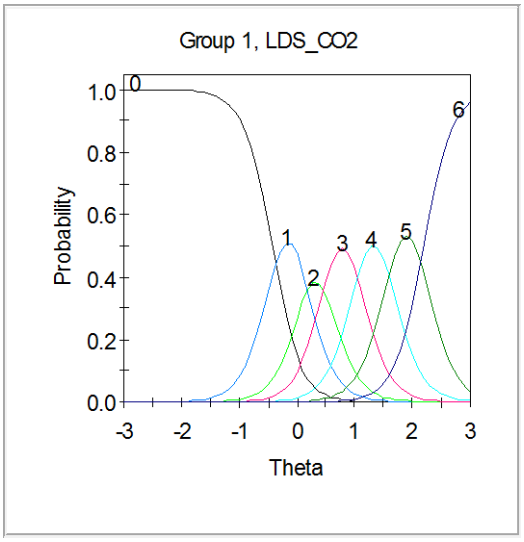


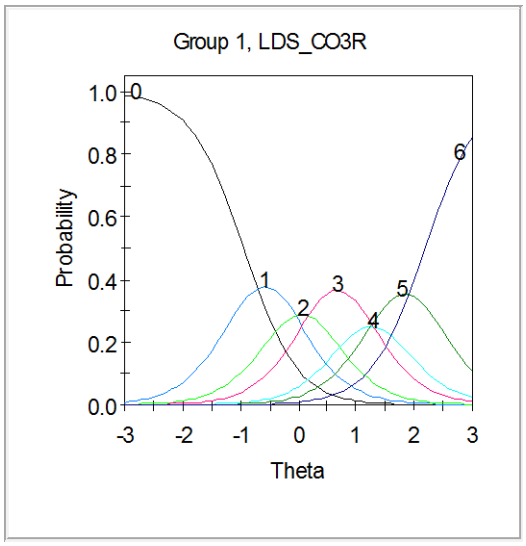
Figure 1.4 Graphical Representation of Two Approaches to Dichotomizing Polytomous Item Responses

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage Publications, Inc. (pp. 12).

