Assessing the Differential Functioning of Items and Tests of a Polytomous Employee Attitude Survey

By
Carl Swander

Thesis Submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Psychology

Robert J. Harvey, Chair
John Donovan
Roseanne Foti

22 March 1999
Blacksburg, Virginia

Keywords: Differential Item Functioning, DIF, DFIT, IRT, Attitude Assessment

Copyright 1999, Carl Swander

Assessing the Differential Functioning of Items and Tests of a Polytomous Employee Attitude Survey

Carl Swander

Abstract

Dimensions of a polytomous employee attitude survey were examined for the presence of differential item functioning (DIF) and differential test functioning (DTF) utilizing Raju, van der Linden, & Fleer's (1995) differential functioning of items and tests (DFIT) framework. Comparisons were made between managers and non-managers on the 'Management' dimension and between medical staff and nurse staff employees on both the 'Management' and 'Quality of Care and Service' dimensions. 2 out of 21 items from the manager/non-manager comparison were found to have significant DIF, supporting the generalizability of Lynch, Barnes-Farell, and Kulikowich (1998). No items from the medical staff/nurse staff comparisons were found to have DIF. The DTF results indicated that in two out of the three comparisons 1 item could be removed to create dimensions free from DTF. Based on the current findings implications and future research are discussed.

ACKNOWLEDGMENTS

I would like to thank my thesis committee John Donovan and Roseanne Foti for taking the time to help with this project. Special thanks to the committee chairman, Robert J. Harvey, for his initial instruction and guidance throughout this project. Thanks to the all the faculty at Virginia Tech for the instruction and guidance that has made all of this possible.

Thanks to Nambury S. Raju for taking the time to help me through multiple stages of this project. Without his help I would still be attempting to analyze the data with a program that didn't work. Also, I would like to thank William Collins for sending me his dissertation and offering his help. Thanks to Oscar Spurlin and Candy Penix for helping me to obtain the data for this study. Without the data this study would have been impossible.

I would also like to thank my family. My parents have been an incredible help. Not only have they helped me to get where I am but they continue to help me get where I want to be. Their suggestions and contributions throughout this project have been invaluable. Finally, I would like to thank Megan Hamaker for giving me a tremendous amount support throughout the entire process.

TABLE OF CONTENTS

Assessing the Differential Functioning of Items and Tests of a Polytomous Employee Attitude Survey

For over 70 years employee attitude surveys have been used in organizations to evaluate how employees feel about a variety of organizational variables. During the onset of their popularity they attracted a considerable amount of attention. Researchers evaluated many measurement issues, especially those pertaining to the construction of the surveys. Until recently, however, attitude surveys had not been evaluated using advanced statistical methods. The application of newly advanced measurement procedures can improve the usefulness of employee attitude surveys (Collins, 1996; Lynch, Barnes-Farrell and Kulikowich, 1998a).

Typically, survey results are aggregated across all levels of the organization to obtain an overall rating. These ratings are then used to evaluate the overall level of satisfaction that the employees have with different organizational issues. Aggregated ratings may provide the best overall picture of how employees feel about the organization, but there are often important distinctions between groups as well. Survey results often show that there are mean differences between functionally different groups within the organization. For this reason, researchers and practitioners, such as Scarpello and Vandenberg (1990) and Scheimann (1990), recommend analyzing results within specific groups.

Lynch et al. (1998a) provide evidence indicating that these mean differences may not be true population differences of the attitude but that the survey items are functioning differently for different populations within an organization. The differences between certain groups may lead to inappropriate aggregation across these groups or, even worse,

wrong interpretation of the results for certain groups. Lynch et al. (1998a) were first to explore these differences using a differential item functioning (DIF) technique, a recently developed procedure for detecting bias in testing.

DIF techniques are employed to determine if measures are truly measuring the same construct across different groups of respondents. DIF procedures can be carried out in a variety of ways but the assumption when using these procedures is that individuals with the same level of the construct of interest should answer an item in the same manner regardless of group membership. If a significant difference occurs, it may be interpreted as an indication that a different construct is being measured for the two groups.

Lynch et al. (1998a) found significant DIF between non-manager and manager groups for two dimensions of an attitude survey. Different conceptualizations or frame of reference in interpretation of the items may explain all or part of the mean differences exhibited between these groups. Managers and non-managers are exposed to different aspects of the organization. Also, managers may view various sections of a survey as self-rating (e.g., a dimension that deals primarily with attitudes about management). These results indicate that a difference between groups is not purely a function of different levels of attitude but that the items differentially measure the attitude between groups. While it has been established that manager/non-manager groups exhibit DIF, it is important to evaluate item functioning with regard to other organizational subgroups.

The current study employs the results of an employee attitude survey given in a large medical service organization to explore DIF between functionally different groups. The survey has been implemented for several years and results are usually reported at the

aggregate level and also broken down by division and other subgroups within the organization.

The purpose of the present study is threefold. First, it will attempt to assess the generalizability of findings of Lynch et al. (1998a) by evaluating DIF for manager and non-manager groups utilizing a different sample and attitude scale. Second, this study will further explore the evidence for differential item functioning between groups by examining medical staff and nurse staff employees in the health care industry. This division of labor is one of the major distinctions made within this industry. While these positions are not equivalent to that of the manager/non-manager relationship, they both perform different roles and subsequently have different functions and relationships within the organization. Third, DIF will be explored at both the item and dimension level using Raju, van der Linden and Fleer's (1995) differential functioning of items and tests (DFIT) framework. Differential test functioning (DTF) is the term used to indicate bias at the test or dimension level. Collins (1996) has demonstrated the usefulness of the DFIT framework for polytomous employee attitude surveys. The unique contribution of the DFIT framework is the added ability to detect DTF, which allows for the identification of items to be removed to create a measure free of DTF. While the item level analyses allow for the examination of specific items, the test level analyses are used to explore the additive contribution of DIF to determine which items should be removed from the test to improve interpretability.

Findings from the current study should have implications for employee survey construction, analysis, and interpretation. Further confirmation of clear patterns of survey DIF among functionally different groups would suggest that the particular frame

of reference of employee groups should be taken into account in the initial construction of surveys. Care should be taken either to try to minimize this impact or to explicitly take advantage of this factor. For example, perhaps it is not always a good idea to construct a survey using a limited set of items that are intended to apply to all employees.

Definition of Attitude

In 1935, Allport defined an attitude as "a mental and neural state of readiness, organized through experience, exerting a directive or dynamic influence upon the individual's response to all objects and situations to which it is related" (p. 810). Although there is not a complete union of the definition of attitude, the notion of an attitude as a *state of readiness* is shared by the majority of the definitions (Campbell, 1950; Cattell, Maxwell, Light & Unger, 1949; McNemar, 1946; Sherif & Cantril, 1945; Stagner, 1950; Viteles, 1953). Recent definitions of attitude have become more general, referring to attitude as "general evaluations"(Petty, 1995, p. 196) or "dispositions" (Greenwald & Banaji, 1995, p. 7) towards stimuli. Current work in the field of Industrial/Organizational psychology tends to focus on the attitude of job satisfaction, which is commonly defined as work experiences that cause positive affect towards one's job. Consistent with attitude assessment surveys, the investigator of job satisfaction is interested in how different components are related to different criterion variables (Brief, 1998). The *state of readiness* component is not included in these modern definitions, but the importance of this aspect is implied. These attitudes are commonly evaluated in relation to behaviors and how they can help to predict or explain certain behaviors (Brief, 1998; Campbell, 1950; McGuire, 1976). Viteles (1953) explains that "psychologists have

used the measurement of attitudes as a device both for making inferences concerning motives and as predictors of behaviors" (p74).

The concept of behavior prediction is important to the measurement of attitudes. The theory driving attitude measurement is that attitudes can be used to predict behavior and by changing attitudes, behaviors can also be changed (Viteles, 1953). This theory guides many current issues within psychology, such as cognitive dissonance theory (Festinger, 1957), attitudes and attitude change theories (Petty, 1995), implicit social cognition theory (Greenwald & Banaji, 1995) and public opinion poll research (Katz, 1941; Katz, 1942; Kornhauser, 1946). Moreover, recent work indicates that job satisfaction is related to organizational citizenship behaviors or contextual performance (Brief, 1998). Although the evidence for the accuracy of predicting behavior from attitudes has not been particularly strong, this fundamental relationship was the driving force behind the use of employee attitude assessments.

