Insufficient Effort Responding on Multisource Feedback Surveys: A Mixed-Model Item Response Theory Investigation

by

ALLISON B. SIMINOVSKY

(Under the Direction of Nathan T. Carter)

ABSTRACT

In order to provide employees with performance feedback from multiple stakeholders with different perspectives, 360° feedback has become a popular tool in organizations. However, there have been questions as to whether these ratings actually help to improve employee performance. A recent stream of research on insufficient effort responding (IER), or a lack of rater attention when responding to survey items, could potentially help to explain problems with 360° feedback. This study uses mixed model item response theory (MM-IRT) to group raters on three 360° feedback surveys into latent classes based upon their response behavior, and then associates class membership with presence of IER measured through different indices. Findings indicate a strong presence of systematic responding on behalf of raters on all three surveys. In addition, there were some instances of systematic responding (e.g., lenient responding or central tendency responding) relating back to specific forms of IER. These results demonstrate the need to further consider rater behavior on performance feedback instruments, as well as the effectiveness of these instruments in shaping employee performance.

INDEX WORDS: Insufficient effort responding, multisource feedback, item response theory, mixed models, partial credit model, latent class, 360-degree feedback, surveys, performance appraisal

Insufficient Effort Responding on Multisource Feedback Surveys: A Mixed-Model IRT Investigation

by

ALLISON B. SIMINOVSKY

Bachelor of Business Administration, CUNY Baruch College, 2010

Master of Science, The University of Georgia, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2017

Insufficient Effort Responding on Multisource Feedback Surveys: A Mixed-Model IRT Investigation

by

ALLISON B. SIMINOVSKY

Major Professor: Nathan T. Carter
Committee: Malissa A. Clark
Gary J. Lautenschlager

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School The
University of Georgia
May 2017

DEDICATION

I wish to dedicate this dissertation to my parents, Robin and Paul Siminovsky. You have always pushed me to keep moving while being unbelievably supportive and loving – I could not have reached this point without your words of reassurance, hugs, and hours and hours on the phone. Thank you, thank you, thank you – I will never be able to say it enough. This is for you.

ACKNOWLEDGEMENTS

I wish to thank Dr. Nathan Carter for picking me up when I was academically orphaned and always pushing me to do better. You made this what it is, and I am very appreciative. I also wish to thank Drs. Malissa Clark and Gary Lautenschlager for their patience and wisdom in helping me accomplish this dissertation.

I want to thank my work colleagues for their constant inspiration, friendly ears, and willingness to believe in me and what I could accomplish. I want to thank my friends and family for their support throughout this process.

Finally, I want to thank Anna, Bob, and Cavan – my dissertation buddies. I am so proud of our collective for getting through this process together, and I could not have done it without you!

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER

1 Introduction 1

2 Literature Review and Research Questions 8

Mixed Model Item Response Theory 8

Insufficient Effort Responding 13

Detection of IER 17

3 Method 21

Sample 21

Measures 21

IER Indices 23

Analyses 24

4 Results 27

Scale Construction and Pre-Screening Statistics 27

MM-IRT 28

Associations with Class Membership 32

5 Discussion 35

Limitations and Directions for Future Research 41

Conclusion 44

REFERENCES 45

LIST OF TABLES

Page

Table 1: Reliability Estimates 55

Table 2: Intraclass Correlation Coefficient Estimates by Rater Type 56

Table 3: MM-IRT Model Fit Statistics 57

Table 4: Means and Standard Deviations of Survey Scales by Class 58

Table 5: Item Misfit Rate by Class 59

Table 6: MM-IRT Latent Class Size Estimate (π) by Class Type 60

Table 7: χ^2 and φ Coefficients of Class Consistency 61

Table 8: Reliability Estimates by Class 62

Table 9: Intraclass Correlation Coefficients by Rater Type and Class 63

Table 10: χ^2 and φ Coefficients of Rater Type by Class Membership 64

Table 11:One-Way Analysis of Variance of IER by Class 65

Table 12: IER Means and Standard Deviations by Class 67

Table 13: Presence of IER by Class 69

LIST OF FIGURES

Page

Figure 1: Example Category Characteristic Curve for a Three-Option Item.....70

Figure 2: Threshold Parameter Plot for the Goal Setting Scale (Non-Manager).....71

Figure 3: Threshold Parameter Plot for the Interpersonal Scale (Non-Manager).....71

Figure 4: Threshold Parameter Plot for the Interpersonal Scale (People Manager).....72

Figure 5: Threshold Parameter Plot for the Task Performance Scale (People Manager)..72

Figure 6: Threshold Parameter Plot for the Interpersonal Scale (Executive Development)….....73

Figure 7: Category Probability Histogram for Class 1 of the Goal Setting Scale (Non-Manager).....74

Figure 8: Category Probability Histogram for Class 2 of the Goal Setting Scale (Non-Manager).....74

Figure 9: Category Probability Histogram for Class 1 of the Interpersonal Scale (Non-Manager).....75

Figure 10: Category Probability Histogram for Class 2 of the Interpersonal Scale (Non-Manager).....75

Figure 11: Category Probability Histogram for Class 3 of the Interpersonal Scale (Non-Manager).....76

Figure 12: Mean Values of the LongString Index 77

Figure 13: Mean Values of the Outlier Index 78

CHAPTER 1

INTRODUCTION

The multisource feedback survey has become a ubiquitous tool for performance evaluation. Multisource feedback, and in particular, 360° feedback, differs from traditional performance appraisal in that performance ratings are gathered from multiple individuals rather than solely from the employee's direct manager. The individuals surveyed in multisource feedback are often the employee's direct manager and higher level managers, subordinates, peers, clients, and the employee him- or herself (Atwater, Brett, & Charles, 2007). The incorporation of ratings from various stakeholders is intended to provide a more complete overview of an employee's performance, as it is assumed that each rater group is privy to distinct performance incidents and can therefore provide unique insights (London & Smither, 1995). This should provide a more holistic picture of employee performance than can be gauged by the employee's manager alone, as managers often do not have the opportunity to witness an employee's performance in all work situations (Oh & Berry, 2009).

It has been demonstrated that 360° feedback provides advantages to an organization. 360° feedback surveys have repeatedly demonstrated improved employee performance post-survey (Bailey & Austin, 2006; Johnston & Ferstl, 1999; Kluger & DeNisi, 1996; Smither, London, & Reilly, 2005). The purpose of 360° feedback, and of providing employee feedback in general, is often to help the employee build self-awareness regarding his/her strengths and areas of opportunity (Atwater & Brett, 2006).

The inclusion of multiple points of view in performance ratings should give the employee more actionable feedback on which to base development plans, potentially resulting in performance improvement. Seeking out feedback helps leaders to gain self-awareness, which should provide insight as to how others view a leader. This allows for a leader to gain awareness into his/her strengths and weaknesses, engage in development, and subsequently improve performance (Ashford, 1989; Day, Fleenor, Atwater, Sturm, & McKee, 2014).

In addition to individual performance, multisource feedback has been found to impact overall workplace productivity. Kim, Atwater, Patel, and Smither (2016) demonstrated that the use of a multisource feedback system in an organization, particularly when used for both administrative and developmental purposes, impacted employee knowledge sharing. This knowledge sharing directly impacted workforce productivity. This was true in the short term and as far out as four years later. Therefore, multisource feedback positively affects not only the individual rating target, but the organization overall.

Despite the benefits of 360° feedback, there are a number of well-documented issues in demonstrating its role in employee development (Atwater et al., 2007; Bailey & Austin, 2006). The results of 360° feedback surveys are typically weak predictors of year-end performance ratings, if the relationships are significant at all (Murphy, 2008; Smither, London, & Richmond, 2005; Zimmerman, Mount, & Goff, 2008). Should self- and other ratings align, the employee might not find the need to work toward improved performance, as the consistency between self/other perceptions does not motivate the

employee to improve (Korman, 1970). When the rating target overrates him- or her-self compared to other raters, performance can even worsen (Johnson & Ferstl, 1999).

Given the sensitivity of subsequent performance to overall ratings and their deviations from one another in multisource feedback measures, it is of crucial importance to consider the accuracy of collected data when dealing with employee performance feedback. When used in the development context, performance appraisal accuracy can impact those areas in which an employee chooses to direct development efforts, and can subsequently determine whether development has an impact on performance. In line with Ashford's (1989) theorized relationship between feedback and performance, inaccurate performance appraisal data can actually hinder employee self-awareness and have minimal, if not negative effects on employee performance. Furthermore, performance appraisals can be used to make far-reaching administrative decisions, including an individual's continued employment with the organization or the decision to give an employee a promotion or pay raise. Such decisions can have major financial and legal implications if made fallaciously. Therefore, efforts should be taken to ensure that data collected as part of performance feedback programs are as accurate as possible.

There are also problems with data accuracy as they relate to rater behavior. Specifically, raters have been found to provide biased ratings depending on their relationships to the rating target. For example, direct reports provide more lenient ratings of their managers, even under anonymous rating conditions (Bamberger, Erev, Kimmel, & Oref-Chen, 2005). In another study, direct reports were found to engage in more leniency as well as halo than other rater types when providing ratings (Ng, Koh, Ang, Kennedy, & Chan, 2011). Several studies have cited success in the use of different types

of rater training to reduce rating error (e.g., rater error training or frame of reference training; Woehr, Sheehan, & Bennett, 2005). However, conducting such training involves the availability of subject matter experts as well as sufficient time, money, and other resources, which could be prohibitive for some organizations.

Researchers have focused on the collection of data itself, utilizing such approaches as frame of reference rating scales to demonstrate moderate improvements in rating accuracy over traditional multisource rating scales (Hoffman, Gorman, Blair, Meriac, Overstreet, & Atchley, 2012) and Thurstonian forced-choice ranking over the typically used Likert Scale (Brown, Lin, & Inceoglu, 2017). Other studies have considered evaluating 360° feedback survey data using advanced analytical models. For example, Barr & Raju (2003) applied several item response theory (IRT) models to examine measurement equivalence and found that different models provided unique information regarding rater leniency and severity. This existing research focuses on how to best collect and analyze data to demonstrate outcomes, but has thus far failed to address the issue of examining data quality. It is unknown exactly *what* or *who* is leading to difficulty in establishing outcomes. This lack of attention to examining the quality of data *post hoc* could play a role in some of the problems in demonstrating the effectiveness of 360° feedback.

In this dissertation, I draw attention to a recent stream of research on insufficient effort responding (IER) and its accompanying statistical procedures, which have the potential to detect some of the issues present in 360° feedback survey data, allowing practitioners to better estimate performance and provide accurate and meaningful feedback to rating targets.

IER refers to survey respondent behavior marked by little to no motivation to follow survey instructions, process item content, or provide ratings in an effortful manner (Huang, Curran, Keeney, Poposki, & Deshon, 2012). There are a number of ways in which IER can manifest itself, including endorsing multiple consecutive items with the same response regardless of item content. In general, IER is indicative of a misunderstanding of or disregard for survey instructions and/or item content. As methods for screening data for IER are statistically straightforward, both researchers and organizations can incorporate such methods into their practices without the need to master advanced statistical techniques or procure expensive analytic software.

The prevalence of IER in 360° feedback survey data is currently unknown, but findings from studies of IER in traditional self-report surveys have detected prominent rates of the phenomenon (Huang, Liu, & Bowling, 2015; Meade & Craig, 2012). These findings are not altogether surprising: a recent Gallup survey noted that only 32.7% of U.S. workers report being engaged at work (Adkins, 2016). If the majority of workers are not engaged, it is unlikely that they will take part in voluntary organizational activities, such as providing feedback to others, in an enthusiastic or effortful manner.

Even when the prevalence of IER is minimal, the removal of IER-flagged data has proven to be psychometrically beneficial (Huang et al., 2012). It follows that by screening 360° feedback data for IER, one could identify those raters who answered questions with insufficient effort and eliminate these cases in order to improve upon data quality. Given the rates of IER detected in previous research and organizational surveys, it is likely that at least some survey respondents engage in IER, which could be distorting data in such a way that makes it challenging to draw definite conclusions about 360°

feedback effectiveness. It is unknown whether IER is prevalent in multisource feedback surveys and which rater types might be more prone to engage in IER. To date, no study has examined IER in the context of multi-rater surveys, nor has any study considered the role of IER in performance feedback surveys.

In the present study, I apply methods from IER research to 360° feedback survey data used for employee development from a mid-size pharmaceutical organization. I consider three different survey forms: one for non-managers, one for managers of people, and a survey used at the beginning of an executive development program involving leadership coaching and group training. I first use mixed-model item response theory (MM-IRT) to determine whether a class of respondents emerges that shows response patterns indicative of IER, as well as the size of this class. Next, I determine the likelihood that individuals from different rater groups (e.g., self vs. manager vs. direct report) will fall into each of these classes. It is my intention to answer several questions regarding the intersection of 360° feedback surveys and IER through this study. First, to what extent is 360° survey data affected by IER? Second, in what ways does IER manifest itself in the data? Third, does the presence of IER in 360° survey data affect the statistical properties of the data, such as interrater reliability? Finally, are certain rater groups more likely to engage in IER? Regardless of the findings, the outcomes of this study have the potential to improve research and organizational practices alike in their treatment of performance data. Additionally, this study will provide deeper insights into the behaviors of survey respondents, particularly in the context of 360° feedback surveys.

In the following chapters, I first describe MM-IRT, its past applications, and its relevance to the issue of IER. Next, I review the current literature on IER and popular

techniques for its detection; research questions are proposed on the basis of this review. I will then discuss the methods for this study, followed by an overview of the analytic results. Finally, I will discuss the implications of the findings, limitations, and directions for future research.

CHAPTER 2

LITERATURE REVIEW AND RESEARCH QUESTIONS

Mixed Model Item Response Theory

Item response theory (IRT) is a family of statistical models in which the probability of a response to an item (called the item response function, or IRF) is proposed to be dependent on a latent variable or trait (referred to as θ). The choice of an IRT model is a function of item format and the research question(s) one is seeking to answer.

Most IRT models account for the concepts of item *discrimination* and *extremity* (sometimes referred to as *difficulty*). Item discrimination can be conceptualized as the IRF's slope (as seen in each category characteristic curve in Figure 1) and describes the extent to which the item discriminates between individuals at different points on the latent continuum. Item extremity is the maximum point of the IRF's slope and represents the item's location on the latent continuum. In the case of dichotomous data, the item's extremity reflects a 0.50 probability of a respondent selecting one response option over the other (e.g., this is the point on the continuum at which the respondent has 50:50 odds of answering an item correctly or incorrectly). When dealing with polytomous data, as in the current proposal, the item's extremity is reflected as an average of the locations of *K*-1 *thresholds*, where K equals the number of response options to the item. A threshold between two response options represents the point on the latent continuum at which the probability of endorsing a certain response becomes greater than endorsing a different

response. As seen in Figure 1, the thresholds are the points at which the curves intersect (i.e., at θ = -2, 0, and 2).