Figure 8. ICRFs of each item







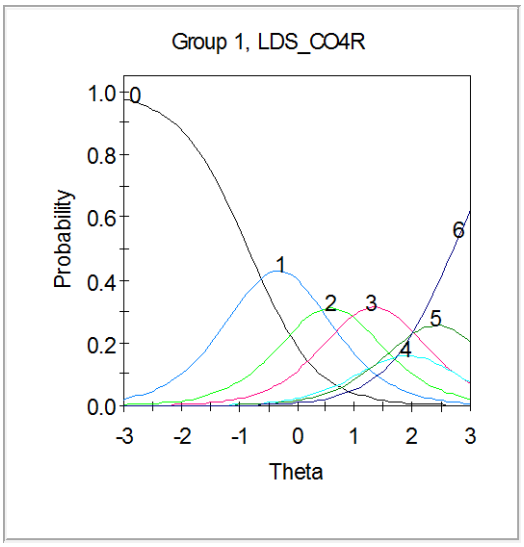


Figure 9. IIC graphs of each item

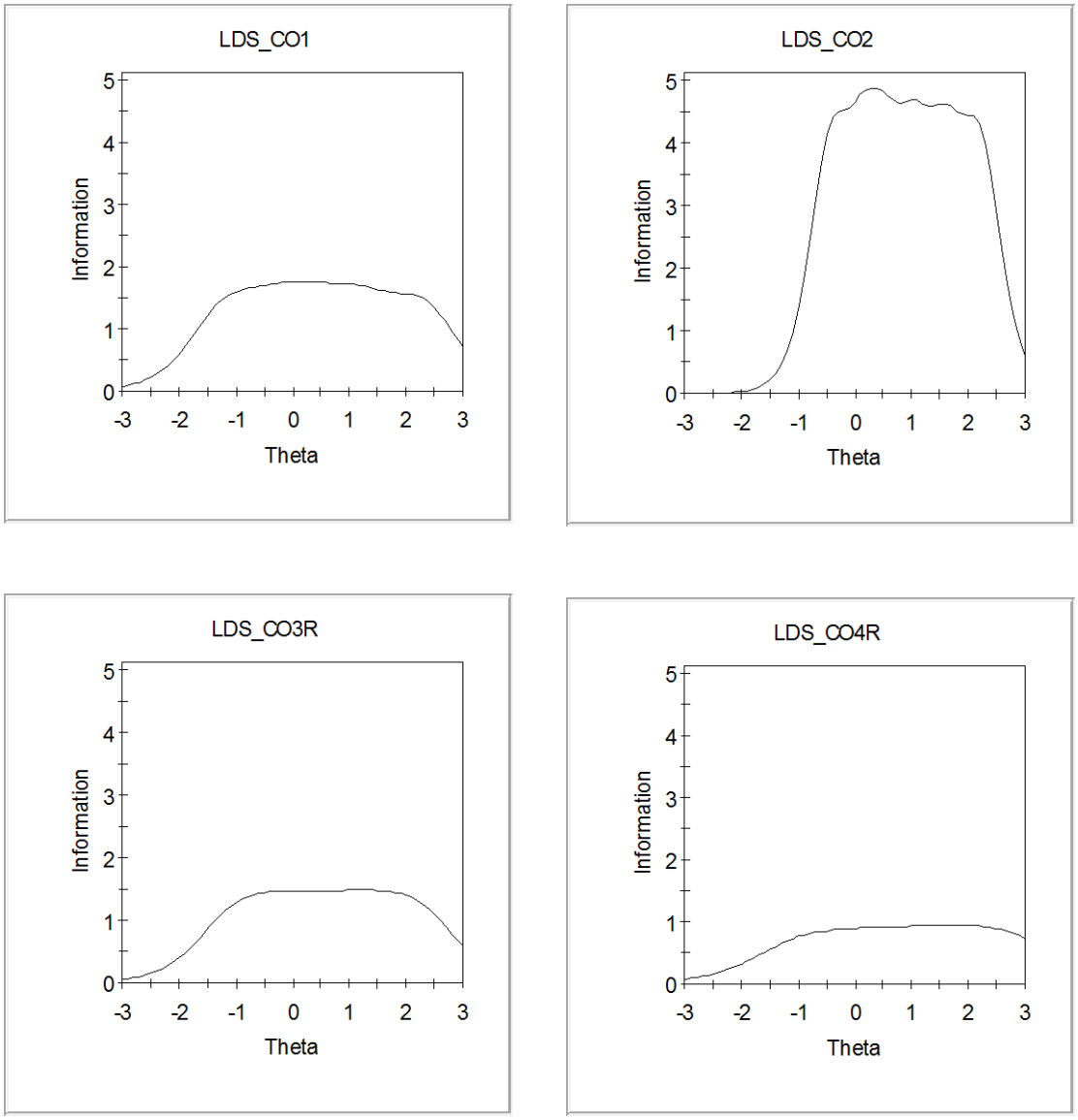