Attitude Assessment Background

In the social sciences, surveys are a commonly employed research method used to assess people's behaviors and/or attitudes. More than one hundred thousand surveys are utilized yearly (Rosenfeld, Edward & Thomas, 1993). Surveys have also become an integral part of organizational management. Attitude assessment surveys are the most frequently conducted type of survey implemented in organizations.

The use of attitude surveys within organizations began in the 1920's and has continued to flourish in the present. Narrow management focus on the direct behaviors of their employees impeded the use of attitude assessment until after World War I. As a result of uncontrollable labor disputes organizations were persuaded to take a different

approach to understanding and dealing with employees (Jacoby, 1988). The measurement of attitudes helped to identify ways that organizations could relieve negative situations for employees before behavioral responses occurred. It was also demonstrated that managers did not have an accurate view of employee attitudes. Attitude assessment allowed managers to more accurately measure attitudes, which helped them identify and deal with problems before they occurred (Jacoby, 1988).

Attitude surveys were heavily promoted after World War II when consultants claimed the utility of such surveys in easing union tensions. These surveys allowed employees to divulge their true feelings about the organization. This enabled the organization to identify and deal with problems before unions were necessary and enabled the organizations to "safely reject any union demands which employees did not support or did not care about" (Jacoby, 1988, p. 77). Although unions became less involved during the 1950's, survey use remained popular. Canter (1948) surveyed 103 Industrial psychologists and found that 36 percent of the psychologists surveyed were involved with organizational attitude surveys. In addition, a study conducted in 1951 by the "Conference Board" reported that 223 companies had used employee attitude surveys (Viteles, 1953).

Not only were attitude assessments effective in relieving union tensions, they also helped employers to understand their employees better. One of the primary purposes of the attitude assessment is to open communication channels between employee and upper management. Communication between management and employees is imperative to solving human problems in an organization (Irwin, 1945; Marks, 1982; McMurry, 1932; William, 1979).

The popularity of attitude assessment has continued to grow. In 1988, Gallup reported that within the previous ten year period, 70% of organizations had conducted an employee attitude survey. Also, 69% of the organizations surveyed agreed that the survey was helpful and that they wanted to conduct another survey in the future. These findings suggest that organizations are convinced that these surveys are useful tools that they will continue to use.

Surveys have recently become recognized as effective tools for increasing organizational effectiveness. This is concurrent with early research on attitude assessment by Putnam (1930) who found that attitudes of employees were more related to the effectiveness of the employee than any other variable. Organizations involved with employee attitude surveys are almost invariably trying to find some general attitude that the employee holds about various aspects of the organization (Viteles, 1953). These general attitudes can be used to identify problems in many organizational areas such as culture, planning and assessment, communication, feedback, differences between groups, productivity (Scheimann, 1990), effectiveness of personnel, training needs (Irwin, 1945; Rothwell, 1983), employer-employee relations (Cole, 1940) (supervision; Scarpello & Vandenberg, 1990), working conditions and organizational policies and procedures (Scarpello & Vandenberg, 1990). Identifying problem areas is essential to properly addressing and relieving the causes of worker dissatisfaction (McMurry, 1932). Attitudes can also help predict potentially detrimental factors such as employee counterproductivity (Moretti, 1986), turnover (Mobely, Griffeth, Hand & Meglino, 1979; Rothwell, 1983; Scheimann, 1990; Steers and Mowday, 1981; William, 1979), lateness and absence (Koslowsky, Sagie, Kruasz & Singer, 1997) and turnover intentions (George & Jones,

1996). The goal of attitude assessment within organizations is to identify problem areas and resolve the issues, thus reducing negative attitudes and their negative behavioral consequences.

Attitude assessment surveys were first designed to ensure that managers had accurate perceptions of employee attitudes about the organizations. Knowledge of these perceptions is used to help improve the organization based on what employees actually want. This has proved successful in derailing many of the negative effects of unions. The knowledge of employee attitudes can be extremely valuable in spotting problem areas within the organization. Not only do employee attitude surveys provide information about attitudes, but they also give employees a chance to voice their opinions and ideas to management. This recognition can serve as a motivational force for employees.

Measurement of Employee Attitudes

Despite the wide use and importance of attitude surveys, some methodological problems are of concern. Given the great advancements attitude assessment has made for organizations, the focus of methodological research has been fairly limited (Rosenfeld et al., 1993). In 1930, Kornhauser identified the importance of worker's attitudes and feelings, but also indicated the need to improve attitude assessment questionnaires. Subsequently, researchers have evaluated certain methodological issues faced when constructing employee attitude surveys.

Many articles have focused on issues related to the construction of attitude assessment scales. Researchers have examined issues such as reliability, reproducibility (Campbell, 1950; Edwards and Kilpatrick, 1948; Edwards and Thomas, 1993; Uhrbrock,

1934), length issues, statement form, scoring of medium or neutral responses, sorting methods, arrangement of items, methods of scoring (Day, 1940), readability level, rating scales, length (Edwards and Thomas, 1993), validity (Cattell et al., 1949), measurement error (Dutka and Frankel, 1993), and comparison of internal versus external norms (William, 1979).

It is surprising that more advances in measurement have not been applied to attitude assessment (e.g., DIF, IRT). Thissen, Steinberg, Pyszczynski and Greenberg (1983) noted this limited use of statistical techniques being applied to the construction of attitude scales given their frequent use. Since that article, a few other authors have shown that these techniques can be properly used to evaluate attitude surveys (Collins, 1996; Koch, 1983; Lynch et al., 1998a; Lynch et al, 1998b; Thissen et al., 1983). However, Collins (1996) and Lynch et al. (1998a) were the only researchers to apply these techniques to further understanding of employee attitude assessments. In their studies, they examined the degree to which attitude scales differentially measured attitudes for different subgroups of the population. Collins (1996) found significant differences for ethnic and gender divided subgroups for some items. Lynch et al. (1998a) used subgroup differences that were created by the organization. Specifically, they examined the difference between managers and non-managers on the 'Management' dimension and 'Customer Orientation' dimension and found that attitude surveys differentially measure attitudes for these groups. Their interest in examining subgroup difference within the organizational level was based on consistent mean differences between managers and non-managers level of attitude about the organization.

Mean Differences

The most common method of reporting results of an attitude assessment survey is to aggregate scores across all employees to determine the overall attitudes within the organization. The data is then broken into department level analyses or by demographic variables such as gender, length of service, ethnic, and age (Lynch et al., 1998a; Schiemann, 1991; William, 1979). Based on previous research and suggestions by many others, one would conclude that this might not be the most appropriate method of extracting information from the data (Bernstein, 1981; Cole, 1940; Korhauser, 1947; Scarpello & Vandenberg; 1990; Schiemann, 1990; Stagner, 1950; Uhrbrock, 1934; William, 1979).

Uhrbrock (1934) illustrates a perfect example of these differences. He used an attitude scale to assess employees' opinions and found that in six out of six factories surveyed, foremen had a mean rating higher than that of the other employees. Also, Uhrbrock found significant differences between mean scores for three classifications of employees: Factory workers, clerks, and foremen. Significant attitude differences were also noted based on the amount of time an employee had been with the organization. Uhrbrock was not certain as to why the differences were present. He suggested that the foremen may have been promoted because of their positive attitudes towards the organization. The clerks, who were also higher in status and pay than the factory workers, gave ratings more closely related to those of the managers. Uhrbrock suggested that the clerks might have had more knowledge about plans and procedures of the

organization. He also suggested that the differences in mean rating could be due to the fact that the more a group was the paid, the higher the group rated the organization.

Cole (1940) also demonstrated that mean differences between members of job categories would be lost if aggregated across the whole sample. He found that group differences in attitudes about the job emerged between variables such as length of service (the longer, the more satisfied), seeing opportunity for advancement, personal relationships with employers (those who talked to their employers were generally more happy with their jobs), qualification for other work (workers more qualified for other jobs were more satisfied with their jobs), and satisfaction with pay.