In traditional IRT models, it is assumed that respondents are drawn from the same homogeneous subpopulation – that is, respondents that are at the same level on the latent continuum will respond to items in a similar manner. It is possible to examine manifest differences in a population in which subgroups are identified *a priori*, such as gender, ethnicity, or other readily-observable traits. These differences are often considered through the analysis of *differential item functioning*, or DIF, which occurs when individuals at the same level on the latent trait have different probabilities of endorsing items in a certain way (Mellenbergh, 1982). When DIF is present, the assumption of invariance is violated as it is clear that the population is not homogenous, and the groups cannot be represented using a single set of item parameters. Raju, van der Linden, and Fleer (1995) proposed the DFIT (differential functioning of items and tests) framework in IRT to examine not only item-level differentiation, but variation at the scale or test level.

DIF analysis has been applied to 360° feedback to examine measurement invariance across groups in numerous studies. Maurer et al. (1998) considered measurement invariance across peer and subordinate ratings for two samples on a multisource feedback assessment. Using the DFIT framework, they did not find significant differences in either of their samples, suggesting that it is reasonable to compare the ratings of peer and subordinate raters.

Facteau and Craig (2001) used the DFIT approach in their examination of invariance across various rater types. Across 276 differential functioning indexes, they found only five occurrences of noncompensatory DIF (NCDIF) and one occurrence of

differential test functioning (DTF), constrained to a single scale across three items for the subordinate rater group. The differential functioning was not only localized to a small portion of the assessment, but was also very minimal in its effect, suggesting that ratings across raters groups are generally invariant and can be fairly compared against one another.

Barr and Raju (2003) considered whether self-raters and other raters were rating 360° feedback items on the same metric in order to determine the fairness of self-other rating comparisons. They looked at three different models: the generalized partial credit model (GPCM; Muraki, 1993), the rater's effect (RE) model, and the hierarchical rater model (HRM). The GPCM shows the interaction between rater and item through the estimation of different item parameters by source, whereas the RE and HRM show rater tendency towards leniency or severity. In addition, the HRM provides a rater reliability index. The results showed evidence of DIF on some items, but this was almost always compensatory when considering the entire scale. In addition, it was found that self-raters were always more lenient in providing ratings than their managers. Under the HRM, self-raters were the most reliable rating source, followed by their managers, peers, and direct reports, who were the least reliable rating source. Overall, these results demonstrated that while different rater groups are generally providing ratings on the same metric, there are differences in leniency across rater groups.

Craig and Kaiser (2003) used DIF analysis to determine the effect of the violation of independence inherent in multisource ratings. They compared random samples of a popular 360° feedback instrument that varied based on scale length and content domain. Using the DFIT framework, it was found that no item or scale met criteria for differential

functioning, regardless of scale length or sample size. This finding demonstrated that despite the fact that using multiple raters per rating target violates the assumption of independence, this should not affect IRT parameter estimates. Taken together, these studies demonstrate the appropriateness of IRT as a method to examine multisource feedback, providing rich information about the functioning of individual items and scales as a whole.

Although DIF is useful when considering manifest differences, researchers often want to observe latent differences, or those fault lines in the population that cannot be as easily observed as demographic characteristics, if they can be observed at all. Therefore, one must rely on other means to identify latent differences in the population *a posteriori*. Mixed model IRT (MM-IRT) provides researchers with such an opportunity.

MM-IRT is a hybrid approach combining principles of latent class analysis (LCA) with IRT in that individuals are divided into classes on the basis of different item parameters. When using MM-IRT, it is possible to see various measurement models hold for different subgroups within the same population, or the same measurement model hold across a single population with different parameter estimates for unique subgroups (Maij-de Meij, Kelderman, & van der Flier, 2008). In this way, MM-IRT "unmixes" the data and allows for the estimation of parameters unique to various subgroups – it does not force parameters that might only be appropriate for some respondents onto an entire population (Rijmen, Tuerlinckz, De Boeck, & Kuppens, 2003; Spiel & Glück, 1998). When it is suspected that groups of participants are responding to a measure differently, one can employ MM-IRT to consider (a) whether any subgroups exist and (b) the underlying differences between subgroups. In other words, it is possible to examine

previously unknown groupings within a population, making MM-IRT unique from standard IRT model estimation (Maij-de Meij et al., 2008).

There are numerous of examples of researchers employing MM-IRT to consider subgroup differences, particularly regarding response style. Eid and Rauber (2000) detected qualitative differences based on gender, organizational tenure, and job level in determining whether individuals were likely to use the entire response scale versus extreme responses on a leadership satisfaction survey. Hernández, González-Romá, and Drasgow (2004) used the mixed-partial credit model to examine whether there were subgroups with a higher likelihood of using the middle-response option on Likert scale items (e.g, ?, undecided) on the 16PF personality inventory. For all subscales, they found that either two- or three-class solutions fit the data better than one-class solutions and there were indeed respondents who favor the middle choice on response scales versus those who use such responses infrequently. Zickar, Gibby, and Robie (2004) found a three-class solution in their analysis of applicant faking – those who did not fake, those who were instructed to fake, and those who faked responses on their own volition. Austin, Deary, and Egan (2006) used MM-IRT with NEO-FFI data to examine individual differences in responding to personality scale items (extreme versus middle-of-the-scale responses). Maij-de Meij et al. (2008) found differential use of a "?" category on the response scale of a personality inventory and demonstrated that social desirability and ethnic background played roles in latent class membership. Egberink, Meijer, and Veldkamp (2010) demonstrated that the personality trait of conscientiousness might function differently based on an individual's levels of perfectionism and neuroticism. Carter, Dalal, Lake, Lin, and Zickar (2011) found that respondents to the Job Descriptive

Index fell into one of three classes – those who tend to positively endorse items, those who tend to negatively endorse items, and those who tend to select the middle option (?) of the scale. These studies all demonstrate the value of MM-IRT measurement models in shedding light on otherwise unobservable class differences.

In the current study, I will attempt to uncover classes of raters engaging in insufficient effort responding – as this classification of raters is based on a latent trait, MM-IRT is an appropriate method. In the next section, I will explain the concept of insufficient effort responding, its history in psychological and organizational literature, and methods for detection.

Insufficient Effort Responding

Insufficient effort responding (IER; Huang et al., 2012) refers to those survey responses made in a careless, random, and/or inconsistent manner. Huang et al. (2012) provide a comprehensive definition of the phenomenon:

> …a response set in which the respondent answers a survey measure with low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses (p. 100).

The broadness of this term is intentional as it includes those responses which were haphazardly selected as well as those which were selected systematically, but without sufficient effort on the part of the respondent (i.e., selecting the middle choice on each of a series of Likert scale items regardless of item content). However, purposeful false responding, such as faking for impression management, does not fall under the IER spectrum, as this is a highly effortful process (Snell, Sydell, & Lueke, 1999). In fact, Meade & Craig (2012) found that a social desirability scale provided no insight into

which respondents would engage in IER, suggesting that faking for social desirability and IER are unique phenomena. The study of IER does not speculate about the root of a respondent's motivation (or lack of it) for his/her response pattern: IER includes random responding in addition to systematic response error to allow for the examination of those responding without attentively following the survey's directions, regardless of motivational mechanisms at play (Huang et al., 2012; Liu, Bowling, Huang, & Kent, 2013).

The notion of screening data for survey response accuracy is not new to psychology. On the contrary, there is a wealth of literature examining validity scales (e.g., Archer, Handel, Lynch, & Elkins, 2002) and participant faking (e.g., Bagby, Nicholson, Bacchiochi, Ryder, & Bury, 2002; Donovan, Dwight, & Hurtz, 2003), for both impression management (Donovan et al., 2003) and less positively-motivated reasons, such as malingering (Bacchiochi & Bagby, 2006).

Many prominent clinical diagnostic and personality measures make use of validity scales in order to detect response patterns indicative of lack of respondent attention or intentional faking. Such scales typically use covertly-worded items to inconspicuously detect responses indicative of social desirability (Detrick, Chibnall, & Call, 2010) or carelessness (Archer et al., 2002). The consequences of inaccurate responding on a clinical measure can be devastating: not only might misdiagnoses be harmful to the targeted individuals and those around them, but judicial decisions often rely on the results of diagnostic personality surveys (Bacchiochi & Bagby, 2006). If these surveys fail to detect intentionally or inadvertently inaccurate responses, court decisions might be made under fallacious or misguided premises.

In organizational research, the historical focus on inaccurate survey responding has centered on faking and general self-misrepresentation, often by job applicants (Allen, Facteau, & Facteau, 2004; Donovan et al., 2003; Snell et al., 1999). For example, many integrity tests use covertly worded items such that respondents cannot gauge the "correct," or more desirable, answer (Alliger, Lilienfeld, & Mitchell, 1996). Situational judgment tests provide respondents with response options such that it is difficult to distinguish "good" versus "bad" behaviors, foiling attempts on behalf of the respondent to fake his/her responses (Arthur, Glaze, Jarrett, White, Schurig, & Taylor, 2014). The focus on response inaccuracy in organizational research has typically revolved around faking, and has only recently begun to venture into a thorough examination of lack of respondent effort, regardless of motivational forces (Huang et al., 2012). While it is of paramount importance to use measures resistant to faking in order to make well-reasoned decisions and ensure data accuracy, it is equally important to consider the possibility that respondents take part in organizational survey measures with insufficient effort and can be similarly tainting the data collected from such efforts.

The potential consequences of a response set plagued by IER underscore the importance of thoroughly screening one's data when using surveys. In an examination of the criterion-related validity of various constructs for predicting job-related criteria, Hough, Eaton, Dunnette, Kamp, and McCloy (1990) found that careless survey responding moderated the predictive validity of all examined relationships, such that predictive validity was lower for careless responders. The presence of these attenuated relationships poses a threat to research in that it increases the likelihood of committing Type-II error (retaining a false null hypothesis; Liu et al., 2013). Huang, Bowling, Liu,

and Li (2015) demonstrated that when IER introduces systematic error to a dataset, relationships between constructs may be *inflated* instead of attenuated, increasing the likelihood of committing Type-I error (falsely rejecting a null hypothesis). Furthermore, Huang et al. (2012) used confirmatory factor analysis of their full data set versus the IER-trimmed dataset to determine that the trimmed data yielded better model fit. Taken together, these results show not only how damaging the effects of IER can be on scientific inquiry, but also that the nature of these effects are not necessarily predictable, highlighting the need to screen data for IER in both research and practical contexts.

A recent study also demonstrated the importance of screening data for IER on a regular basis. Bowling, Bragg, Liu, Huang, Khazon, and Blackmore (2016) conducted surveys at multiple time points across numerous samples and found that those respondents who engaged in IER at time 1 were likely to do so again at time 2. In addition, the study demonstrated that the personality traits of conscientiousness and agreeableness as rated by acquaintances were negatively related to IER; that is, the more conscientious or agreeable the respondent, the less likely he/she was to engage in IER. These findings demonstrate the temporal consistency of IER, or that those who in engage in IER are likely to do so repeatedly. Due to this consistency, it is important to regularly screen for IER, even when dealing with the same sample as a previous survey.

To date, no study has examined the effect of IER on survey-based performance feedback, where insufficient effort in providing ratings could have major consequences for an individual's employment status and developmental focus areas. The notion of using performance feedback plagued by IER to make organizational decisions is threatening to the continued use of 360° feedback, and therefore calls for immediate

investigation. As previous IER research has not examined performance ratings, it is unknown the extent to which IER pervades 360° feedback survey data, as well as how IER affects data quality.

Research Question 1: What is the prevalence of IER in 360° feedback survey data?

Research Question 2: What differences in data quality in terms of classical statistics are evident between sufficient and insufficient effort rater groups?

Detection of IER

Given the various possible response patterns that constitute IER, there are numerous methods available for its identification and detection in survey data. These varied approaches target different aspects of IER and have been found to be correlated to a low to moderate extent (Meade & Craig, 2012). Although this provides some evidence for convergent validity, it also highlights the benefits of using multiple approaches to detection in order to have the most complete picture of the prevalence of IER in a given data set.

Some methods are not entirely practical when considering organizational surveys and may even be inappropriate in the context of performance feedback surveys. For example, the infrequency approach, which presents bogus items that respondents would be highly unlikely to truthfully endorse, has been found to be effective at determining which respondents are engaging in IER (Huang et al., 2015), but could potentially be demotivating or frustrating to raters in the case of performance feedback. Although Huang et al. (2015) found that respondents did not have negative reactions to the use of these bogus items, the survey in question was not based around performance and was not tied to employee consequences – their findings might have been different had the survey

been tied to employee outcomes. The detection mechanisms to be used in this study are all examined *after* data collection; they have no effect on survey design or participant experience. These detection mechanisms are discussed in greater detail below.

Outlier analysis. Data sets are often examined for outliers in order to flag those responses that constitute extreme scores, falling outside of the pattern reflected by the rest of the reported responses to a given item. These univariate analyses do not take into account the entirety of the pattern of an individual's responses, however, and do not consider the fact that someone might respond very high or very low across an entire survey. Meade & Craig (2012) recommend the use of a multivariate outlier index, specifically Mahalanobis distance, for the detection of IER. Instead of flagging individual responses which fall outside the range of expected responses for one item, Mahalanobis distance considers an individual's full pattern of responses when determining what constitutes an outlier. In a latent profile analysis, Meade & Craig (2012) found that the mean Mahalanobis distance value for the non-IER class was 53.09, while the IER class had a mean of 100.20, demonstrating the usefulness of this index in identifying those respondents engaging in IER.

Consistency indices. It is to be expected that individuals should respond to items containing similar content in a consistent manner when applying sufficient effort to a survey. Following this line of reasoning, consistency indices match items based on the relationships between their content, whether these relationships are built into the survey at the time of its design or are determined empirically after data collection. The *post-hoc* empirical method has been used with success in various studies of IER as it applies to organizational surveys (Huang et al., 2012; Meade & Craig, 2012).

By examining within-person correlations between items, one can gauge consistency in similarity of responses on items with high positive correlations (psychometric synonyms; Meade & Craig, 2012) or differences in responses on those items with high negative correlations (psychometric antonyms; Johnson, 2005 as cited in Meade & Craig, 2012). It is to be expected that when responding with sufficient effort, a rater will answer positively correlated items similarly, and negatively correlated items somewhat differently. Substantial deviation from this pattern is indicative of IER. Each of these indices functions to demonstrate that respondents are answering all survey questions in a consistent manner.

Response pattern indices. On a scale with questions targeting different behavioral aspects of employee performance, it is unlikely that even a superlative performer would merit the same rating for all behaviors described in the survey. The response pattern index identifies consecutive strings of the same responses (i.e., providing a rating of "Strongly Agree" for six items in a row). Meade & Craig (2012) and Huang et al. (2012) found that a response pattern index of this nature identifies different respondents engaging in IER than would other indices, as this is a unique form of IER. This index is used to examine the possibility that a rater would likely only truthfully endorse consecutive items with the same rating so many times in a row before it is probable that the rater is responding with little to no regard to item content.

On the whole, these detection mechanisms do not require advanced statistical skill and can be easily employed by researchers and organizational scientists in order to confirm that data were provided under effortful conditions. Therefore, there should be

further exploration into the specific forms of IER demonstrated by 360° feedback survey raters.

Research Question 3: What forms of IER do raters use when responding to 360° feedback surveys?