Figure 10. TIC for the TLDS

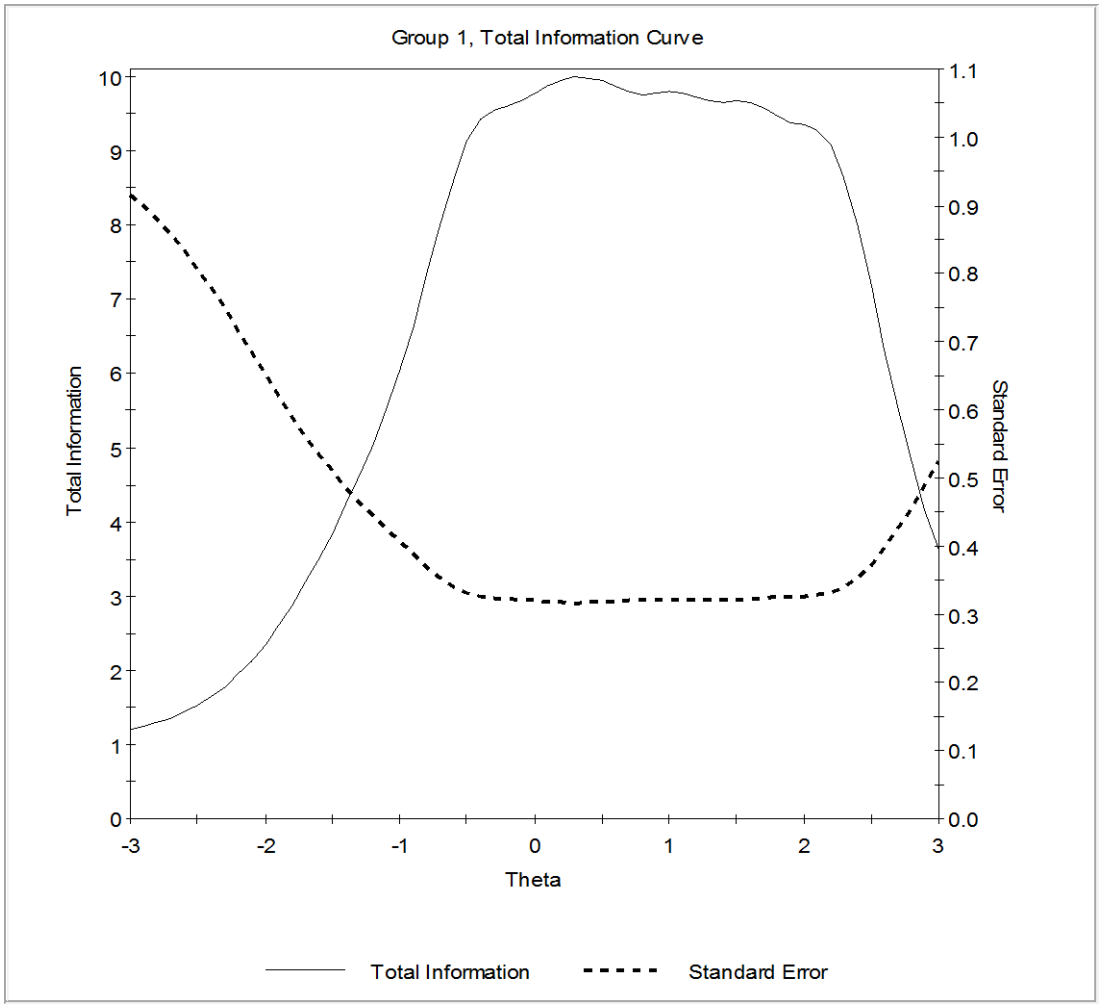


Figure 11. ICRFs of Each Item Separated by Groups

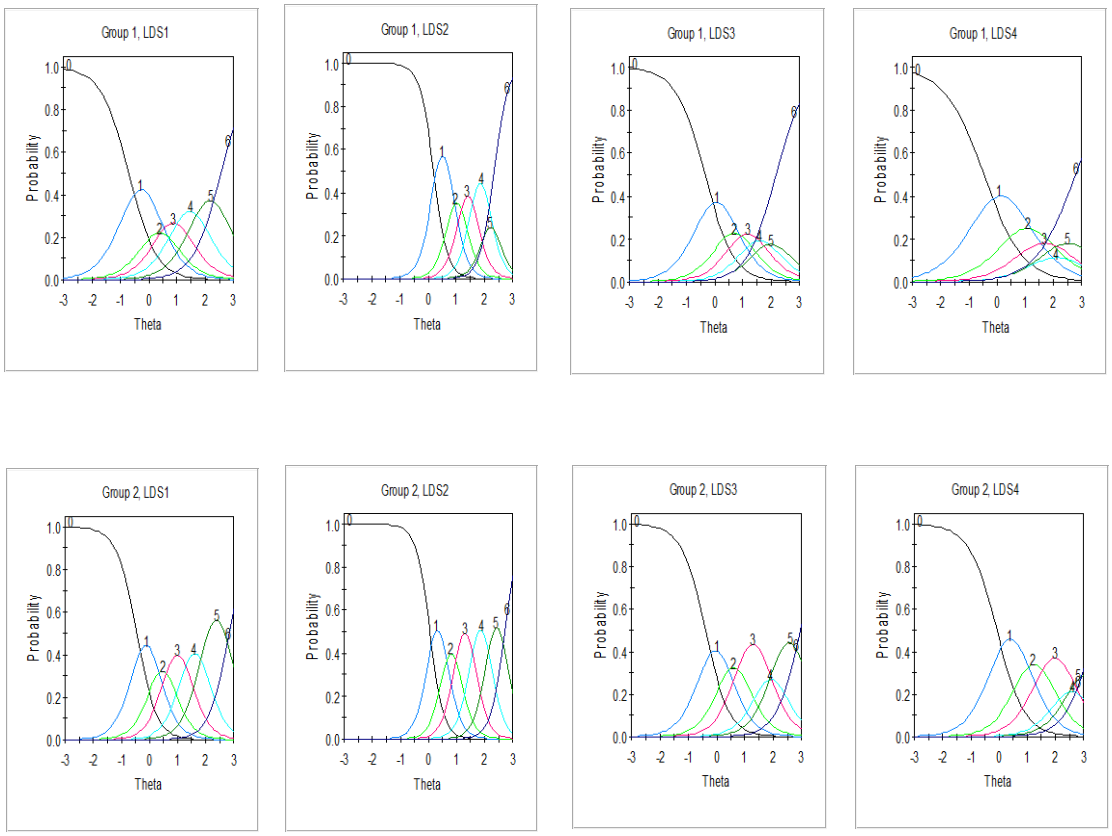