Thus, it appears that the analysis of attitude survey data must be done within specific groups, not at the organizational level (Stagner, 1950). Scarpello and Vandenberg (1990) and Schiemann (1990) recommended classifying individuals into specific groups for the examination of the results. Bernstein (1981) also identified the need to determine how management levels and demographics of employees affect their attitudes. William (1979) reported that morale is highest for managers because they have more authority and responsibility to make things happen. Problem areas may be different for employees at different levels of the organization. To the extent employee realities are different, their ratings are likely indicative of different problems. Survey data will not identify these different problems if results are aggregated across everyone (Bernstein, 1981). Kornhauser (1947) also argued that it is important to identify what each of the different groups desire in their relations with the organization. Attitudes can help to determine what expectations employees have and how this will impact the effectiveness of an organization.

It is quite clear that there are often mean differences between employees who hold unique positions within an organization. This phenomenon is commonly dealt with by analyzing results by separate groups. Although this approach appears to be a reasonable solution, it is only valid if the attitudes are truly being measured in the same way. The problem with evaluating mean differences is that it lacks the theoretical reasoning of why these differences are occurring, in effect assuming that the only impact is from true attitudinal differences. The assumption that these questions are measuring the same attitude equivalently is a challenging assumption to make and may be an incorrect statement in some cases (e.g., Lynch et al., 1998a). Although relatively few studies have examined mean differences, ratings of certain groups, such as management groups, are consistently higher than that of the non-management group.

In a comparison of managers and non-managers, Lynch et al. (1998a) concluded that different perceptions and interpretation of the items may explain the mean differences exhibited between these groups. They proposed that management may view various sections of a survey as a self-rating (e.g., Management dimension). Also, managers experience different aspects of the organization and are exposed to more information than non-managers. Additionally, Lynch et al demonstrated that differential functioning can play a role even when mean differences are not present. This was demonstrated by Lynch et al.'s study of the customer orientation scale. Similar mean scores could indicate the same level of attitude, but the scores should be interpreted in relation to other factors affecting the meaning of the rating for that group.

Self-Ratings

If managers interpret certain sections of an attitude assessment as self-ratings the mean differences exhibited are consistent with the performance appraisal literature. Differences between types of raters are often examined within the performance appraisal literature. Obtaining multiple raters has been used to increase the reliability of performance appraisals because it allows for multiple people to observe different job behaviors (Houston, Raymond & Svec 1991). This has led to research on the differences between raters when rating the same person. Of particular concern in this study is the difference between self and other ratings of performance. The literature suggests that self-ratings are consistently higher, more lenient than that those of supervisors and peers (Farh, Dobbins and Cheng 1991; Harris & Scheobroeck, 1988; Yu & Murphy, 1993).

Harris and Schaubroeck (1988) conducted meta-analyses to determine the relationship between supervisory, self, and peer ratings. In their analysis, peer ratings correlated highly with supervisory ratings (ρ=.62), while self appraisals correlated only moderately with supervisory appraisals (ρ=.35) and peer appraisals (ρ=.36). Self-ratings were the most lenient, on average half a standard deviation higher than supervisory ratings and one-quarter standard deviation higher than peer ratings.

Farh et al. (1991) and Yu et al. (1993) also found that self-ratings were consistently higher than those of the supervisors. The relationship between self and supervisor ratings has also been fairly weak. Research on self-ratings has found them to be consistently higher than those of the supervisor and peers and they also have consistently low to moderate correlations with each other.

Although leniency errors introduced during a rating of the self may account for the mean differences, it may also be the case that mean differences indicate a unique conceptualization of the questions not just a favorable distortion of responses. This would necessitate the differential interpretation of the results. For example, results of the 'Management' dimension for managers would not be useful in identifying attitudes concerning higher level management or senior management, while this may be the proper assessment of the non-managers' results.

Organizational level

Lynch et al. (1998a) not only found that survey items function differently between managers and non-managers; specifically, the results indicated that the items were differentially discriminating for the two groups. This would suggest that there may be more knowledge applied to answering an item by the management group. This knowledge base could be purely a function of what level a person is in the organization. Organizations are structured; this structure provides definitions of jobs and roles associated with them. These roles affect the relationships and knowledge employees have with regards to the organization (Illgen & Hollenbeck, 1991). The higher individuals are within the organization the more knowledge they have regarding policies and procedures and the more likely they are to be involved with the decisions made about policies and procedures.

Uhrbrock (1934) concluded from the findings of mean differences between levels of employees that the clerks might have more knowledge about plans and procedures of the organization than the workers, which may have led to the higher ratings. Mean differences are not the only indicator that attitude assessment surveys may function

differently for different groups within the organization. Managers have different relationships and roles within the organization than non-managers. These roles determine what organizational information the employee is provided. The assumption of different roles within the organization has also been applied to other theories of organizational functioning. For example, the leader-member exchange approach to leadership is based on the theory that members of an organization negotiate or develop roles (Dansereau, Graen, & Haga, 1975; Yukl, 1998). That is, members of organizations fill roles in order to accomplish the required tasks of that organization. While this theory posits how the relationship between leaders and members form, the key assumption is that employees form unique roles within the organization. These roles determine the relationship the employee has with the organization and the amount of information the employee has about the functioning of the organization. This information could lead to different conceptualizations of the survey items. The implication of different conceptualizations would be to analyze the results differently based on group membership.

The importance of the interpretation of the results has largely been ignored. Employee attitude surveys can be used for many different purposes within an organization, but for them to be of any value the results must be interpreted correctly. Results that are wrong or misleading due to improper interpretations can have harmful effects within an organization. Erroneous interpretations can lead to action plans that are not related to problem areas within the organization. Also, employees' trust in the organization is likely to decrease if surveys do not lead to helping resolve problem situations (Scarpello & Vandenberg, 1990).

It is likely that attitude survey items could unequivocally measure other functionally different groups beside the distinction of managers and non-managers. If organizational roles lead employees to know more about the organizational functioning, view questions as self-ratings, or interpret the meaning of the items differently in any manner, then measurement equivalence would not exist. The current survey was conducted in a large medical service organization in which there were more distinctions among employees than those of managers and non-managers. The other groups of interest in the current study are the medical staff and nurse staff. The medical staff employees are doctors. The dimensions that are of particular interest for these subgroups are the 'Management' and 'Quality Care of Service'. These dimensions would be likely to demonstrate the same differences found in the Lynch et al. (1998a) study of managers for similar reasons. Doctors are placed at a higher organizational level than nursing employees and are likely to view themselves as managers of the direct patient care employees. These differences will be examined in two dimensions of the attitude assessment. These dimensions include the 'Management' and the 'Quality of Care and Service' dimensions. Doctors may have more information regarding patient care initiatives within the organization and may feel more directly responsible for this dimension. The manager/non-manager distinction will also be evaluated on the 'Management' dimension to corroborate the findings of Lynch et al. (1998a).

Illustrative Examples

The following are two examples which demonstrate how attitude surveys would differentially measure an attitude between groups. First, an example of a question in the management scale is 'Innovation is rewarded at my organization.' As a line employee,

this question might be interpreted as the question intended. In essence, employees would be rating how well the organization acknowledges and rewards innovation (e.g., through financial incentives). However, a manager may interpret the question as being how effective he or she as an individual communicates appreciation for innovation to employees. This interpretation would clearly be a self-appraisal of effective communication and supervision. While this may in part lead to leniency, an error component, the actual true score of what the question is measuring may also be different. In this situation the results of the survey for the different people would not be the same. For the employee, conclusions regarding their feelings towards the entire innovation reward system may be warranted. On the other hand, the same conclusions may not be appropriate for managers. The interpretations of the results could only be applied to the individual manager, not to the entire organizational management.