While 360° feedback research provides researchers and practitioners an exhaustive list of considerations to make when attempting to demonstrate the effects of feedback on individual and organizational outcomes, very little attention is paid to the actual quality of the data itself. There is much discussion of rating *accuracy* (Atwater et al., 2007), but only insofar as ensuring an organizational climate supportive of honest feedback based on actual performance. To my knowledge, the concept of screening the data resulting from 360° feedback surveys is novel to this area of research, and therefore demands exploration, regardless of the impact of the findings.

The intention of this inquiry is not to "fix" multisource data – as has been demonstrated in the vast majority of previous studies, this is not a realistic goal. Rather, it is my intention to examine a potential barrier to the accuracy and usefulness of 360° feedback ratings. It is crucial to understand whether IER plays a role in the 360° feedback process, and if so, to examine the effects of IER as they apply to performance data. This intervention can potentially improve not only the reliability of source ratings, but also the relationships between 360° feedback and other organizational outcomes for which explaining validity has been problematic.

CHAPTER 3

METHOD

Sample

The sample comes from an American subsidiary of a mid-size international pharmaceutical company. Rating targets fall into one of three categories: non-managers, managers, or executives (managers of departments/functions). At the time of each survey, rating targets and other raters were either based at the company's American headquarters in the Northeastern United States, as field employees at various locations around the country, or as employees of another subsidiary of the parent corporation. Raters are split into four categories: self, manager, direct report, and other. The "other" category consists of second-line managers, human resources professionals, trainers, peers, internal clients, and other stakeholders of the rating target.

Measures

Non-manager survey. The first 360° feedback survey was designed to assess non-managers on eight competencies related to communication skills, strategic thinking, and collaborative ability. The survey contains 34 items, with between four and five items for each competency. Each item is rated on a five-point rating scale with 1 indicating "Significant Development Needed" and 5 indicating "Role Model" (i.e., the rating target is an exemplary performer on the behavior being rated).

People manager survey. The second survey was designed to assess managers of people on eight competencies related to leadership abilities. The survey contains 35

items, with between four and five items for each competency. Each item is rated on a five-point rating scale with 1 indicating "Significant Development Needed" and 5 indicating "Role Model."

Executive development survey. The third survey was designed for high-level executives as the first step of an executive development training program. After participating in the 360° feedback survey, each rating target met with an executive coach to develop an action plan based on survey results prior to engaging in a 6-month training program. The executive development survey consists of 44 behavioral statements based on nine leadership-related competencies with between three and six items for each competency. Each item is rated on a 7-point agreement scale with 1 indicating "Strongly Disagree" and 7 indicating "Strongly Agree."

On each of the three surveys, there is a "Cannot Assess" option for each item in addition to the standard rating scale. Raters are instructed to only select "Cannot Assess" if they do not feel they have sufficient information to provide an accurate rating on that behavior.

It is not a requirement for any employee at this organization to engage in the 360° feedback process, whether as a rating target or as a rater – it is a purely developmental process. In order to serve as a rating target, an individual must be in role for at least six months and must have been in the organization for at least one year. The non-manager and manager surveys are administered by human resources whenever a rating target and his/her manager request for the process to take place. The executive development survey is administered to all those executive-level employees selected to participate in the training program prior to their first coaching sessions. Regardless of survey type, each

rating target is required to select potential raters representing their direct supervisor, peers, and direct reports, if applicable. The rating target's manager then reviews the list of potential raters to confirm that they should have had sufficient opportunity to witness the target at work, as well as suggest potential raters that might have been missing from the initial list. Each selected rater is expected to have worked with the rating target for at least six months in the past year. Once raters have been approved and submitted, the survey is sent out to all raters, including the rating target, who is required to provide self-ratings in order to receive survey results.

IER Indices

Outlier index. Meade and Craig (2012) found that Mahalanobis distance was an effective means of distinguishing those who responded with effort from those who engaged in IER. A Mahalanobis distance measure can be computed for each rater for each scale. If the correlations between the measures for each scale are statistically significant, the Mahalanobis distance measures are averaged to a single index.

Consistency index: psychometric antonyms. The psychometric antonym index is computed by creating item pairs from those items whose correlations are stronger than -0.60 (Meade & Craig, 2012, adapted from Johnson, 2005). The within-person correlations for each of these pairs are then averaged[1] to compute a Psychometric Antonym index for each rater.

Consistency index: psychometric synonyms. This index was developed by Meade and Craig (2012) in a similar manner to the psychometric antonym index, except with item-pair correlations greater than +0.60. The within-person correlation for each item pair

[1] In order to reduce the impact of sampling bias, all correlations will be averaged using Fisher's r to z transformation, which has been found to result in a less biased statistic than achieved by averaging r values (Corey, Dunlap, & Burke, 1998).

is then averaged using Fisher's r to z transformation to create a psychometric synonym index for each rater.

Response pattern index. The LongString protocol (Meade & Craig, 2012) is used to detect strings of a rater endorsing the same response option across multiple consecutive questions. In this study, the LongString index consists of the longest consecutive run of the same item response for each rater. For example, if a rater has strings of 4, 8, and 6 of the same response in a row, his or her index is 8.

Analyses

Data preparation and pre-screening analyses. In order to ensure that each scale is unidimensional, all items underwent exploratory factor analysis (EFA) to empirically construct scales (3 separate analyses – one for each survey). Based on familiarity with the survey content, it was expected that there would be fewer empirical scales for each survey than the number of competencies. The scales referred to in the analysis stage will be those uncovered by the EFA, rather than the competencies on which the surveys are based.

Research question 2 involves considering differences between the reliability indices of the pre-screened data with screened data. Prior to engaging in IER detection, I examined internal consistency for each scale and interrater reliability (ICC) for self, manager, subordinate, and other ratings for all scales on each of the three surveys.

IER detection. All IER indices were computed used the "Careless" program (Yentes, 2016) in R (R Core Team, 2015). Rather than using cut-off values to classify IER vs. non-IER raters, these categories are determined on the basis of the MM-IRT analysis for this study.

MM-IRT. As discussed in the previous chapter, I conducted MM-IRT analyses for each scale. The computer program R was used for all MM-IRT procedures.

Using the MM-IRT method employed by Rost (1991), one first fits a model under the assumption that the data represents one homogeneous population. This model is then compared with alternate models, each of which contains gradually more subpopulations. This process ends when successive models cease to produce improved model fit, at which point model fit is compared for the existing estimated models. Should a model with multiple subgroups yield better fit than the homogenous population model, it is implied that the population contains subgroups who respond to items differently. It is then possible to consider the theoretically-relevant qualitative differences across the subgroups in a matter analogous to that of determining the meaning of factors in exploratory factor analysis.

The three datasets for the proposed study involve polytomous ordinal Likert scales. Therefore, I use the partial credit model (PCM; Masters, 1982), which is an extension of the Rasch model for ordered polytomous data. As a Rasch model, the PCM assumes equal discrimination across items. Under the standard PCM as described by Carter et al. (2011), the probability of person *j* endorsing the *k*th response (*h)* option on the item *i* is

$$P(U_{ij} = h) = \frac{\exp(h\theta_j - \delta_{ih})}{\sum_{s=0}^{M} \exp(s\theta_j - \delta_{is})} \qquad (1)$$

where δ_{jh} is the intercept parameter. Essentially, each pair of response options is examined using the standard Rasch model and then considered together additively.

Using the mixture version of the PCM as described by Carter et al. (2011):

$$P(U_{ij} = h) = \sum_{g=1}^{G} \pi_g \frac{\exp(h\theta_{jg} - \delta_{ihg})}{\sum_{s=0}^{M} \exp(s\theta_{jg} - \delta_{isg})} \qquad (2)$$

The inclusion of the *g* term makes the intercept parameters class-specific.

In accordance with the method described by Rost (1991) and the seven-step model employed by Carter et al. (2011), I analyzed the PCM using equation 2 for all identified scales on each of the three surveys with an increasing number of latent classes until model fit got worse.

I then compared the models on the basis of the Consistent Aikake's Information Criterion (CAIC; Bozdogan, 1987), which has been used in other MM-IRT studies (e.g., Carter et al., 2011; Zickar et al., 2004). The CAIC favors more parsimonious models by correcting for the number of parameters estimated, as well as sample size. Lower values of the CAIC indicate better model fit. Since this step of the procedure is intended to examine relative model fit, it is logical to use a fit statistic that favors parsimony. I also examined standardized infit and outfit to determine instances of item misfit within the determined latent classes. Then, I interpreted the latent classes based on the response patterns they revealed. I used ANOVA and χ^2 tests to determine the roles of IER and rater type with regards to class membership.

Meade and Craig (2012) found evidence of three classes in their study of IER using latent profile analysis – one class for those who generally responded with effort, a second class for those engaging in IER as detected by the consistency indices, and a third class for those with higher value on the LongString index. Accordingly, it was anticipated that the current study would yield multiple subgroups of raters engaging in IER on the basis of the type of IER demonstrated.

CHAPTER 4

RESULTS

Scale Construction and Pre-Screening Statistics

Each of the three surveys underwent EFA using maximum likelihood estimation with promax rotation, as it was expected that factors would be correlated due to significant correlations between the survey items. As anticipated, the EFAs revealed fewer factors than the competencies on which the surveys were based. For each scale, the first unrotated factor accounted for at least 20% of common variance, providing evidence of unidimensionality for all scales (Reckase, 1979). All three surveys yielded interpersonal and task performance factors. In addition, the non-manager survey revealed a goal setting factor, the people manager survey revealed a coaching factor, and the executive development survey revealed team leadership and team reputation factors.

After the scales were constructed, coefficient alphas were computed for each scale. Across all three surveys, all scales yielded reliability estimates with a minimum of 0.87, suggesting strong reliability across all survey scales. The coefficient alphas can be found in Table 1.

In addition, two-way mixed average measures intraclass correlation coefficients (ICC-3) were estimated for each rater type for each scale. These values can be found in Table 2. The ICC values for the non-manager survey, people manager survey, and direct report and other raters of the executive development survey range from fair to excellent

reliability, while self and manager ratings on the executive development survey generally indicate poor reliability[2] (Cicchetti, 1994).

All IER indices were computed using the R package "Careless" (Yentes, 2016). As discussed in the methods chapter, I computed LongString, Mahalanobis d (outlier), and psychometric synonym indices. However, the psychometric antonym index could not be computed for any scale on any survey, as no items had negative correlations. Once all preliminary analyses took place, I began MM-IRT analyses.

MM-IRT

First, individuals using "Cannot Assess" ratings were eliminated from each scale's sample; this allowed for individuals to be put into classes for some scales where they had full responses, but not others when they used the "Cannot Assess" option. All models were fit using the "mixRasch" package in R (Willse, 2014). After fitting a one-class partial credit model for each scale, I gradually increased the number of classes until relative fit as per the CAIC failed to improve.

The non-manager survey yielded a three-class solution for the interpersonal scale and two-class solutions for the task performance and goal setting scales. Similarly, the people manager survey yielded a three-class solution for the interpersonal scale and two-class solutions for the task performance and coaching scales. The executive development survey yielded two-class solutions for the interpersonal, team leadership, and team reputation scales, but only a one-class solution for the task performance scale. This is somewhat inconsistent with past literature, where three-class solutions have generally been found to have the best fit (e.g., Hernández et al., 2004; Carter et al., 2011). The

[2] Cicchetti (1994) provides the following guidelines for reliability: A value below .40 indicates poor reliability, a value between .40 and .59 indicate fair reliability, a value between .60 and .74 indicate good reliability, and a value of .75 or above indicate excellent reliability.

CAIC values and best-fitting class solutions can be found in Table 3. The scale means and standard deviations for each class can be found in Table 4. All means were considerably higher than scale midpoints, providing evidence of generally lenient responding.

Next, I considered absolute fit for the various class solutions. This was accomplished by looking at item misfit rates in terms of infit and outfit for each class in each scale. The infit and outfit statistics are both based on the chi-square distribution. Infit is the weighted residual based on observed versus expected values from the model, sensitive to the individual's pattern of responses, while outfit is outlier sensitive (Linacre, 2002). The standardized (z-transformed) infit and outfit statistics for each item by class were considered, with significant p-values indicating item misfit. I first considered uncorrected fit statistics. I then performed two Bonferroni corrections: the first by number of items in the scale, and the second by number of items in the scale multiplied by number of classes in the model.

As evidenced in Table 5, uncorrected item misfit rates were high across all surveys for both infit and outfit. After performing the Bonferroni corrections, some item misfit rates dropped dramatically, such as for the interpersonal scale of the non-manager survey. However, most item misfit rates did not change dramatically post-correction, if they changed at all. This indicates that absolute fit was often poor, and therefore item parameters should be interpreted with caution. It is worth noting, however, that a previous study determined that infit and outfit in Rasch models are largely influenced by sample size and are prone to type-I error (Smith, Rush, Fallowfield, Velikova, & Sharpe,

2008). Given the large sample sizes employed in this study, it is possible that in this case, the high item misfit rates are reflective of type-I error rather than poor model fit.

I plotted threshold parameters for each scale by class to determine if these thresholds were ordered. The non-manager survey generally displayed properly ordered thresholds, as indicated by parallel lines on the threshold plots (e.g., Figure 2). There were two instances of items with disordered thresholds for one class on the non-manager survey (e.g., Figure 3). The people manager survey showed disordered thresholds, but in a fairly consistent pattern, in that the thresholds generally appeared in the same order across items on a given scale, even though this order wasn't in the same order as the rating scale options (e.g., Figures 4 and 5). The executive development survey plots also revealed disordered thresholds, but in an extremely erratic pattern across all scales and classes (e.g., Figure 6). The reasoning behind these disordered thresholds will be considered in the discussion section.

After considering fit and threshold parameters, I plotted category probability histograms for each scale by class. These histograms make response patterns visible and comparable between classes, making it possible to categorize and name unique classes. Across all surveys, each scale showed evidence of a "central tendency" class, predominantly responding in the middle of the scale, and a "lenient" class, responding at the higher end of the scale. For example, Figures 7 and 8 show the central tendency and leniency classes, respectively, of the non-manager goal setting scale. When comparing the two histograms, it is clear that two distinct response patterns emerge. For the three-class solutions, there was evidence of a "mixed lenient" class for each scale, showing generally lenient solutions, but with higher probabilities of using the full scale than the

lenient classes. Figures 9 through 11 show the category probabilities of the interpersonal scale on the non-manager survey. When comparing classes 2 and 3 (Figures 10 and 11), it is clear that lenient responses are more likely in both classes, but there is more evidence of lower scale options being used in class 3. It is worth noting that in general, responses could be characterized as fairly lenient across all classes on all surveys, with respondents tending to favor higher-value response options. This will be discussed further in the discussion section.

In addition to the histograms, class size estimates can be found in Table 6. In almost all cases, the lenient classes were larger than the central tendency classes. However, the interpersonal scale on the people manager survey had a considerably larger central tendency class (49% of the sample) than lenient class (29% of the sample).

In order to determine the consistency of class membership (and the prevalence of systematic responding), I conducted three contingency table analyses for each survey for every scale combination, except for the task performance scale in the executive development survey, which only had one class. The results of these analyses can be found in Table 7. I found significant χ^2 statistics for all but one of the scale combinations (task performance by goal setting on the non-manager survey). The φ coefficients are generally above .10 for all survey scales. Therefore, it appears that class membership is generally consistent across scales and is indicative of systematic responding.