Figure 12. Example IRFs for a dominance model and an ideal point model.

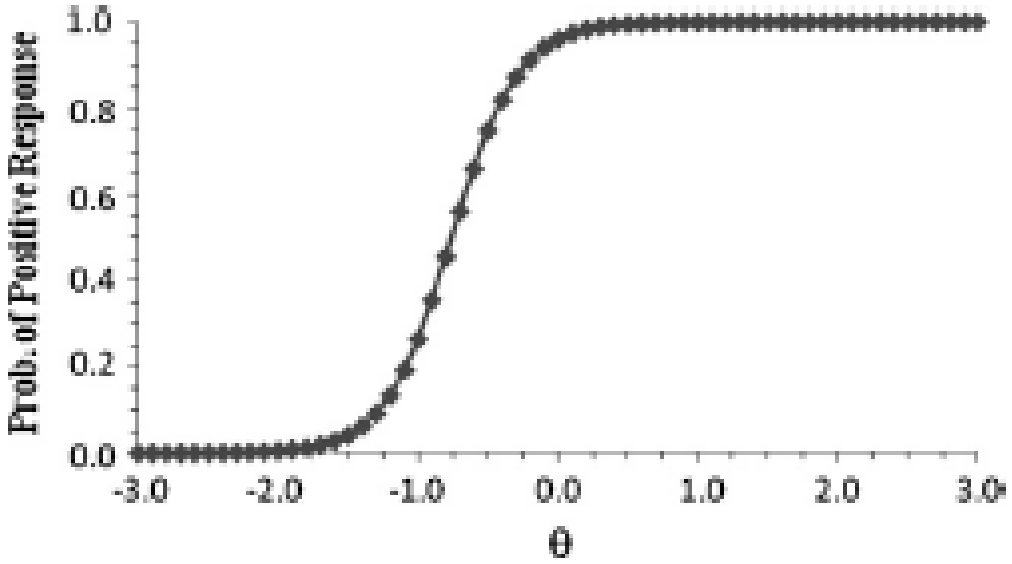


Fig. 1. Example of a dominance response process model.