A second example of DIF for attitude surveys may be illustrated using a question from the 'Quality of Care and Service' dimension and the distinction between the medical staff and the nurse staff employees. The item of interest reads: "Our organization is committed to continuously improving patients' and other customers' satisfaction with healthcare outcomes." In this situation it would be possible that the roles of the individuals would be related to how they interpret the question. The nurse staff employees might answer this question taking into account a number of variables including level of staffing, pay, employee working conditions such as hours, complaints or compliments of customers and condition of customer facilities such as waiting rooms. Medical staff could have a completely different set of criterion for answering this question. They might evaluate this question with regard to organizational recognition in

the field of research, physician prestige, administrative delay or denial of physician recommended procedures, and age and type of medical equipment available to them. If this question were answered according to these variables it would be clear that the work roles of the individual clearly influence the interpretation of the items. The interpretation and identification of solutions to these questions would have to be considered separately for each group.

These differences can be evaluated using a relatively new technique for assessing bias in testing labeled differential item functioning (DIF). The following section will address the methodological issues regarding DIF and will demonstrate how these techniques can be used to evaluate employee attitude surveys.

Differential Item Functioning

Although rigorous examination of bias in testing began in the 1960's, it is important to emphasize the current dissatisfaction with the term bias (Angoff, 1991; Hambleton, Swaminathan & Rogers, 1991). This term has two meanings when evaluating measurement instruments. First, social bias is that of differences between members of different groups because of their group membership. Second, statistical bias refers to difference in performance for a certain group. The problem with interpreting the dual meaning of bias is that the social implication of bias is one of fairness or prejudice and is labeled as a cause while statistical bias is purely a measurement issue. It is important that these two meanings, social and statistical, are not mixed. The term differential item functioning (DIF) is an attempt to solve this problem. Differential item functioning (DIF) refers to "the simple observation that an item displays different statistical properties in different group settings." (Angoff, 1991, p. 4) The cause of DIF is

unknown and may or may not be able to be determined by theoretical/social means. This distinction is very important in the current paper. The DIF exhibited in an attitude assessment survey would not be one of social bias or prejudice but rather different conceptualization or frame of reference used to answer survey items.

The term impact should also be distinguished from DIF. As mentioned previously in the paper, mean differences do not necessarily indicate DIF because this could be a true population difference. The term impact is used to indicate these real population differences. Impact often occurs but this does not mean a test is biased. True performance on a given test can be truly different between groups. DIF refers to the statistical properties of the test that are different for different groups at the *same* level of the variable being measured (Millsap & Everson, 1993).

DIF studies have been commonly used to study differences in test scores between subgroups in a population. Typically, the minority group is referred to as the reference group and the majority as the focal group. With the increased popularity of DIF, a growing body of literature on the techniques used for assessing DIF has become available. The majority of the literature concerned with the detection of DIF has focused on evaluation of DIF for dichotomously scored items. This has led to the creation of many techniques for the identification of DIF for dichotomously scored items and tests. However, many tests are not dichotomously scored. Recent attempts have been made to create procedures that can be used to identify DIF in polytomous scales.

Potenza and Dorans (1995) and Millsap and Everson (1993) have written fairly extensive reviews of the models of DIF that are applicable to polytomous items. They identified two broad categories of DIF procedures. These procedures can be broken into

either observed score approaches or latent variable approaches. Observed score and latent variable approaches are different with respect to the criterion that is chosen to match subgroups for evaluation of DIF.

Observed score methods have the same assumption regarding the null hypothesis of DIF. The assumption is that if each group has the same proportion of getting an item correct and has the same overall score on a homogeneous sample of items that includes the item being studied, then the item is unbiased (Potenza & Dorans 1995). Hence, the observed score is used as a proxy for the underlying variable, which creates the need to rely on the observed score when matching people at certain levels of the variable being measured. Observed score models that have been adapted for polytomous items include the Mantel-Haenszel (MH) procedure, Standardization (STND) procedure and the logistic regression procedure (LRDIF). Although these procedures are commonly used, overall test score used as a matching criterion can cause a few problems (Dorans & Holland, 1993). First, when using the observed score as the criterion it must be assumed that that what is being measured is reliable. While many tests have adequate reliability, perfect reliability is relatively impossible. This would require the assumption that the errors of measurement will not affect DIF detection. Second, there has been serious concern as to whether it is appropriate to include or exclude the items that demonstrate DIF when matching the observed scores. While some researchers remove items that have large DIF (Dorans and Holland, 1993), others suggest that exclusion of the item being studied can make assessment of DIF problematic (Potenza & Dorans, 1995). Finally, Dorans and Holland (1993) explain that these techniques need to be adapted for the use in identifying DIF at a larger level of analysis (e.g., test) than just the item level.

Latent variable DIF procedures are based on the idea that test scores are a measure of both a reliable portion and an unreliable portion (Potenza & Dorans, 1995). That is, a score can be broken down into true score and error components. Within this framework it is possible to derive estimates of the latent trait to use as a matching variable. Drasgow and Hulin (1991) explained that models that compare probabilities of getting an item right between two groups at the same level of ability are often referred to as "theoretically preferred" models. Both the simultaneous item bias test (SIBTEST) and item response theory (IRT) based models of DIF detection are latent variable models.

The SIBTEST model is a nonparametric model that is based on multidimensional IRT, but also relies on classical test theory assumptions in obtaining a matching variable. SIBTEST has recently been adapted for use with polytomous items. This model has been developed for both test and item level DIF analysis but empirical tests have been fairly limited. Sijtsma (1998) expressed concern about the lack of evaluation of nonparametric polytomous models that led him to omit these models from his study. IRT methods of detecting DIF are also preferred to other methods because the models hold the property of sample independence (invariance), this helps to clarify DIF from true population differences (Park & Lautenschlager, 1990).

For an understanding of detecting DIF using IRT methods it is necessary to explore the fundamental relationships proposed in IRT. IRT is a fairly new technique that gives item and test level statistics that are sample invariant within a linear transformation. Item response theory identifies an item characteristic curve (ICC), a monotonically increasing function of the relationship between the latent variable of interest and the probability of success, for each item. The latent variable being measured

is referred to as Theta (θ). The ICC's are derived from item parameters, which are specified according to the IRT model chosen. A test characteristic curve (TCC) is derived from adding all of the ICC's together, which can be used to evaluate the entire test.

Before using an IRT model, a few assumptions must be evaluated to determine if IRT is the appropriate technique and if so, which model is appropriate for the evaluation of the test. The assumptions that apply to the majority of the models include unidimensionality and local independence.

When using IRT models it is assumed that a single trait or ability is underlying the test in question (Hambleton, 1989). Unidimensionality is evaluated as the degree to which items are statistically dependent. This dependence is defined as a single trait or ability (Crocker & Aligna, 1986). When the assumption of unidimensionality is met, the item covariances in a variance/covariance matrix should be relatively close to zero when θ is partialled out. Essentially, unidimensional refers to a factor structure consisting of one major factor. Judgment of dimensionality should be based on statistical models, such as factor analysis along with theoretical evaluations of the content of the measurement instrument (Camilli, Wang & Fesq, 1995). The assumption of unidimensionality is also a critical assumption made when using most techniques for assessing DIF (Potenza & Dorans, 1995). Factor analysis is a powerful technique for assessing the unidimensionality of a scale.

When using IRT it is also assumed that responses to items are independent of one another if the ability that is influencing performance is held constant (i.e., local independence). That is, if ability is partialed out from the items then the items will not be

correlated (Hambleton, Swaminathan & Rogers, 1991). A subject's answer on one item should not affect that subject's answer on another item. So, the only factors that influence performance are the item characteristics and the person's ability. The assumption of local independence is met when the off diagonals of the variance/covariance matrix are close to zero after partialling out θ (Hambleton, 1989).

Essentially, local independence and unidimensionality are relatively similar and if the assumption of unidimensionality is met then local independence can be assumed. However, they are not completely equivalent because local independence can be met without having a unidimensional test. In this case, local independence can be achieved by accounting for all the abilities that are supposed to be measured and then evaluating the variance/covariance matrix.

After the basic assumptions are met, an IRT model must be chosen. Finding the appropriate IRT model requires that assumptions be made about item functioning. These assumptions will help to determine the number of parameters that are needed in the model (Harvey & Thomas, 1996). The parameters are then used to derive the item characteristic curve (ICC). The one-, two- and three-parameter logistic models are the most common IRT models (Hambleton et al., 1991).