After classes were identified, named, and checked for consistency of membership, I once again considered reliability, but by class rather than the entire sample. Table 8 shows coefficient alphas for each survey scale by class. In general, coefficient alpha went down as compared to the entire sample. This is likely due, at least in part, to the decrease

in sample size when considering individual classes versus the entire sample. Despite decreases, coefficient alpha values remained high across all classes for all scales. I also reconsidered ICC, taking into account class membership. These values can be found in Table 9. The effects were variable: in some cases, ICC went up, whereas in other cases, the value decreased. Taking the coefficient alphas and ICCs into account, it seems that class membership has little to no effect on the statistical properties of the data, addressing research question 2.

Associations with Class Membership

Rater type. In order to determine the effect of rater type (self, manager, direct report, or other rater) on class membership, I conducted a series of χ^2 analyses. The results of these analyses can be found in Table 10. With the exceptions of the non-manager interpersonal scale and the team reputation executive development scale, all scales revealed significant relationships between class membership and rater type. This indicates that rater type plays a role in one's propensity to engage in lenient versus central tendency responding.

Based upon relationships with effect sizes of .10 or higher, raters in the self and other categories tended to fall in the lenient class. In most relationships, there was no substantial difference in class membership for manager and direct report raters. This is contrary to previous findings, where direct reports were found to frequently be the most lenient raters (Bamberger et al., 2005).

IER. In order to determine the association between class membership and the different forms of IER, I ran a series of ANOVAs. In order to test the assumption of homogeneity of variance, I ran Levene's test, which was significant across all scales,

suggesting a violation of the assumption. Therefore, I used Welch's F to determine relationships and, in cases where there were three-class solutions, I used the Games-Howell post-hoc test to determine differences between classes. ANOVA results can be found in Table 11. Mean IER values and standard deviations can be found in Table 12.

Although most relationships were found to be statistically significant, effect sizes (η^2) varied in magnitude across the various scales and indices of IER[3]. Class membership predicted IER with medium to large effect sizes for several survey scales. LongString was higher in the lenient class for all scales except task performance on the people manager survey, where the central tendency class showed higher LongString values (see Figure 12). The outlier index was higher in the central tendency class for all scales except task performance on the people manager survey (see Figure 13). Across both the LongString and outlier indices, most substantial effect sizes were in the people manager survey. In addition, the task performance scales of both the non-manager and people manager surveys showed substantial effect sizes for both LongString and outlier. The psychometric synonym index, however, did not display any clear patterns in terms of which class engaged more strongly in IER across scales.

This set of analyses help us to address the first research question – what is the prevalence of IER in 360° feedback? These findings suggest that while IER is present, it is not pervasive. It was expected that three classes would emerge: a no-IER class, a high-IER class in terms of the outlier and psychometric synonym indices, and a high-IER class in terms of LongString. Contrary to what was anticipated, there was no single class characterized by a lack of IER; rather, the two main classes (central tendency and lenient)

[3] Cohen (1988) provides guidelines for magnitude of effect size as determined by η^2: .01to .05 indicate low effect size, .06 to .13 indicate medium effect size, and .14 and above indicate large effect size.

were each associated with a particular form of IER, but only in some instances. Lenient responders seemed to be more prone to engage in the type of IER that involves using the same response for numerous consecutive items. In addition, the outlier index tended to be higher for the central tendency classes, indicating a stronger presence of this particular form of IER for the central tendency class. Based upon these findings, it is not clear exactly what role IER plays in 360° feedback ratings – it is present in some instances, but does not seem to characterize any raters consistently. However, it is clear that systematic responding plays a major role in 360° feedback ratings and could potentially be affecting the accuracy of results.

CHAPTER 5

DISCUSSION

In this dissertation, I examined the presence and role of IER in 360° feedback surveys through the application of MM-IRT. The results of the study provide interesting and novel insights into rater behavior on organizational surveys. There is clear evidence that raters respond in one of several systematic patterns to 360° feedback surveys. The presence of IER was significantly different across class types for some survey scales, and the different indicators of IER tended to primarily correspond with specific classes. In addition, rater type (self, manager, direct report, or other) seems to play a role in respondent behavior. I will examine each of these findings in more detail.

First, raters tended to respond systematically, employing either a lenient (favoring higher scale values) or central tendency (favoring middle scale values) response style. This indicates that rater behavior influences the results of 360° feedback. As mentioned earlier in this paper, these surveys were comprised of items representing unique behaviors – even very strong performers are likely to have a few areas of weakness, and poor performers are likely to have a few areas of strength. Raters tended to be consistent in the ratings that they gave across all items, as class membership tended to be consistent across survey scales. For example, a rater that was in the lenient class for the interpersonal scale of the non-manager survey tended to be in the lenient class for the task performance and goal setting scales, as well. This is aligned with past findings about

rater behavior on 360° feedback surveys, particularly with regard to rater leniency (e.g., Antonioni & Park, 2001; Barr & Raju, 2003).

On the interpersonal scales of the non-manager and people manager surveys, a mixed lenient response style also emerged, tending to favor higher scale values with a somewhat stronger preference than other classes for the rest of the scale. It is not clear why this third class emerged, or why it did not emerge for the interpersonal scale of the executive development survey. It is possible that this is related to the highly variable nature of interpersonal relationships: whereas a concept such as task performance will likely remain relatively consistent in various raters' eyes due to its objective nature, the interpersonal construct is more likely to elicit different types of responses from different individuals, which may have been the reason for the third class.

In addition, the task performance scale of the executive development survey was the only scale to yield a one-class solution. This was highly unexpected based on previous research using MM-IRT to examine rater behavior. It is possible that there was a one class solution due to the rating targets for this survey, high-level executives. These individuals have a larger focus on coaching, strategic planning, and larger management duties rather than actually performing tasks as a lower level employee would. This could have resulted in similar responses across the various raters for these rating targets, resulting in a single latent class.

The consistency of class membership is one of the more crucial findings of this dissertation. It suggests that, although not necessarily related to IER, there are strong rater effects present in 360° feedback ratings. This could be indicative of a lack of rater concern or buy-in for the performance appraisal process, or even rater malaise with the

process altogether. This provides some credence for the movement to eliminate the performance appraisal process altogether, which has become a trend in organizations worldwide (Rock & Jones, 2015). If employees are not providing accurate ratings on performance appraisal surveys, and instead are consistently relying on systematic response styles when providing their ratings, this provides evidence to support the elimination of such processes, or, at a minimum, a dramatic shift in how we conduct employee performance appraisal. Based upon the findings of this dissertation, it is clear that raters are consistently relying on response styles rather than actual behavioral incidents in order to provide performance ratings.

Next, it was interesting to see that in some cases, different forms of IER tended to correspond with patterns of responding detected by MM-IRT. Table 13 shows the pattern of correspondence across survey scales and IER type. The LongString index tended to be higher in the lenient class, and the outlier index tended to be higher in the central tendency class. There was no clear pattern for the psychometric synonym index. It was anticipated that there would be a class displaying more IER in terms of outlier and psychometric synonym indices, as well as a class higher in terms of the LongString index, as was found in Meade and Craig's (2012) latent profile analysis. My findings were generally consistent with the earlier research – the lenient class was usually associated with LongString, while the central tendency class was usually associated with the outlier index. However, there was no class associated with a lack of IER, which is different than what was anticipated. This makes it impossible to say that any group of raters is more prone to engage in IER than others – rather, just that certain types of IER

are higher among raters engaging in particular response patterns. This limits the usefulness of the overall findings.

In terms of IER, mean LongString values for the mixed lenient tended to be lower than the other classes, or in between the central tendency and lenient classes. For the outlier index, however, values were higher for the mixed lenient class than the other two classes. Although effect sizes were low for these analyses, this suggests that the mixed lenient class may still represent a systematic response style, rather than just "normal" or unbiased rating.

The psychometric synonym index was not consistent in its association with class (see Table 13), making it difficult to characterize this index. In addition, this index had the most instances of nonsignificant associations with class membership and showed the lowest effect sizes, with many η^2 values at .01 or below. Therefore, in this dissertation, the psychometric synonym index provided relatively little value in identifying IER or systematic responding. However, this is likely due to the properties of the survey in question and the organization from which it stemmed. The survey items were all positively worded and positively correlated, making it straightforward for the rater to detect that all items are related to positive attributes of the rating target. In addition, it is the practice of the organization to use positively worded items across all surveys, so raters were likely accustomed to a survey written in this style. Had the survey been written with some negatively worded items, the psychometric synonym index may have provided more value. In addition, it may have been possible to use the psychometric antonym index to provide further insights into IER.

It is interesting to note that for the task performance scale of the people manager survey, there were substantial effect sizes for the LongString and outlier indices with class membership, yet in the opposite direction as other scales – LongString was associated with the central tendency class, while outlier was associated with the lenient class. It is difficult to speculate as to what caused this survey scale to function differently than the others. It is also interesting that the task performance scale of the non-manager survey showed substantial effect sizes for both the LongString and outlier indices. This calls to attention that the task performance scale might be different in general. When considering the content of all scale types, the task performance scale is the only scale that deals with objective, results-oriented behavior. The other types of scales, such as interpersonal and coaching, are based on "softer" skills, and individuals' impressions of these skills can vary based on relationship. Therefore, it is possible that task performance showed consistently strong results due to the nature of the item content.

The results were largely consistent across the various scales of three distinct surveys with different rating targets, providing further evidence of consistency for these findings. The rating targets for each of the surveys were at different levels of their careers, and the surveys were administered in different contexts: the non-manager survey is an ad hoc survey for individual contributors at lower levels of the organization, the people manager survey is an ad hoc survey for managers at higher levels of the organization with direct reports, and the executive development survey is for very high-level senior professionals participating in a large-scale training program. Given that class membership's relationship to IER was generally the same across the three types of

surveys, this provides some generalizability for the results in terms of the rating target in question – IER should not be affected by the type of rating target of a survey.

It is difficult to speak to the prevalence of IER as the result of these findings (i.e., what is the rate of IER in the data?), as well as which types of raters were more prone to engaging in IER. It was anticipated that IER would be different for the various rater types. While the χ^2 tests showed evidence of differential relationships between rater type and class membership, it seemed as though self and other raters tended to fall into the lenient classes, without a clear pattern for manager and direct report raters. As mentioned previously, this is contrary to findings of earlier studies on rater leniency.

Unlike in previous IER studies, it was difficult to see if the separation of IER-afflicted data affected the statistical properties of the data set overall. This is likely because both classes (as most models had two-class solutions) each had a prevalent form of IER: LongString in the lenient class, and outlier in the central tendency class. Although this study was inconclusive in determining how the screening of IER could impact the quality of the data, it did reveal the overwhelming presence of systematic responding, which suggests that interventions should take place to prevent this phenomenon from occurring, such as rater training and data screening. While a single IER-focused class did not emerge, as was initially anticipated, both classes revealed some evidence of IER. In addition, all respondents (with the exception of those respondents to the task performance scale of the executive development survey) fell into a class indicating some form of systematic responding.

This study makes numerous contributions to research and practice alike. First, it further demonstrates the usefulness of MM-IRT in examining systematic differences in

response patterns. Based on my research, MM-IRT had not been applied to multisource data in the past, nor had it been applied to performance data. This study helps to prove that MM-IRT is an effective method for detecting uniqueness in different types of datasets. In addition, IER had not been studied in multisource or performance-based surveys – this study makes it clear that IER is present beyond self-report personality measures. Next, this study establishes that different forms of IER are associated with different patterns of systematic responding. This suggests that researchers and practitioners should consider response patterns, and how these patterns alone might provide clues as to what form of IER to expect to find in the data. The ubiquity of systematic responding present in all three samples in this study could provide evidence for why 360° feedback data is not always viewed as useful. For example, the fact that most respondents tended toward leniency would eliminate the ability of the feedback to provide rating targets with accurate areas of developmental opportunity, reducing the usefulness of a multisource feedback intervention.

Limitations and Directions for Future Research

There are a number of limitations to the current study. First, there are several factors that could have contributed to elevated type-I error, including the large sample sizes present in all three surveys. While IRT requires a large sample size in order to reach a solution, this large sample size also increases the likelihood of small differences appearing to be statistically significant. In addition, I underwent a large number of analyses, such as numerous chi-square tests and ANOVAs, in order to determine the significance of the findings. This large number of small-scale analyses could also contribute to type-I error. Finally, the large sample size also could have influenced the

high item misfit rate across all three surveys – χ^2 values for these tests tended to be significant, which is a characteristic of large sample sizes for Rasch analyses that can be indicative of high type-I error.

Next, these results are specific to one organization. In order for these findings to be generalizable on a broader level, similar analyses should be conducted at organizations of varying sizes in different industries across the world. It is possible, and worth exploring, that insufficient effort responding could be an individual or cultural phenomenon within an organization, or even within parts of an organization, and future research should focus on the personal and organizational factors that could potentially lead to the occurrence of IER. These factors could potentially include trust in management, individual performance level, or demographic characteristics. It is also would have been possible to find more generalizable results using data from a commercially available multisource feedback assessment, such as BENCHMARKS™, which includes assessment results from numerous organizations.

In addition, this particular study was unable to examine the psychometric antonym index due to positive correlations between all items on each survey, regardless of scale membership. This index has proved to be useful in previous studies of IER (e.g., Meade & Craig, 2011) and might have been able to provide interesting insights to the current study had the survey been constructed differently. Furthermore, the psychometric synonym index yielded little information in this study, as there was no observable pattern in the index across classes. This could be, at least in part, due to high positive correlations amongst the items. Therefore, future studies should consider adding more varied content

and negatively-worded items to surveys in order to determine the effectiveness of the psychometric antonym index.

Although there was evidence of differing types of response patterns evidenced in the class histograms, responses still tended to be lenient across all classes, with extremely limited use of the lower end of the scale. This likely contributed to the high item misfit rates for most of the scales across surveys. This could possibly be improved upon in future research by having raters go through training or by providing more specific behaviorally anchored rating scales in the assessment.

The high item misfit rates, even after Bonferroni corrections, also provide evidence that relative fit indices, such as the CAIC, are not sufficient to determine model fit. Although all models selected in this study had favorable model fit as per the CAIC, most class solutions showed poor model fit in terms of standardized infit and outfit, making it crucial that parameter estimates are interpreted cautiously. This draws attention to the importance of considering both relative and absolute fit when interpreting results.

On a similar note, many models showed disordered thresholds. The disordered thresholds could be attributed to the choice of model, as disordered thresholds are not an uncommon issue when using Rasch-based models. According to Adams, Wu, and Wilson (2012), disordered thresholds are not necessarily a problem and are not always an indication of model misfit. The disordered thresholds could be attributed to the low probability of responses in certain categories, as was the case in this study – responses tended to be lenient, with very little usage of the lower end of the scale.

While this study makes it clear that there is evidence of IER, it was difficult to pinpoint the rate with which IER occurred. Future studies should generate rule-of-thumb

cut-off values to determine the presence of IER. In this way, it might be possible to screen and remove respondents who show a great deal of evidence of having engaged in IER.