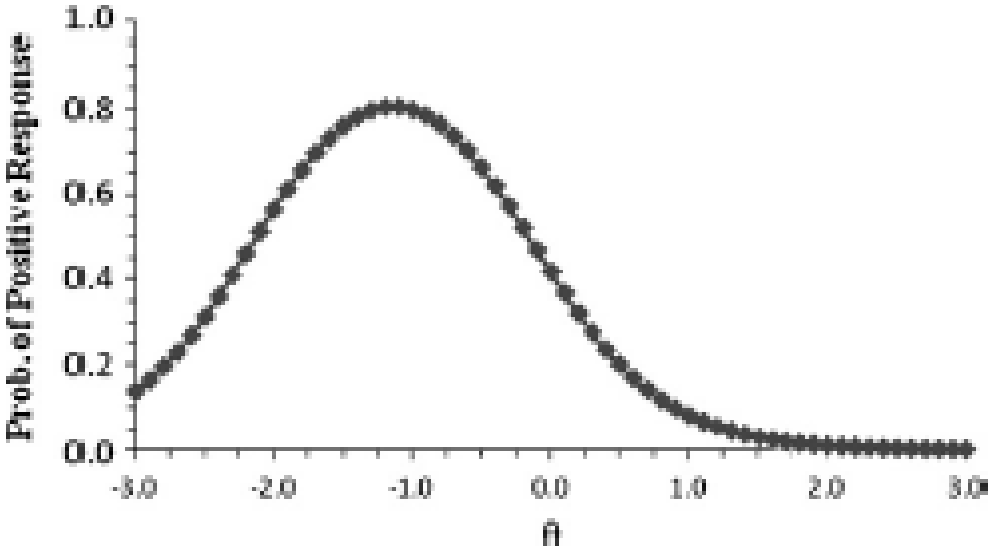
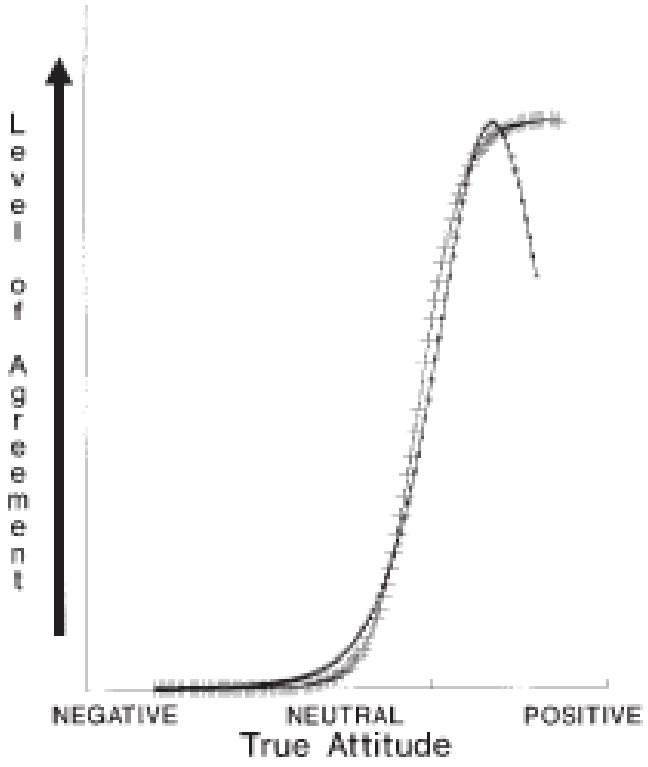


Fig. 2. Example of an ideal point response process model.

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI Work satisfaction scale. *Personality and Individual Differences*, *49*(7), 743-748.

Figure 13. Superimposed theoretical ICRFs from a dominance model and a ideal point model



Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, *59*(2), 211-233.

APPENDIX D

INSTITUTIONAL REVIEW BOARD

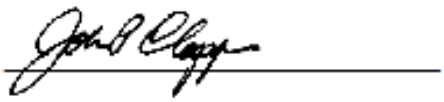
Human Subjects Review Board
Department of Psychology
California State University,
San Bernardino

PI:	Lee, Jung-Jung; Kottke, Janet
From:	John P. Clapper
Project Title:	Item Response Theory Analysis of the Top Leadership Direction Scale
Project ID:	H-16WI-26
Date:	2/27/16

Disposition: Administrative Review

Your IRB proposal is approved. The research involves the study of existing data, and no new data will be collected under this approval. This approval is valid until 2/27/2017

Good luck with your research!



John P. Clapper, Co-Chair
Psychology IRB Sub-Committee

REFERENCES

Baker, J. G., Rounds, J. B., & Zevon, M. A. (2000). A comparison of graded response and rasch partial credit models with subjective well-being. *Journal of Educational and Behavioral Statistics*, *25*(3), 253-270. doi:10.3102/10769986025003253

Bennis, W. (1999). The end of leadership: Exemplary leadership is impossible without full inclusion, initiatives, and cooperation of followers. *Organizational Dynamics*, *28*(1), 71-79. doi:10.1016/S0090-2616(00)80008-X

Bentler, P. M., & Wu, E. J. (2005). *EQS 6.1 for Windows*. *Structural equations program manual.* Encino, CA: Multivariate Software.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261–280. doi:10.1177/014662168801200305

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, (12), 253–260.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, *10*(1), 1-19.

Chaleff, I. (1995). *The courageous follower*, Berett-Koehler. CA: San Francisco.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523-562.