The one-parameter logistic or Rasch IRT model uses only the difficulty (b) parameter to distinguish between the items. That is, an item will only be expected to be different in difficulty from other items. Therefore, it must be assumed that the other two parameters, the discrimination (a) and pseudo-guessing (c) parameters, will be constants for all items. The ICC for the one-parameter model is determined by the sum of the probabilities of choosing an answer across all levels of θ, which can be expressed as

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

In this model b_i is the difficulty parameter and e is a constant (2.718). This equation gives the probability that a person with a certain level of θ will answer the item correctly or incorrectly. A probability function is defined for all items of the scale. The one-parameter model can rarely be considered 'adequately' fit because items usually differentially discriminate (Hambleton et al., 1991).

The same model represents the two-parameter logistic model but with discrimination (a) parameter included. The equation for the ICC is written as

$$P_i(\theta) = \frac{e^{D_{a_i}(\theta - b_i)}}{1 + e^{D_{a_i}(\theta - b_i)}}. \quad (2)$$

The addition of D_a indicates that the discrimination parameter has been included in the equation and that it has been scaled to fit a cumulative normal distribution or normal ogive function (Hambleton et al., 1991). The addition of the a parameter helps to distinguish items based on how well they identify θ. Even when using the a and b parameters it still must be assumed that the asymptote of the ICC will always reach zero. If the ICC was to always reach zero then a person with a given level of θ, defined by the ICC, would have a zero probability of getting the answer right. This immediately causes problems when using any kind of test where there is a possibility of guessing the correct answer (Hambleton, 1989).

The three-parameter logistic model adds the pseudo-guessing or c parameter. The possibilities of guessing make it possible to have people with very low levels of θ get the

question right, even when they would have a very low probability of getting it right given their level of θ (Harvey & Thomas, 1996). The three-parameter ICC is defined by

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{D_{a_i}(\theta - b_i)}}{1 + e^{D_{a_i}(\theta - b_i)}}. \quad (3)$$

Although these three models are designated for dichotomously scored items, many models that are similar to these have been developed to fit polytomous items. Samejima (1969) presented a polytomous two-parameter model. Samejima's graded response model is based on an assumption that the scale has ordered response categories, such as a likert scale. The graded response model is relatively similar to the two-parameter logistic model. The equation is expressed as

$$P^*_{x_i}(\theta) = \frac{e^{D_{a_i}(\theta - b_{x_i})}}{1 + e^{D_{a_i}(\theta - b_{x_i})}}. \quad (4)$$

This model is different from that of the logistic dichotomous model because it is the probability of responding to one of the categories or higher. Therefore, the probability of getting a certain score is computed by taking the difference of one probability level from the probability of responding to the next category or higher. This can also be expressed as

$$P_{x_i}(\theta) = P^*_{x_i}(\theta) - P^*_{x_i+1}(\theta) \quad (5)$$

According to the graded response model, if there are x_i response categories and these categories are ordered from zero to m_i, then there are $(m_i + 1)$ response categories, although there are only m_i response functions derived for each item. There is no need to calculate the final response function because if any other response category is chosen the

final response category must be equal to zero. Each response function is analogous to the dichotomous IRT model. Equation 4 identifies the difficulty parameter for one response category. This indicates that there are m_i difficulty parameters for each item. The number of parameters estimated in the graded response model becomes very large when there are more than a few items (Hambleton et al., 1991).

Collins (1997), Koch (1983), Lynch et al. (1998a), Lynch et al. (1998b) and Thissen et al. (1983) have demonstrated that IRT models are appropriate for polytomous attitudinal scales. Koch (1983) was the first to fit Samejima's (1969) graded response model to attitudinal data. The results of his study indicated that the graded response model was appropriate for use in analyzing such data. Furthermore, it was evident that the two-parameter graded response model was a better model than the typical Rasch IRT model. This model allowed for greater precision in the measurement of the attitude. Collins (1996), Lynch et al. (1998a) and Lynch et al. (1998b) supported Koch's results.

Once the appropriate model has been defined, a method for estimating the item parameters and θ must be chosen. First, when using a sample of data it is assumed that the model will not exactly fit the data. This requires the search for a 'best fitting' curve for the data. Although the least squares procedure is commonly used in regression, it is not appropriate for the use in IRT because these are nonlinear models (Hambleton et al., 1991). Nonlinear regression could be used to identify the parameters if θ was known, but in most cases θ must also be estimated. Maximum likelihood and Bayesian estimation procedures are the most commonly used methods for estimating both θ and the item parameters (Drasgow & Hulin, 1991).

Maximum likelihood procedure attempts to estimate θ and the item parameters that maximize the likelihood function. These variables are treated as unknowns when estimated. The joint maximum likelihood estimation procedure is a variation of maximum likelihood procedure in which the parameters and θ are simultaneously estimated through a series of iterations. Problems occur when applying these techniques to short scales or small samples. When samples and test lengths are small using an estimated θ will produce poorly estimated item parameters. The marginal maximum likelihood method is preferred to these other methods because it helps to reduce the impact of scale length and sample size through the assumption that the distribution of θ is normal. This is similar to Bayesian estimation procedures which also apply a known distribution to θ.

After the estimation of the parameters, the ICC or item parameters for the reference and focal group are compared. As mentioned above, when comparing people from different groups they must be matched on the variable being measured. After the two groups are matched it is possible to examine differences between the groups that are *not* related to the variable of interest. Since IRT is a latent variable approach to DIF, the estimate of the latent variable θ is used to match the subgroups.

Matching the subgroups at a level of θ for comparisons is referred to as linking. The first major choice to be made when evaluating the linking processes is the difference between one or multiple iterations. Lord (1980) first developed a single iteration or noniterative linking process. This process includes estimating the item parameters using b as the standardized parameter. Thus, b has a mean of zero and a standard deviation of one. The next step is to standardize the a parameter so that the estimation of b is

possible. Then, all items that are found to have DIF are removed and θ is estimated with the reduced set of items. The θ value is then used to re-estimate all item parameters within each group, including the biased ones. The new parameter values are then used to examine DIF. One apparent problem of this approach is that only items that do not exhibit DIF should be included in the linking procedure yet these items are used to identify the items that have DIF. The iterative approach offers a solution to this problem by estimating θ and the item parameters multiple times without the items that exhibit DIF.

Park and Lautenschlager (1990) presented an iterative approach to Lord's noniterative procedure. This process first requires the estimation of θ. Then, all the item parameters are estimated for each group. The items that demonstrate bias are removed and θ is re-estimated. The item parameters are then re-estimated for each group using that value of θ. When consecutive iterations identify the same biased items the iterations stop.

The iterative approach to linking has been shown to be more effective at detecting DIF than the noniterative approach (Park & Lautenschlager, 1990). If items that are exhibiting DIF are not excluded from the linking measurement then false positive (fp) and false negative (fn) indications of DIF are more likely to occur.

Cohen and Kim (1998) evaluated three iterative linking methods in relation to Samejima's graded response model. These procedures included characteristic curve, minimum chi-squared (χ^2) and mean and sigma methods. The characteristic curve method is used to estimate the a and b parameters while trying to minimize the difference between the TCC's of each of the subgroups. The minimum χ^2 approach to linking

includes estimates of standard errors along with the a and b parameters. The mean and sigma methods utilize the distributions of the discrimination and difficulty parameters. Cohen and Kim (1998) found that all of these procedures work well when using polytomous data. They concluded that researchers should feel comfortable in using any one of the methods. Cohen and Kim (1993) used test characteristic curve for linking process and found that it was better for small sample sizes than either the minimum chi-squared (χ^2) methods or mean and sigma methods.

Now that the reference group and the focal group are measured on the same metric, it is possible to examine DIF. The measurement of DIF using IRT is commonly assessed using either of two measures. First, it is possible to compare the item parameters between each group. Second, the area under the ICC can be measured and compared across subgroups.