Conclusion

This study used MM-IRT to find evidence of IER in 360° feedback survey data. Although the presence of IER was somewhat limited, it appeared that almost all raters engaged in some form of systematic responding, whether lenient or central tendency. In addition, raters tended to use the lower end of rating scales on an extremely limited basis. This suggests that (a) performance ratings are not entirely accurate, and (b) raters may be somewhat disillusioned with the performance appraisal process. This could potentially speak to some of the issues and common complaints associated with multisource feedback, namely the lack of improvement in rating target performance post-feedback. Overall, although IER was not prevalent, this study provides clear evidence for the presence of rater effects on 360° feedback surveys and the need to further investigate rater behavior.

REFERENCES

Adams, R.J., Wu, M.L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*, 547-573. doi:10.1177/0013164411432166

Adkins, A. (2016, June 10). U.S. employee engagement slips below 33% in May [Web log post]. Retrieved from http://www.gallup.com/poll/192575/employee-engagement-slips-below-may.aspx

Allen, T.D., Facteau, J.D., & Facteau, C.L. (2004). Structured interviewing for OCB: Construct validity, faking, and the effects of question type. *Human Performance, 17*, 1-24. doi:10.1207/s15327043HUP1701_1

Alliger, G.M., Lilienfeld, S.O., & Mitchell, K.E. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7*, 32-39. doi:10.1111/j.1467-9280.1996.tb00663.x

Antonioni, D., & Park, H. (2001). The relationship between rater affect and three sources of 360-degree feedback ratings. *Journal of Management, 27*, 479-495. doi:10.1177/014920630102700405

Archer, R.P., Handel, R.W., Lynch, K.D., & Elkins, D.E. (2002). MMPI-A validity scales uses and limitations in detecting varying levels of random responding. *Journal of Personality Assessment, 78*, 417-431. doi:10.1207/s15327752jpa7803_03

Arthur, W., Glaze, R.M., Jarrett, S.M., White, C.D., Schurig, I., & Taylor, J.E. (2014). Comparative evaluation of three situational judgment test response formats in terms of

construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology, 99,* 535-545. doi:10.1037/a0035788

Ashford, S.J. (1989). Self-assessments in organizations: A literature review and integrative model. In L.L. Cummings, & B.M. Staw (Eds.), *Research in organizational behavior, Vol. 11* (pp. 133-174). Greenwich, CT: JAI Press.

Atwater, L., & Brett, J. (2006). Feedback format: Does it influence manager's reactions to feedback? *Journal of Occupational and Organizational Psychology, 79,* 517-532. doi:10.1348/096317905X8656

Atwater, L.E., Brett, J.F, & Charles, A.C. (2007). Multisource feedback: Lessons learned and implications for practice. *Human Resource Management, 46,* 285-307. doi:10.1002/hrm.20161

Austin, E.J., Deary, I.J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences, 40,* 1235-1245. doi:10.1016/j.paid.2005.10.018

Bacchiochi, J.R., & Bagby, R.M. (2006). Development and validation of the malingering discriminant function index for the MMPI-2. *Journal of Personality Assessment, 87,* 51-51. doi:10.1207/s15327752jpa8701_04

Bagby, R.M., Nicholson, R.A., Bacchiochi, J.R., Ryder, A.G., & Bury, A.S. (2002). The predictive capacity of the MMPI-2 and PAI validity scales and indexes to detect coached and uncoached feigning. *Journal of Personality Assessment, 78,* 69-86. doi:10.1207/s15327752jpa7801_05

Bailey, C., and Austin, M. (2006). 360 degree feedback and developmental outcomes: The role of feedback characteristics, self-efficacy and importance of feedback dimensions to focal

managers current role. *International Journal of Selection and Assessment, 14*, 51-66. doi:10.1111/j.1468-2389.2006.00333.x

Bamberger, P. A., Erev, I., Kimmel, M., & Oref-Chen, T. (2005). Peer assessment, individual performance, and contribution to group processes: The impact of rater anonymity. *Group & Organization Management, 30*, 344-377. doi:10.1177/1059601104267619

Barr, M.A., & Raju, N.S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods, 6*, 15-43. doi:10.1177/1094428102239424

Bowling, N.A., Bragg, C.B., Liu, M., Huang, J.L., Khazon, S., & Blackmore, C.E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality & Social Psychology, 111*, 218-229. doi:10.1037/pspp0000085

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370. doi:10.1007/BF02294361

Brown, A., Lyn, Y., & Inceoglu, I. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods, 20*, 121-148. doi:10.1177/1094428116668036

Carter, N.T., Dalal, D.K., Lake, C.J., Lin, B.C., & Zickar, M.J. (2011). Using mixed-model item response theory to analyze organizational survey responses: An illustration using the Job Descriptive Index. *Organizational Research Methods, 14*, 116-146. doi:10.1177/1094428110363309

Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. doi:10.18637/jss.v048.i06

Cicchetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized instruments in psychology. *Psychological Assessment, 6,* 284-290. doi:10.1037/1040-3590.6.4.284

Cohen, J. (1988). *Statistical power analyses for the behavioral sciences (2nd ed.)* Hillsdale, NJ: Lawrence Erlbaum Associates.

Corey, D.M., Dunlap, W.P., & Burke, M.J. Averaging correlations: Expected values and bias in combined Pearson *r*s and Fisher's *z* transformations. *Journal of General Psychology, 125,* 245-262. doi:10.1080/00221309809595548

Craig, S.B., & Kaiser, R.B. (2003). Applying item response theory to multisource performance ratings: What are the consequences of violating the independent observations assumption? *Organizational Research Methods, 6,* 44-60. doi:10.1177/1094428102239425

Day, D.V., Fleenor, J.W., Atwater, L.E., Sturm, R.E., & McKee, R.A. (2014). Advances in leader and leadership development: A review of 25 years of research and theory. *The Leadership Quarterly, 25,* 63-82. doi:10.1016/j.leaqua.2013.11.004

Detrick, P., Chibnall, J.T., & Call, C. (2010). Demand effects on positive response distortion by police officer applicants on the Revised NEO Personality Inventory. *Journal of Personality Assessment, 92,* 410-415. doi:10.1080/00223891.2010.497401

Donovan, J.J., Dwight, S.A., & Hurtz, G.M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level application faking using the randomized response technique. *Human Performance, 16,*81-106. doi:10.1207/s15327043hup1601_4

Egberink, I.J., Meijer, R.R., & Veldkamp, B.P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality, 44,* 232-244. doi:10.1016/j.jrp.2010.01.007

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychology Assessment, 16,* 20-30. doi:10.1027//1015-5759.16.1.20

Facteau, J.D., & Craig, S.B. (2001). Are performance appraisal ratings from different sources comparable? *Journal of Applied Psychology, 86,* 215-227. doi:10.1037/0021-9010.86.2.215

Hernández, A., González-Romá, V., & Drasgow, F. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model. *Journal of Applied Psychology, 89,* 687-699. doi:10.1037/0021-9010.89.4.687a

Hoffman, B.J., Gorman, C.A., Blair, C.A., Meriac, J.P., Overstreet, B., & Atchley, E.K. (2012). Evidence for the effectiveness of an alternative multisource performance rating methodology. *Personnel Psychology, 65,* 531-563. doi:10.1111/j.1744-6570.2012.01252.x

Hough, L.M., Eaton, N.K., Dunnette, M.D., Kamp, J.D., & McCloy, R.A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581-595.

Huang, J.L., Bowling, N.A., Liu, M., & Li, Y. (2014). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology, 30,* 299-311. doi:10.1007/s10896-014-9357-6

Huang, J.L., Curran, P.G., Keeney, J., Poposki, E.M., & DeShon, R.P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*, 99-114. doi:10.1007/s10869-011-9231-8

Huang, J.L., Liu, M., & Bowling, N.A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. Advanced online publication. *Journal of Applied Psychology*. doi:10.1037/a0038510

Johnson, J.A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39,* 103-129. doi:10.1016/j.jrp.2004.09.009

Johnson, J.W., & Ferstl, K.L. (1999). The effects of interrater and self-other agreement on performance improvement following upward feedback. *Personnel Psychology, 52,* 272-303. doi:10.1111/j.1744-6570.1999.tb00162.x

Kim, K.Y., Atwater, L., Patel, P.C., & Smither, J.W. (2016). Multisource feedback, human capital, and the financial performance of organizations. *Journal of Applied Psychology, 101,* 1569-1584. doi:10.1037/apl00000125

Kluger, A.N., & DeNisi, A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254-284. doi:10.1037/0033-2909.119.254

Korman, A.K. (1970). Toward a hypothesis of work behavior. *Journal of Applied Psychology, 54*, 31-41. doi:10.1037/h0028656

Linacre, J.A. (2002). *What do infit and outfit, mean square and standardized mean?* Retrieved from https://www.rasch.org/rmt/rmt162f.htm.

Liu, M., Bowling, N.A., Huang, J.L., & Kent, T.A. (2013) Insufficient effort responding to surveys as a threat to validity: The perceptions and practices of SIOP members. *The Industrial-Organizational Psychologist, 51*(1), 32-38.

London, M., & Smither, J.W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance-related outcomes? Theory-based applications and directions for research. *Personnel Psychology, 48,* 803-839. doi:10.1111/j.1744-6570.1995.tb01782.x

Maij-de Meij, A.M., Kelderman, H, & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32,* 611-631. doi:10.1177/0146621607312613

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. doi:10.1007/BF02296272

Maurer, T.J., Raju, N.S., & Collins, W.C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology, 83*, 693-702. doi:10.1037/0021-9010.83.5.693

Meade, A.W. & Craig, S.B. (2012) Identifying careless responses in survey data. *Psychological Methods, 17*, 437-455. doi:10.1037/a0028085

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-118. doi:10.2307/1164960

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*, 351-363. doi:10.1177/014662169301700403

Murphy, K.R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial Organizational Psychology, 1*, 148-160. doi:10.1111/j.1754-9434.2008.00030.x

Ng, K.Y., Koh, C., Ang, S., Kennedy, J.C., & Chan, K.Y. (2011). Rating halo and leniency in multisource feedback ratings: Testing cultural assumptions of power distance and individualism-collectivism. *Journal of Applied Psychology, 96,* 1033-1044. doi:10.1037/a0023368

Oh, I., & Berry, C.M. (2009). The five-factor model of personality and managerial performance: Validity gains through the use of 360 degree performance ratings. *Journal of Applied Psychology, 94,* 1498-1513. doi:10.1037/a0017221

R Core Team (2015). R: A language and environment for statistical computing [Computer software]. Vienna, Austria. https://www.R-project.org/

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,* 353-368. doi:10.1177/014662169501900405

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Education Statistics, 4*(3), 207-230. doi:10.2307/1164671

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185-205. doi:10.1037/1082-989X.8.2.185

Rock, D., & Jones, B. (2015). Why more and more companies are ditching performance ratings. *Harvard Business Review Digital Articles,* 2-4.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44,* 75-92. doi:10.1111/j.2044-8317.1991.tb00951.x

Rost, J., & von Davier, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement, 18,* 171-182. doi:10.1177/014662169401800206

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8,* 33. http://doi.org/10.1186/1471-2288-8-33,

Smither, J.W., London, M., & Reilly, R.R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology, 58,* 33-66. doi:10.1111/j.1744-6570.2005.514_1.x

Smither, J.W., London, M., & Richmond, K.R. (2005). The relationship between leaders' personality and their reactions to and use of multisource feedback: A longitudinal study. *Group & Organization Management, 30,* 181-210. doi:10.1177/1059601103254912

Snell, A.F., Sydell, E.J., & Lueke, S.B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9,* 219-242. doi:10.1016/s1053-4822(99)00019-4

Spiel, C., & Glück, J. (1998). Item response models for assessing change in dichotomous items. *International Journal of Behavioral Development, 22,* 517-536. doi:10.1080/016502598384252

Willse, J. R. (2014). mixRasch: Mixture Rasch Models with JMLE. R package version 1.1. http://CRAN.R-project.org/package=mixRasch

Woehr, D. J., Sheehan, M. K., & Bennett, J. W. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal Of Applied Psychology, 90,* 592-600. doi:10.1037/0021-9010.90.3.592

Yentes, R.D. (2016). Careless: Procedures for Computing Indices of Careless Responding. R package version 1.0. http://github.com/ryentes/careless

Zickar, M.J., Gibby, R.E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7,* 168-190. doi:10.1177/1094428104263674

Zimmerman, R.D., Mount, M.K., & Goff, M. (2008). Multisource feedback and leaders' goal performance: Moderating effects of rating purpose, rater perspective, and performance dimension. *International Journal of Selection & Assessment, 16,* 121-133. doi:10.1111/j.1468-2389.2008.00417.x

Table 1. Reliability Estimates

Survey	Scale	Coefficient Alpha	Items	n
Non-Manager	Interpersonal	.96	11	2460
	Task Performance	.97	16	1386
	Goal Setting	.92	4	1176
People Manager	Interpersonal	.94	9	3414
	Task Performance	.94	8	2984
	Coaching	.93	7	2437
Executive Development	Interpersonal	.94	13	2208
	Team Leadership	.90	8	1647
	Task Performance	.92	10	2016
	Team Reputation	.87	4	2386

Table 2. Intraclass Correlation Coefficient Estimates by Rater Type

		Self		Manager		Direct Reports		Others	
Survey	Scale	ICC	n	ICC	n	ICC	n	ICC	n
Non-Manager	Interpersonal	.461	173	.617	150	-	-	.688	2137
	TP	.434	163	.497	126	-	-	.666	1097
	GS	.584	165	.557	145	-	-	.780	866
People Manager	Interpersonal	.421	252	.469	216	.681	1105	.608	1841
	TP	.452	249	.517	220	.708	1067	.659	1456
	Coaching	.422	249	.512	205	.701	1027	.651	956
Executive Development	Interpersonal	.255	123	.309	129	.601	728	.531	1228
	TL	.294	140	.367	119	.555	669	.520	719
	TP	.343	141	.335	130	.595	657	.515	1088
	TR	.521	142	.441	132	.607	792	.637	1320

ICC estimated using two-way mixed effects model. ICC estimates presented are single measures. TP=Task Performance. GS=Goal Setting. TL=Team Leadership. TR=Team Reputation.

Table 3. MM-IRT Model Fit Statistics

		Number of Classes			
Survey	Scale	1	2	3	4
Non-Manager	Interpersonal	34667.53	34475.69	**34382.40**	34574.91
	TP	29821.68	**29068.99**	29407.38	-
	Goal Setting	4110.165	**4095.931**	4306.117	-
People Manager	Interpersonal	43123.12	42435.50	**42311.04**	42629.12
	TP	29176.44	**28771.28**	29069.68	-
	Coaching	21810.79	**21388.75**	21435.97	-
Executive Development	Interpersonal	48748.20	**48572.92**	49097.34	-
	TL	22053.36	**21910.56**	22349.76	-
	TP	**31619.73**	31680.40	-	-
	TR	12219.15	**12023.99**	12131.34	-

CAIC = Consistent Aikake's Information Criterion. TP=Task Performance. TL=Team Leadership. TR=Team Reputation. Bolded CAIC values signify the best-fitting latent class solution.