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology*, *85*(3), 451–461. doi:10.1037//0021-9010.85.3.451

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Cronbach, L. J., & Furby, L. (1970). How we should measure “change”: Or should we? *Psychological Bulletin*, *74*(1), 68-80. doi:10.1037/h0029382

Denis, J.-L., Lamothe, L., & Langley, A. (2001). The dynamics of collective leadership and strategic change in pluralistic organizations. *Academy of Management Journal*, *44*(4), 809–837.

Denis, J.-L., Langley, A., & Rouleau, L. (2010). The practice of leadership in the messy world of organizations. *Leadership*, *6*(1), 67–88. doi:10.1177/1742715009354233

Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin*, *92*(2), 526-531. doi:10.1037/0033-2909.92.2.526

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*(1), 19-29. doi:10.1037/0021-9010.72.1.19

Edelen, M. O., Mccaffrey, D. F., Marshall, G. N., & Jaycox, L. H. (2009). Measurement of teen dating violence attitudes. *Journal of Interpersonal Violence*, *24*(8), 1243–1263.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Psychology Press.

Flannery, W. P., Reise, S. P., & Widaman, K. F. (1995). An item response theory analysis of the general and academic scales of the self-description questionnaire II. *Research in Personality*, *29*(2), 168-188.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, *78*(2), 350–365. doi:10.1037/0022-3514.78.2.350

Holland, W. P., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit structural equation modelling: Guidelines for determining model fit. *Articles*, *6*(1), 53–60.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2014). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*, 486-507. DOI: 0049124114543236.

Kottke, J. L., Pelletier, K. L., & Agars, M. D. (2013). Measuring follower confidence in top leadership direction. *Leadership & Organization Development Journal*, *34*(4), 292–307. doi:10.1108/LODJ-07-2011-0062

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 55.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of likert-type personality data. *Multivariate Behavioral Research*, *40*(2), 261–279.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement*, *18*(3), 245–256. doi:10.1177/014662169401800305

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2 n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009-1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713-732.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *The Journal of Applied Psychology*, *95*(4), 728–743. doi:10.1037/a0018966

Meindl, J. R. (1990). On leadership-an alternative to the conventional wisdom. *Research in Organizational Behavior*, *12*, 159-203.

Michalisin, M. D., Karau, S. J., & Tangpong, C. (2004). Top management team cohesion and superior industry returns: An empirical study of the resource-based view. *Group & Organization Management*, *29*(1), 125-140. doi:10.1177/1059601103251687

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176. doi:10.1177/014662169201600206

Muthén, L. K., & Muthén, B. O. (2012). *Mplus* (Version 7) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100. doi:10.1016/S0022-2496(02)00028-7

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*(4), 289-298.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage Publications, Inc.

Pearce, C. L. (2004). The future of leadership: Combining vertical and shared leadership to transform knowledge work. *Academy of Management Executive*, *18*(1), 47–57. doi:10.5465/AME.2004.12690298

Pelletier, K. L., Kottke, J. L., & Reza, E. M. (2015). During furloughs, who is more attached to a public university? staff? faculty? . . . managers. *Public Personnel Management*, *44*, 120–142. doi:10.1177/0091026014558155

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368. doi:10.1177/014662169501900405

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the likert and thurstone approaches to attitude measurement Review of the Thurstone and Likert Approaches. *Educational and Psychological Measurement*, *59*(2), 211–233.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3-32.

Ryan, A. M., Ployhart, R. E., Schmitt, N., & Slade, L. A. (2000). Hypothesizing differential item functioning in global employee opinion surveys. *Personnel Psychology*, *53*(3), 531–562.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*, 100-114.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph*, *18*, 1-68.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics*. Pearson. NY: Allyn and Bacon

Tay, L., Meade, a. W., & Cao, M. (2014). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3–46. doi:10.1177/1094428114553062

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567-577.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118–128. doi:10.1037/0033-2909.99.1.118

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 26,* 249-269.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for latent GOLD 4.0: Basic and advanced*. Belmont, MA: Statistical Innovations Inc.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*(4), 479-494.

Wielkiewicz, R. M., & Stelzner, S. P. (2005). An ecological perspective on leadership theory, research, and practice. *Review of General Psychology*, *9*(4), 326–341. doi:10.1037/1089-2680.9.4.326