Lord's (1980) χ^2 is used to compare item parameters between the focal and referent groups. This method is based on the null hypothesis that all of the parameters will be equal. This measure can be used to test the difference between the discrimination and difficulty parameters. This model has been demonstrated to be an appropriate measure of DIF for polytomous items (Cohen, Kim & Baker, 1993). This method of assessing DIF relies on the assumption that the item under question is the only item in the scale that exhibits DIF between the subgroups (Collins, 1996).

Area under the curve method for assessing DIF has also recently been advanced to polytomous items. This method requires the comparison of the area under the ICC between the focal and the referent group. The null hypothesis in this circumstance is that the areas will be equal. One problem that occurs when using this method is when DIF is

non-uniform. Non-uniform DIF is present when the ICC's for the two groups cross. In this situation there is both positive and negative DIF between the two groups, which could result in the cancellation of the difference between the area measures. This would lead to the conclusion that there is no DIF present. Unsigned area under the curve methods have been developed to combat this problem. This method also relies on the assumption that all the items, besides the one of interest, are free from DIF. Lord's χ^2 test and area under the curve methods are limited to the identification of DIF and do not lend themselves to test DIF at the test level (Collins, 1996).

Raju et al. (1995) recently developed an IRT measure of differential functioning of items and tests (DFIT) to evaluate the effects of DIF at both the item and test level. The advantages of the DFIT framework have led to a rise in its popularity in recent research (Collin, 1996; Laffite, Raju, Scott & Fasolo, 1998; Maurer, Raju, & Collins, 1998). Research has indicated that the DFIT framework is appropriately suited to evaluate polytomously scored items. Collins (1996) specifically evaluated whether the DFIT framework was suitable for employee attitude assessment data. He concluded that it was sufficiently capable of detecting both DIF and differential test functioning (DTF) when applied to Samejima's graded response model.

The DFIT framework offers a few potential advantages over the other IRT measures of DIF. DFIT offers the test or scale level measure of DIF and also two measures of DIF. One measure of DIF is unique to the DFIT framework while the other measure is similar to those previously introduced (e.g., area under the curve and Lord's χ^2 measures)(Collins, 1996).

The new measure of DIF implies an additive relationship in the identification of DTF. If this is the case, the impact of removing an item that demonstrates DIF can be evaluated by summing the DIF for each item in the scale. This measure of DIF is labeled compensatory DIF (CDIF). Noncompensatory DIF (NCDIF) does not hold an additive property and when it is evaluated the assumption is made that all other items in the scale are free from DIF (Raju et al., 1995). These relationships will be illustrated in the methodology below.

DTF allows for evaluation at the scale level. DTF is calculated by determining two expected proportion correct (EPC) scores, indicated by the sum of the ICCs, one for each group. This can be mathematically expressed by

$$DTF = \in (T_{sF} - T_{sR})^2. \quad (6)$$

In the equation above T_{sF} represents the EPC for the focal group and the T_{sR} represents the EPC for the reference group. According to this equation, the expectation ($\in$) can be summed across either the focal or reference group and if the EPC for the two groups is different from each other then DTF is occurring across all examinees (Raju et al., 1995). In the following illustration the $\in$ will be taken over the focal group as denoted by the subscript $_F$. Now, given that it is possible to substitute D_s (difference measure) for $T_{sF} - T_{sR,}$ the DTF equation takes the form of

$$DTF = \in_F D_S^2 \quad (7)$$

which can be rewritten as

$$DTF = \int_{\theta} D_S^2 f_F(\theta) d\theta \tag{8}$$

when the density function of θ ($f_F(\theta)$)is inserted for the focal group. Finally, the equation can be expressed as a function of the variance of D (σ_D^2) and the squared difference between the mean true scores of the focal and reference groups (μ_D^2) or

$$DTF = \sigma_D^2 + (\mu_{TF} - \mu_{TR})^2 = \sigma_D^2 + \mu_{D.}^2 \tag{9}$$

Because of the additive property of CDIF it is possible to derive a measure from DTF as follows:

$$DTF = \sum_{i=1}^{n} CDIF_i. \tag{10}$$

So,

$$CDIF_i = Cov(d_i, D) + \mu_{d_i}\mu_D. \tag{11}$$

This equation can be interpreted as the covariance between the differences in probabilities of items for item i(d_i) and the difference between the expected proportion correct (D) plus the mean of di (U_{d1}) multiplied by the mean of D (U_D).

Equation (11) clarifies the relationship between CDIF and DTF, in that DTF is an additive function of the CDIF for all the items. This relationship is the first to identify that items may have correlated DIF. NCDIF, on the other hand, is based on the assumption that no items other than the one in question exhibit DIF (Collins, 1996). Because of this assumption NCDIF can be simply expressed by the equation:

$$NCDIF = \sigma_{d_i}^2 + \mu_{d_i}^2. \tag{12}$$

Test of significance for the DFIT framework

Raju et al. (1995) demonstrated how chi-squared (χ^2) tests are used to determine the significance of both DTF and NCDIF. In order to carry out a test of significance it must be assumed that *D* is normally distributed with a mean and standard deviation of μ_D and σ_D, respectively (Raju et al., 1995). Performing a χ^2 test of significance requires that each examinee's score be transformed into a z score as follows:

$$z_s = \frac{\hat{D}_s - \mu_D}{\sigma_D}. \tag{13}$$

This is an important transformation because the distribution of z_s^2 is χ^2 which indicates that the sum of the squared z scores for all examinees will also have a χ^2 distribution, with N_f degrees of freedom. Assuming that DTF is equal to zero, indicating that μ_D must be zero, the χ^2 distribution can be written as follows:

$$\chi^2_{N_F} = \frac{\sum_{s=1}^{N_F} D_s^2}{\sigma_D^2}. \tag{14}$$

The definition of DTF, as presented above (equation 7), implies that the equation can be expressed as

$$\chi^2_{N_F} = \frac{N_F DFT}{\hat{\sigma}_D^2}. \tag{15}$$

Note that a sample-based estimator of variance was substituted in the equation, which changes the degrees of freedom to N_F-1. The significance of NCDIF can be evaluated the same as DTF, the only difference is that item level statistics are used. The χ^2 significance test is expressed as

$$\chi^2_{N_F} = \frac{N_F NCDIF}{\hat{\sigma}^2_{d_i}}. \quad (16)$$

If the χ^2 tests are significant then DTF and NCDIF are significantly different from zero. It is not necessary to test the significance of CDIF because if DTF is significant then CDIF must be present. Removing the items until DTF is not significant identifies the CDIF items. Based on previous empirical studies Raju et al. (1995) and Collins (1996) recommended setting a significance cutoff value for DTF and NCDIF in addition to a χ^2 of .01 because χ^2 tests can be overly sensitive to sample size. Because both methods are relatively sensitive to sample sizes, Raju et al. (1995) recommended that a significance cutoff value be set at greater than .006 for dichotomously scored items. This criterion value is close to a .01 level of significance in that only 1% of the items will falsely be labeled as having DIF. Collins (1996) and Laffitte et al. (1998) determined that a .016 value of NCDIF and a significance level of .01 for the χ^2 test were sufficient for the detection of NCDIF. Furthermore, Collins used a significance cutoff value of .024 and a significance level of .01 for the χ^2 test to determine DTF. This value was derived by multiplying the number of possible responses minus one by .006. Further analysis of cutoff scores has led Raju (personal communication, February 22, 1999) to suggest even more stringent cutoff levels for NCDIF and DTF for polytomous data. It is

currently recommended that a cutoff score of .096 be set for both NCDIF and CDIF. The rationale for these cutoff scores is placed in the utility of the DIF finding. Because the true score of a response in the polytomous case lies between 1 and k (number of response categories), it is critical that the value for DIF reflect the scale of the items. So, the .096 value is identified by multiplying the square of the number of response categories minus one by the .006 cutoff value used for the dichotomous case. The cutoff of .096 indicates that, on a five point scale, there is an absolute difference of .310 between the focal and reference groups (N. S. Raju, personal communication, February 22, 1999). According to Raju, a value of .310 is more practically meaningful on a five point scale.

Hypotheses

Hypothesis I - The 'Management' dimension will demonstrate significant DTF between the management and non-management groups, which will indicate significant CDIF for the items of the 'Management' dimension. Also, items in the 'Management' dimension will have significant NCDIF for the management and non-management groups.