Table 4. Means and Standard Deviations for Survey Scales by Class

		Class 1		Class 2		Class 3	
Survey	Scale	Mean	SD	Mean	SD	Mean	SD
Non-Manager	Interpersonal	2.77	.86	2.91	.77	2.77	.83
	TP	2.73	.72	2.82	.86	-	-
	Goal Setting	2.31	.72	2.79	.64	-	-
People Manager	Interpersonal	2.86	.89	2.77	.75	2.71	.89
	TP	2.75	.83	2.53	.78	-	-
	Coaching	2.75	.98	2.77	.75	-	-
Executive Development	Interpersonal	4.98	.99	5.00	1.07	-	-
	TL	4.75	1.16	4.88	.79	-	-
	TP	5.04	.89	-	-	-	-
	TR	4.82	.95	4.85	.84	-	-

TP=Task Performance. TL=Team Leadership. TR=Team Reputation

Table 5. Item Misfit Rate by Class

	Scale		Class 1		Class 2		Class 3	
			Infit	Outfit	Infit	Outfit	Infit	Outfit
Non-Manager	Interpersonal	Uncorrected	7/11	6/11	8/11	6/11	7/11	7/11
		Corrected[1]	2/11	3/11	7/11	5/11	5/11	4/11
		Corrected[2]	2/11	3/11	3/11	5/11	3/11	3/11
	TP	Uncorrected	11/16	11/16	9/16	6/16	-	-
		Corrected[1]	4/16	4/16	7/16	5/16	-	-
		Corrected[2]	3/16	4/16	5/16	4/16	-	-
	Goal Setting	Uncorrected	3/4	3/4	3/4	3/4	-	-
		Corrected[1]	3/4	2/4	3/4	2/4	-	-
		Corrected[2]	3/4	2/4	3/4	2/4	-	-
People Manager	Interpersonal	Uncorrected	4/9	3/9	7/9	5/9	7/9	3/9
		Corrected[1]	4/9	3/9	6/9	5/9	3/9	3/9
		Corrected[2]	3/9	3/9	6/9	5/9	3/9	3/9
	TP	Uncorrected	6/8	6/8	6/8	4/8	-	-
		Corrected[1]	5/8	5/8	4/8	4/8	-	-
		Corrected[2]	5/8	5/8	4/8	4/8	-	-
	Coaching	Uncorrected	2/7	1/7	6/7	6/7	-	-
		Corrected[1]	2/7	1/7	5/7	6/7	-	-
		Corrected[2]	2/7	1/7	5/7	6/7	-	-
Executive Development	Interpersonal	Uncorrected	5/13	5/13	8/13	9/13	-	-
		Corrected[1]	3/13	3/13	7/13	7/13	-	-
		Corrected[2]	3/13	3/13	6/13	7/13	-	-
	TL	Uncorrected	3/8	5/8	4/8	5/8	-	-
		Corrected[1]	3/8	3/8	1/8	3/8	-	-
		Corrected[2]	3/8	3/8	1/8	3/8	-	-
	TP	Uncorrected	9/10	6/10	-	-	-	-
		Corrected[1]	7/10	6/10	-	-	-	-
		Corrected[2]	7/10	6/10	-	-	-	-
	TR	Uncorrected	2/4	1/4	2/4	2/4	-	-
		Corrected[1]	2/4	1/4	1/4	2/4	-	-
		Corrected[2]	1/4	1/4	1/4	2/4	-	-

TP=Task Performance. TL=Team Leadership. TR=Team Reputation. Corrected[1] signifies Bonferroni correction by number of items. Corrected[2] signifies Bonferroni correction by number of items*number of classes.

Table 6. MM-IRT Latent Class Size Estimate (π) by Class Type

Survey	Scale	Class Type		
		Central Tendency	Lenient	Mixed Lenient
Non-Manager	Interpersonal	.29	.42	.30
	Task Performance	.40	.60	-
	Goal Setting	.45	.55	-
People Manager	Interpersonal	.49	.29	.21
	Task Performance	.25	.75	-
	Coaching	.27	.73	-
Executive Development	Interpersonal	.41	.59	-
	Team Leadership	.41	.59	-
	Team Reputation	.33	.67	-

Table 7. χ^2 and φ Coefficients of Class Consistency

		χ^2	df	φ
Non-Manager	Interpersonal*TP (3x2)	37.084*	2	.18
	Interpersonal*Goal Setting (3x2)	15.106*	2	.12
	TP*Goal Setting (2x2)	.000	1	-
People Manager	Interpersonal*TP (3x2)	104.137*	2	.20
	Interpersonal*Coaching (3x2)	113.127*	2	.23
	TP*Coaching (2x2)	111.419*	1	.23
Executive Development	Interpersonal*TL (2x2)	86.816*	1	.25
	Interpersonal*TR (2x2)	15.190*	1	.09
	TL*TR (2x2)	10.208*	1	.09

TP=Task Performance. TL=Team Leadership. TR=Team Reputation. *p<.01.

Table 8. Reliability Estimates by Class

Survey	Scale	Coefficient Alpha		
		Central Tendency	Lenient	Mixed Lenient
Non-Manager	Interpersonal	.95	.96	.95
	TP	.92	.97	-
	Goal Setting	.91	.87	-
People Manager	Interpersonal	.91	.95	.91
	TP	.94	.81	-
	Coaching	.79	.94	-
Executive Development	Interpersonal	.96	.92	-
	Team Leadership	.80	.93	-
	Team Reputation	.88	.86	-

TP=Task Performance.

Table 9. Intraclass Correlation Coefficient Estimates by Rater Type by Class

		Self			Manager			Direct Reports			Others		
Survey	Scale	CT	L	ML	CT	L	ML	CT	L	ML	CT	L	ML
	I	.356	.436	.512	.620	.658	.650	-	-	-	.628	.688	.631
NM	TP	.253	.529	-	.338	.659	-	-	-	-	.459	.698	-
	GS	.665	.555	-	.643	.572	-	-	-	-	.731	.627	-
	I	.406	.500	.415	.350	.615	.533	.580	.744	.554	.508	.685	.517
PM	TP	.518	.191	-	.667	.187	-	.701	.401	-	.674	.359	-
	C	.202	.513	-	.276	.610	-	.404	.726	-	.296	.662	-
	I	.398	.192	-	.459	.268	-	.738	.532	-	.618	.454	-
ED	TL	.068	.426	-	.310	.427	-	.382	.688	-	.296	.604	-
	TR	.482	.550	-	.481	.545	-	.610	.535	-	.655	.642	-

ICC estimated using two-way mixed effects model. ICC estimates presented are single measures. NM=Non-Manager. PM=People Manager. ED=Executive Development. I=Interpersonal. TP=Task Performance. GS=Goal Setting. C=Coaching. TL=Team Leadership. TR=Team Reputation. CT=Central Tendency. L=Lenient. ML=Mixed Lenient.

Table 10. χ^2 and φ Coefficients of Rater Type by Class Membership

		χ^2	df	φ
Non-Manager	Rater Type*Interpersonal (3*3)	3.243	4	.04
	Rater Type*TP (3*2)	10.423*	2	.09
	Rater Type*Goal Setting (3*2)	18.257*	2	.13
People Manager	Rater Type*Interpersonal (4*3)	35.968*	6	.11
	Rater Type*TP (4*2)	39.642*	3	.12
	Rater Type*Coaching (4*2)	49.587*	3	.15
Executive Development	Rater Type*Interpersonal (4*2)	33.005*	3	.13
	Rater Type*TL (4*2)	30.723*	3	.14
	Rater Type*TR (4*2)	2.497	3	.04

TP=Task Performance. TL=Team Leadership. TR=Team Reputation. *p<.01.

Table 11. One-Way Analysis of Variance of IER by Class

Survey	IER Type	Scale	df_{model}	df_{error}	Welch's F	η^2
Non-Manager	LongString	Interpersonal	2	1372.64	11.385**	.01
		Task Performance	1	1144.87	273.943**	.13
		Goal Setting	1	914.15	2.957	.00
	Outlier	Interpersonal	2	1345.33	11.424**	.01
		Task Performance	1	895.85	353.074**	.23
		Goal Setting	1	843.47	2.837	.00
	Synonym	Interpersonal	2	1340.03	11.925**	.01
		Task Performance	1	922.24	.286	.00
		Goal Setting	1	767.66	7.796**	.01
People Manager	LongString	Interpersonal	2	1627.93	149.354**	.07
		Task Performance	1	2097.31	549.873**	.07
		Coaching	1	1693.25	294.696**	.67
	Outlier	Interpersonal	2	1381.72	57.716**	.04
		Task Performance	1	702.41	319.655**	.12
		Coaching	1	649.36	262.423**	.12
	Synonym	Interpersonal	2	1429.78	7.544**	.00
		Task Performance	1	821.08	5.569*	.00
		Coaching	1	830.34	8.803**	.00
Executive Development	LongString	Interpersonal	1	1267.68	75.626**	.04
		Team Leadership	1	1489.13	162.483**	.07
		Team Reputation	1	1831.24	82.326**	.02
	Outlier	Interpersonal	1	2073.89	101.230**	.04
		Team Leadership	1	730.41	144.769**	.11
		Team Reputation	1	855.33	12.990**	.01
	Synonym	Interpersonal	1	1708.17	7.417**	.00
		Team Leadership	1	1129.49	18.706**	.01
		Team Reputation	1	1013.14	.001	.00

The Games-Howell test for non-manager Interpersonal by LongString revealed significant differences between the lenient and mixed lenient classes. The Games-Howell test for non-manager Interpersonal by Outlier revealed significant differences between the central tendency and lenient class, and between the lenient class and mixed lenient class. The Games-Howell test for non-manager Interpersonal by Synonym revealed significant differences between the central tendency and lenient class, and between the central tendency and mixed lenient class. The Games-Howell test for people manager Interpersonal by LongString revealed significant differences between the central tendency and lenient classes, and the lenient and mixed lenient classes. The Games-Howell test for people manager Interpersonal by Outlier revealed significant differences between the central tendency and lenient classes, and between the lenient class and mixed lenient class. The Games-Howell test for people manager Interpersonal by Synonym revealed significant differences between the lenient and mixed lenient classes. *p<.05. **p<.01.

Table 12. IER Means and Standard Deviations by Class

Survey	IER Type	Scale	Central Tendency		Lenient		Mixed Lenient	
			Mean	SD	Mean	SD	Mean	SD
Non-Manager	LongString	I	8.85	6.68	9.48	7.11	8.09	4.79
		TP	6.73	3.36	12.24	8.32	-	-
		GS	9.81	7.17	10.64	8.04	-	-
	Outlier	I	30.03	19.46	26.24	19.38	30.28	18.93
		TP	26.22	10.08	15.63	8.98	-	-
		GS	22.06	14.45	20.51	14.45	-	-
	Synonym	I	.30	.50	.42	.51	.42	.55
		TP	.31	.50	.29	.46	-	-
		GS	.34	.48	.26	.43	-	-
People Manager	LongString	I	6.82	3.37	10.34	7.30	7.08	4.31
		TP	9.95	6.80	5.60	2.64	-	-
		C	6.32	3.23	10.24	7.22	-	-
	Outlier	I	32.24	19.74	25.02	20.19	33.258	19.74
		TP	22.07	15.30	36.80	17.08	-	-
		C	32.14	15.59	19.29	13.42	-	14.84
	Synonym	I	.30	.31	.27	.35	.32	.31
		TP	.26	.33	.29	.30	-	-
		C	.27	.20	.32	.24	-	-
Executive Development	LongString	I	10.28	8.19	7.54	4.99	-	-
		TL	6.54	3.70	10.36	7.96	-	-
		TR	6.60	3.43	8.71	6.89	-	-
	Outlier	I	30.90	25.42	44.79	37.47	-	-
		TL	50.08	40.49	27.46	22.48	-	-
		TR	49.36	42.72	31.82	37.83	-	-
	Synonym	I	.31	.33	.34	.30	-	-
		TL	.35	.31	.28	.31.	-	-

TR	.31	.30	.31	.32	-	-

I=Interpersonal. TP=Task Performance. GS=Goal Setting. C=Coaching. TL=Team Leadership. TR=Team Reputation.

Table 13. Presence of IER by Class

		LongString	Outlier
Non-Manager	Interpersonal	-	-
	TP	Lenient	Central Tendency
	Goal Setting	-	-
People Manager	Interpersonal	Lenient	-
	TP	Central Tendency	Lenient
	Coaching	Lenient	Central Tendency
Executive Development	Interpersonal	-	-
	TL	Lenient	Central Tendency
	TR	-	-

TP=Task Performance. TL=Team Leadership. TR=Team Reputation.

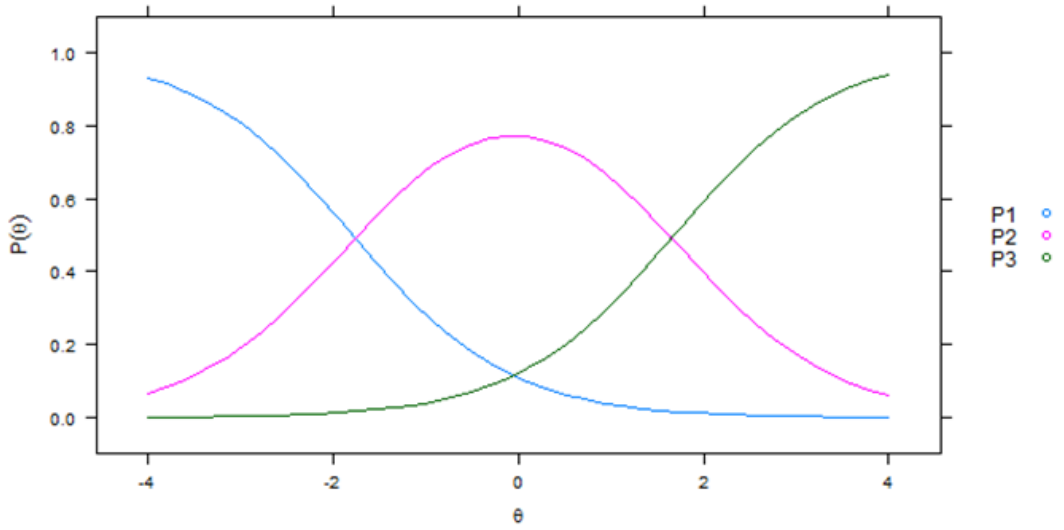

Figure 1: Example Category Characteristic Curves for a Three-Option Item

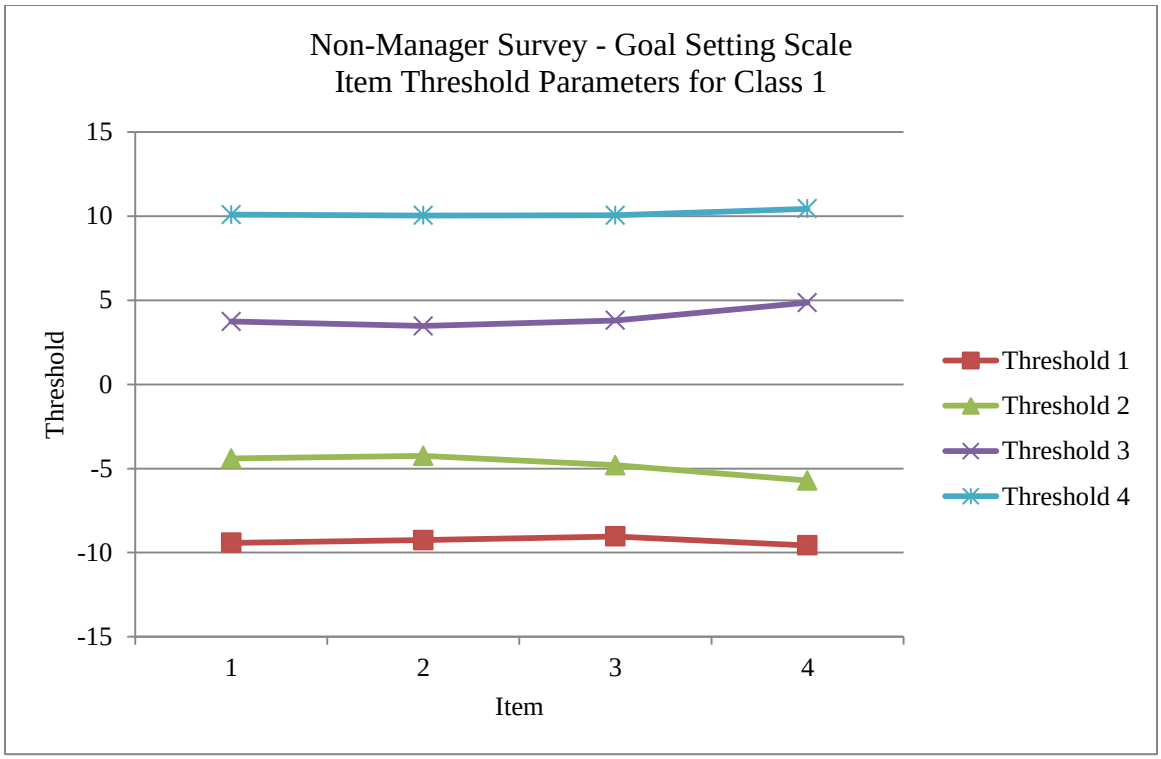

Figure 2: Threshold Parameter Plot for the Goal Setting Scale (Non-Manager)