Hypothesis II - The 'Management' and 'Quality of Care and Service' dimensions will demonstrate significant DIF at both the test and item level for the medical staff and nursing staff.

Method

Participants

A large medical service company conducted a company wide survey in 1993 and 1996. The survey consisted of 100 items measuring attitudes about various parts of the organization and descriptive items. The survey conducted in 1993 sampled 4901 employees and the survey in 1996 sampled 4413 employees. The surveys included a question to distinguish between manager and non-manager groups. In 1996, 650 managers filled out the survey and 3578 non-managers filled out the survey. There were 185 people that failed to answer this item; this data was treated as missing data. Also, the survey had questions identifying groups by medical staff and nurse staff. There were 600 medical staff employees and 1086 nurse staff employees in 1996. There were 77 respondents that checked both medical staff and nurse staff responses and hence were not used in the analysis. Participation in this survey was purely voluntary and anonymous. Surveys were sent directly by the employee to an outside consulting firm where they were analyzed.

Instrument

An Industrial/Organizational psychology firm developed the survey for the purpose of assessment within this organization. The items were specifically designed for the organization, and were developed based on the topics the organization was interested in assessing. The 'Management' dimension of the survey consisted of 21 questions that were used in both 1993 and 1996. The categories within the 'Management' dimension included: 'Career Development, Communication, Decision Making, Performance Feedback and Receptiveness to Innovation in the Work Place'. The 'Quality of Care and Service' dimension consisted of 14 questions that

were used in both 1993 and 1996. The categories within this dimension were 'Commitment to Quality of Care and Service' and 'Quality Improvement Efforts'. The questions included in the 'Management' and 'Quality of Care and Service' dimensions are presented in Appendix A.

Comparisons

Manager/non-manager and medical staff/nurse staff comparisons will be made in this study. The manager/non-manager comparisons will be made on the 'Management' dimension. The medical staff/nurse staff comparisons will be made on the 'Management' and the 'Quality of Care and Service' dimensions. As indicated above, the majority group is referred to as the focal group and the minority group is referred to as the reference group. Based on sample sizes, the non-manager and nurse staff groups will be considered the focal groups and the manager and medical staff groups will be considered the reference groups.

Unidimensionality of Subscales

Although the attitude assessment has been broken into various sections no empirical examination of the unidimensionality has been performed on the instrument. Item factor analysis is the most informative and sensitive method for assessing the dimensionality of a test. Principal components factor analysis will be used to assess the unidimensionality of each subscale. The criterion for unidimensionality of the subscales will be evaluated in two ways. First, as recommended by Reckase (1979), the percent of variance accounted for by the first factor must be greater than 20 percent. While this is a minimal amount of variance for the identification of unidimensionality, Drasgow and Parsons (1983) demonstrated that substantial violations of unidimensionality do not warrant using a multidimensional IRT model. Second, the scree plot of eigenvalues will be examined to determine if there is a dominant first factor. When analyzing

the scree plot a dominant first factor will be clearly separated from the rest of the factors. When using factor analysis the eigenvalues are examined to see if a dominant first factor is present (Hambleton, 1989). Furthermore, these interpretations of unidimensionality are consistent with Camilli et al.'s (1995) determination that dimensionality should be approached from a functionally unidimensional viewpoint instead of the unrealistic standard of statistical unidimensionality. The principal components factor analysis will be conducted on the entire sample for 1993. Collins (1996) demonstrated that these criteria were successful in determining the unidimensionality of the attitude scale used in his study.

IRT Model, Parameter Estimation and Assessment of Fit

As indicated above, Samejima's two-parameter graded response model is appropriate for the current study. MULTILOG, which implements the marginal maximum likelihood estimation procedure, will be used to estimate θ and the item parameters. MULTILOG also provides fit indices to evaluate the model chosen. MULTILOG employs a likelihood ratio χ^2 test of the graded response model (Thissen, 1986). Mckinley and Mills (1985) demonstrated that the likelihood χ^2 ratio resulted in fewer false negatives than other fit indices. Also, it was compatible with the other fit indices in indicating correct rejections of the null. Mckinley and Mills also demonstrated that the likelihood χ^2 ratio correctly rejected the fit of a unidimensional model when applied to multidimensional data, indicating another test of unidimensionality.

Linking Procedure

The program EQUATE 2.1 (Baker, 1993) will be used to match the person and item parameters. EQUATE 2.1 employs a characteristic curve method of linking the subgroups.

Raju's DFITPU4 (1997) FORTRAN program will be used to identify the DFIT indices. This program is derived from the equations presented above. The significance of NCDIF and DTF will be determined using the criterion value of .096 accompanied by a significance level of .01 for the χ^2 test. If items are indicated to have significant DIF then the items will be removed from the sample and new EQUATE constants will be estimated. These new estimates will then be used to re-estimate the DFIT indices to determine if more items demonstrate NCDIF. This method will be repeated until no new items demonstrate NCDIF. Also, the significance of DTF will be evaluated. If the results indicate that DTF is significant then the item with the largest CDIF will be removed and DTF will be re-estimated. This procedure will be continued until DTF is no longer significant.

Results

Mean Differences

Mean differences between various organizational levels are common within attitude assessment research, as indicated previously in the text. A dimension score was computed on the 'Management' dimension and the 'Quality of Care and Service' dimension based on the average of all the items within the dimension. Descriptive statistics were computed to compare the mean rating on the dimension and comparison groups of interest (Figure 1a, 1b, and 1c; see appendix B for item means). Figure 1a, 1b, and 1c are graphical representations of the mean differences between the groups. As can be seen in these figures, the mean differences are the similar to those found in previous research on attitude assessment surveys (e.g., Uhrbrock, 1934). The current survey demonstrates the consistent finding that managers have higher mean rating than those of non-managers and that employees with higher status (e.g., medical staff) have higher mean ratings than those with lower status in the organization (e.g., nurse staff). It is interesting to note that the means of managers on the 'Management' dimension are also higher than that of the medical staff and Nurse. This finding suggests any findings of DIF from the medical staff/nurse staff comparisons are based on other factors besides the nature of superior-subordinate relationships. The mean difference between all groups is at least .2, on a five-point scale. The following results will help to determine if these differences are true attitudinal differences about the same construct or whether these differences are a byproduct of DIF.

Figure 1a

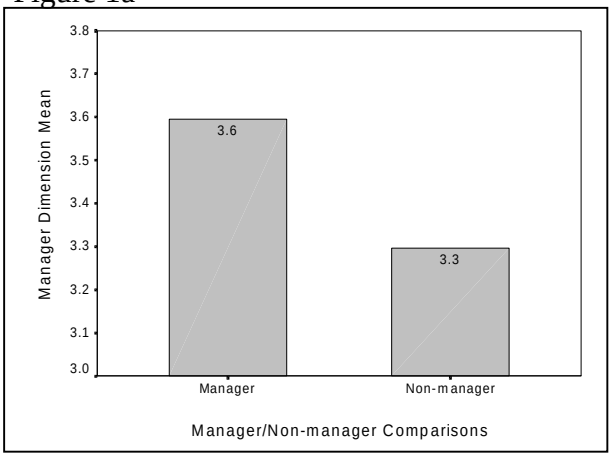

Figure 1a represents the mean differences between the managers and non-managers on the 'Management' dimension.

Figure 1b

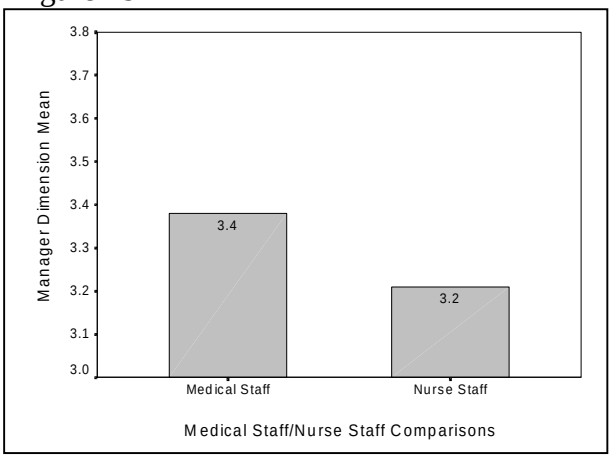

Figure 1b represents the mean differences between the medical staff and nurse staff on the 'Management' dimension.