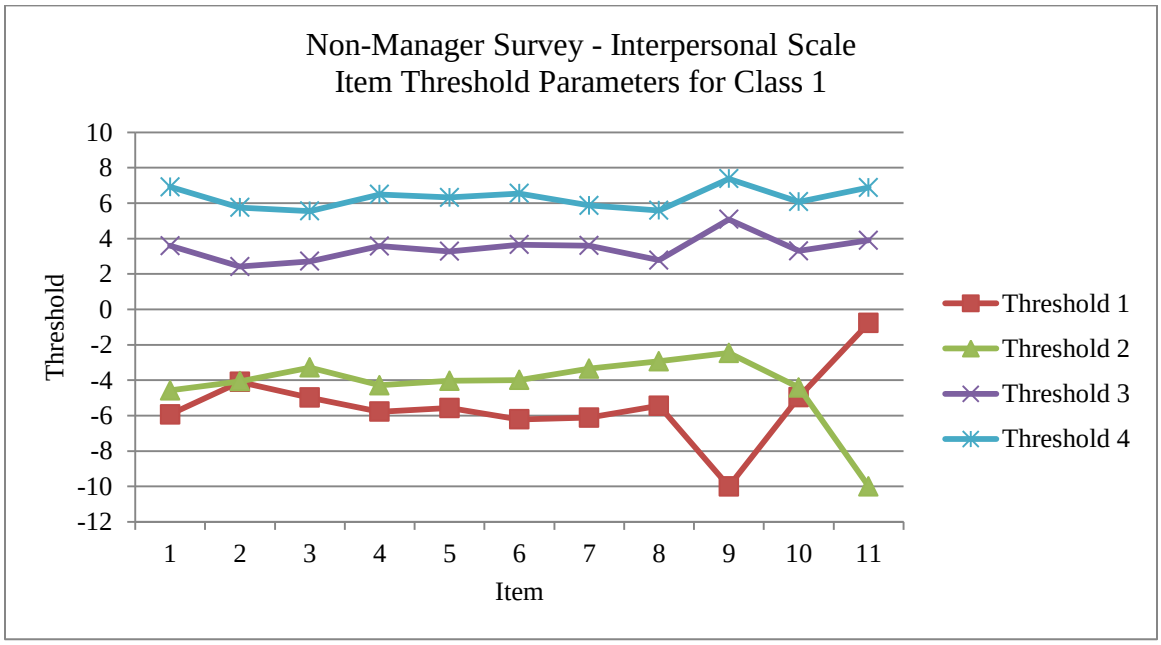

Figure 3: Threshold Parameter Plot for the Interpersonal Scale (Non-Manager)

Figure 4: Threshold Parameter Plot for the Interpersonal Scale (People Manager)

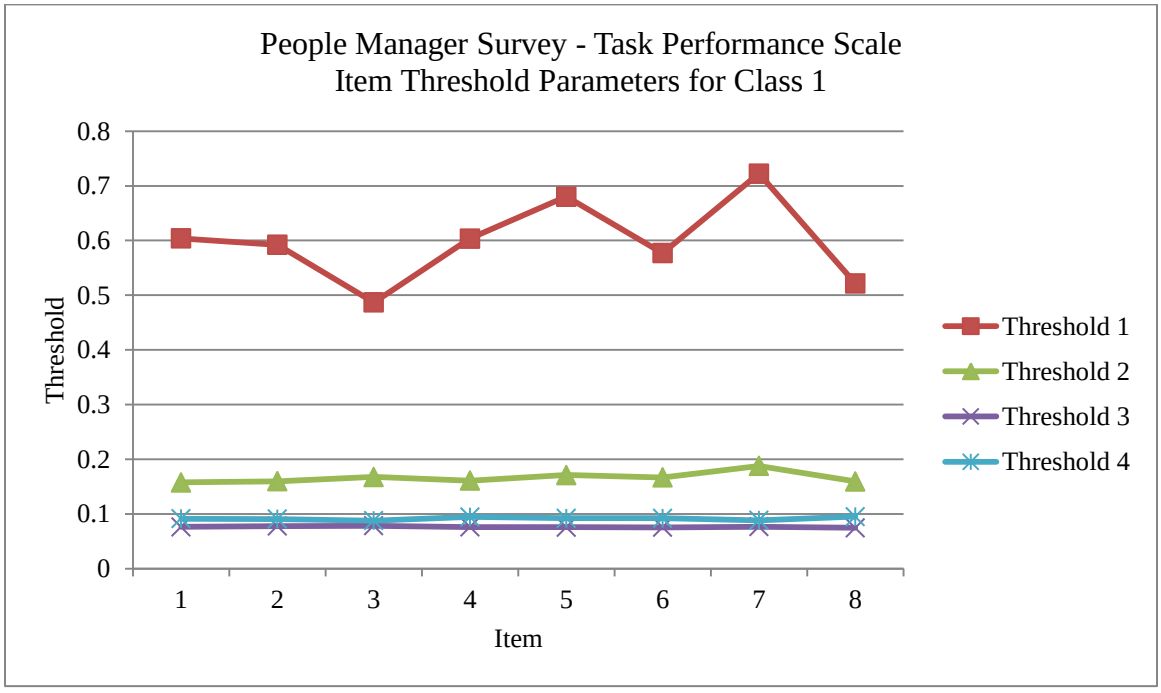

Figure 5: Threshold Parameter Plot for the Task Performance Scale (People Manager)

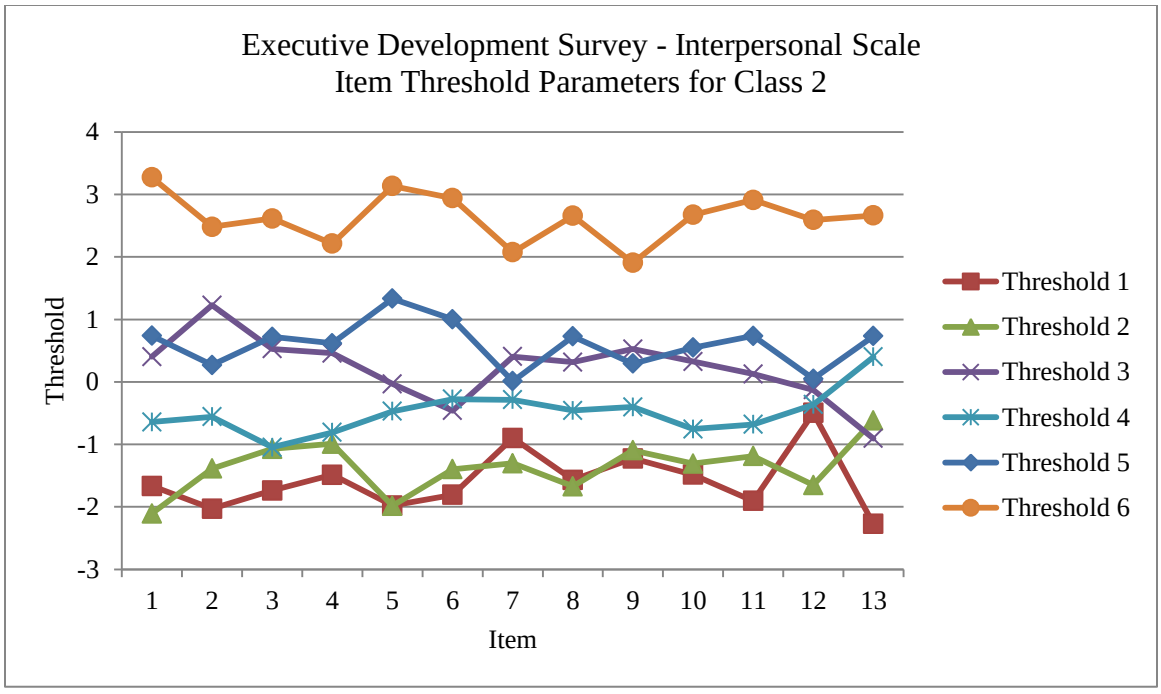

Figure 6: Threshold Parameter Plot for the Interpersonal Scale (Executive Development)

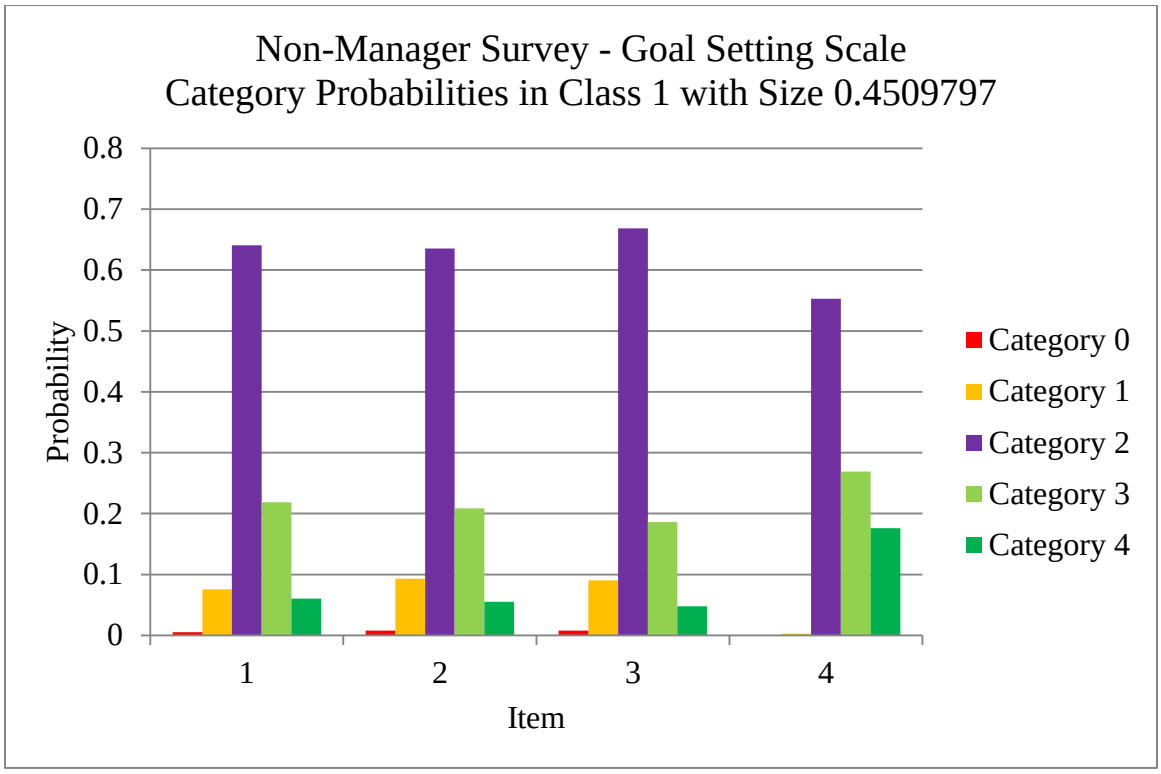

Figure 7: Category Probability Histogram for Class 1 of the Goal Setting Scale (Non-Manager)

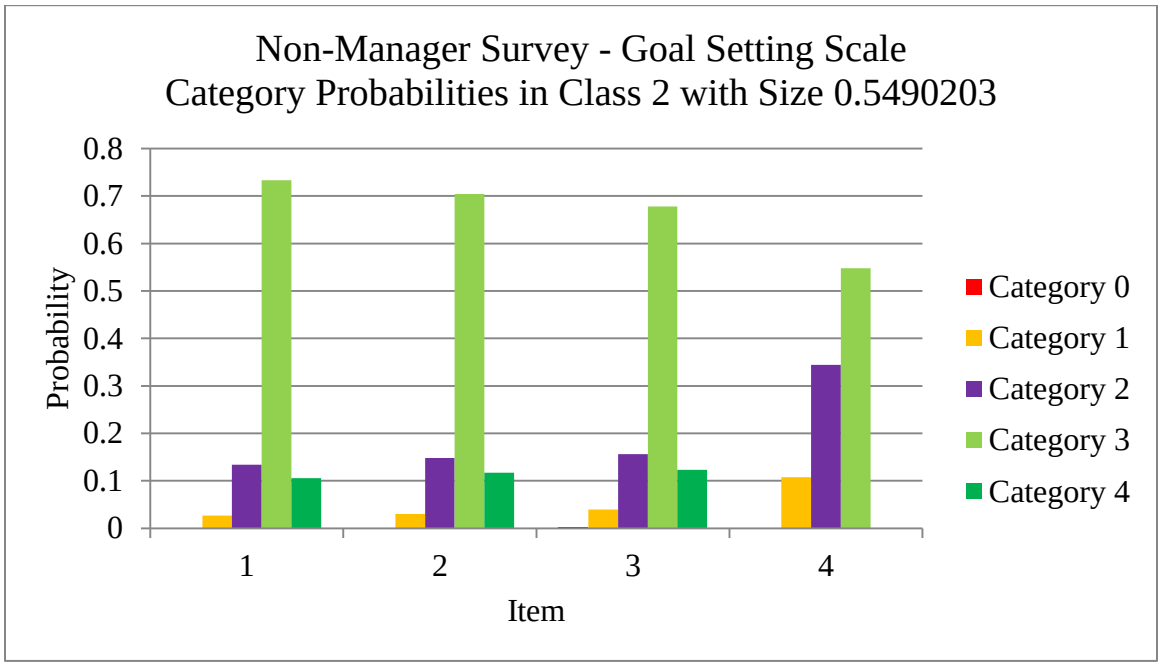

Figure 8: Category Probability Histogram for Class 2 of the Goal Setting Scale (Non-Manager)

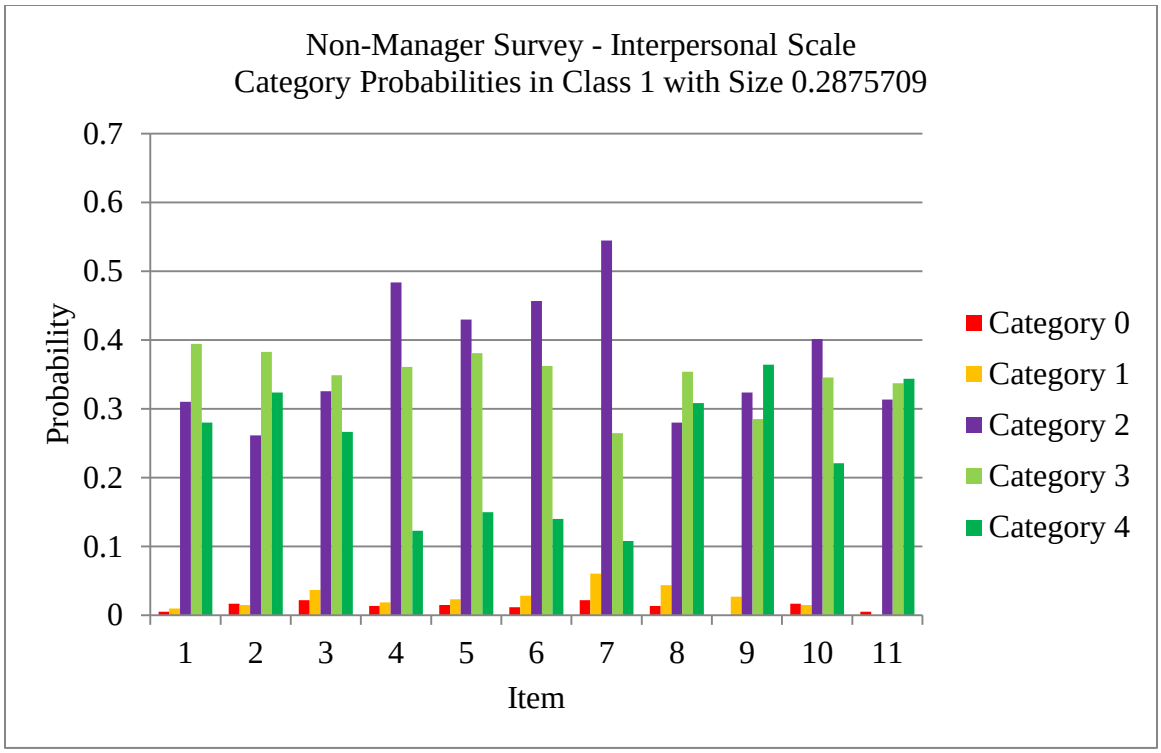

Figure 9: Category Probability Histogram for Class 1 of the Interpersonal Scale (Non-Manager)

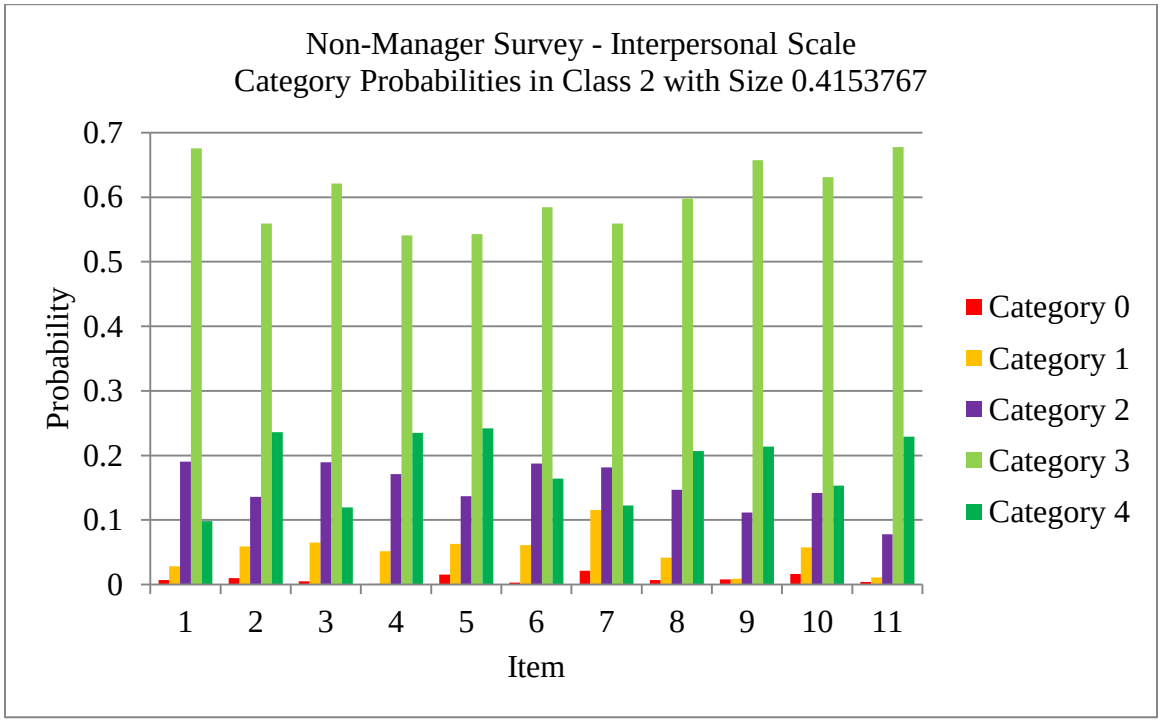

Figure 10: Category Probability Histogram for Class 2 of the Interpersonal Scale (Non-Manager)

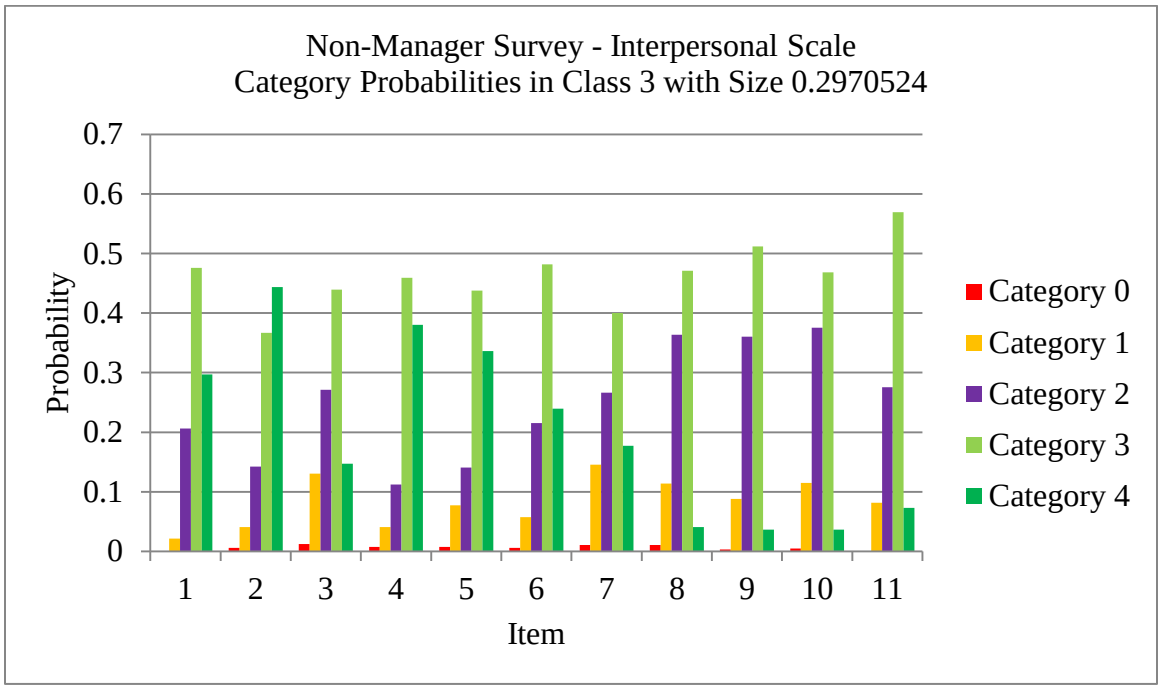

Figure 11: Category Probability Histogram for Class 3 of the Interpersonal Scale (Non-Manager)

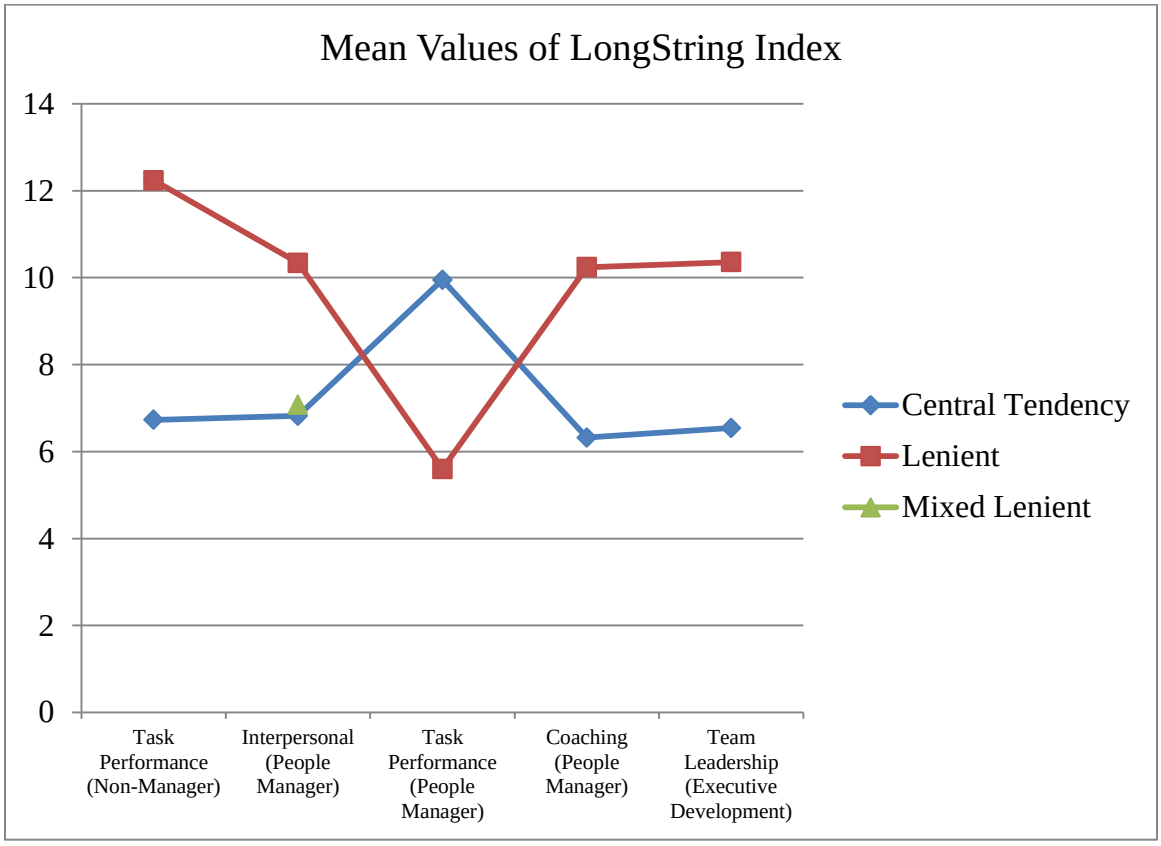

Figure 12: Mean Values of the LongString Index

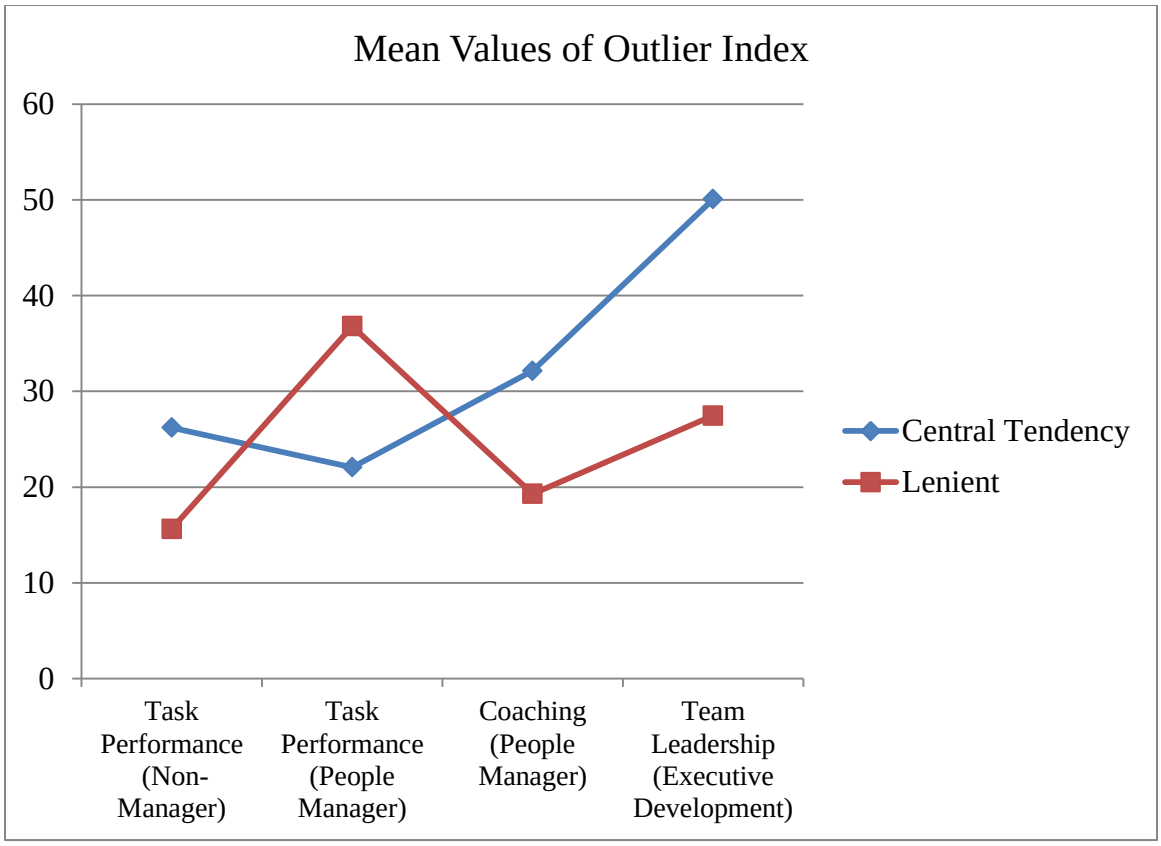

Figure 13: Mean Values of the Outlier Index