Figure 1c

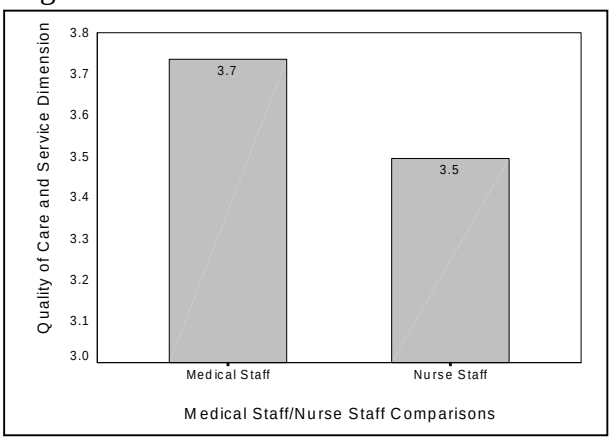

Figure 1c represents the mean differences between the medical staff and nurse staff on the 'Quality of Care and Service' dimension.

Figure 1. Dimension mean differences between groups.

Unidimensionality of Dimensions

Management dimension. Principal components factor analysis was used to assess the unidimensionality of the 21 item 'Management' dimension using the 1993 sample (see Appendix B for factor analysis results and descriptive statistics of the items by group). The first factor accounted for 46.9 percent of the variance. The first criteria of greater than 20 percent of the variance accounted for by the first factor was obtained. The default option used by SPSS includes all factors that have eigenvalues of greater than 1. Given this criterion, 2 factors would have been retained. Based on the results from Reckase (1979) and Drasgow and Parsons (1983) the amount of variance accounted for by the first factor is sufficient for the use of unidimensional IRT models. The second criteria, the first factor being clearly separated from the other factors, was assessed by examining the scree plot. This criterion was obtained, as can be seen in Figure 2.

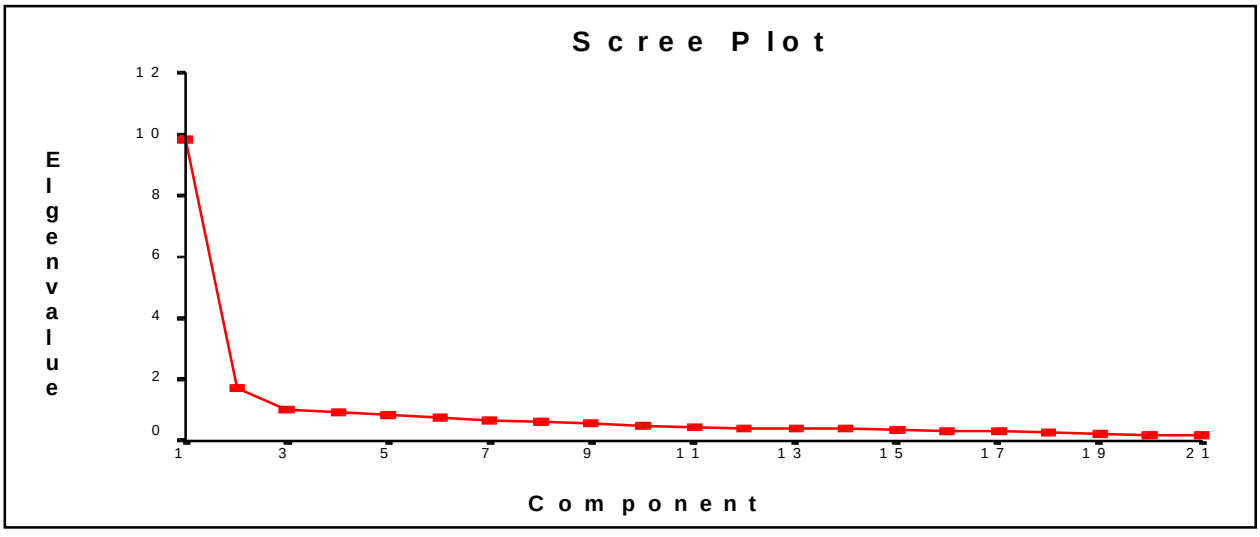

Figure 2. Scree plot the eigenvalues for all components of the 'Management' dimension.

Based on the results of principal components analysis produced by SPSS, it is concluded that the 'Management' dimension is sufficiently unidimensional to proceed with the next step of analysis.

Quality of Care and Service dimension. Principal components factor analysis was used to assess the unidimensionality of the 14 item 'Quality of Care and Service' dimension using the 1993 sample (see Appendix B for factor analysis results and descriptive statistics of the items by group). The first factor accounted for 37.86 percent of the variance. The first criteria of greater than 20 percent of the variance accounted for by the first factor was obtained. The default option used by SPSS includes all factors that have eigenvalues of greater than 1. Given this criterion, 3 factors would have been retained. Based on the results from Reckase (1979) and Drasgow and Parsons (1983) the amount of variance accounted for by the first factor is sufficient for the use of IRT models. The second criteria, the first factor being clearly separated from the other factors, was assessed by examining the scree plot. This criterion was obtained, as can be seen in Figure 3.

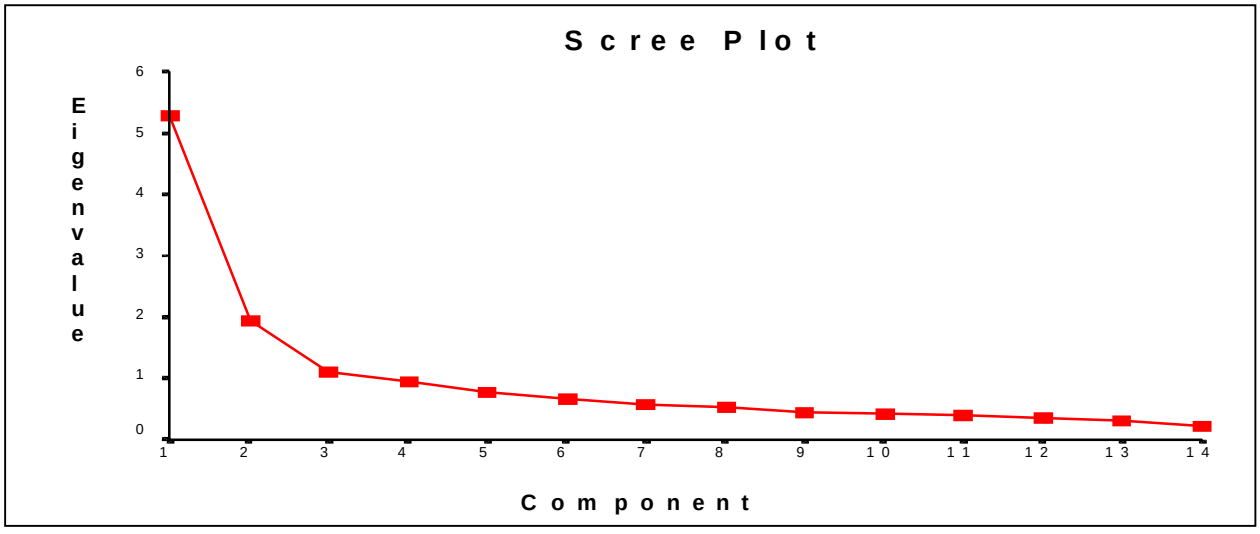

Figure 3. Scree plot the eigenvalues for all components of the 'Quality of Care and Service' dimension.

Although an argument could be made for a break after the second factor, the amount of variance accounted for by the first factor (37.86%) is substantially larger than the amount of variance accounted for by the second factor (13.9%). Based upon the results of principal components analysis produced by SPSS, it is concluded that the 'Quality of Care and Service' dimension is sufficiently unidimensional to proceed with the next step of analysis.

Item Parameter Estimation

The difficulty and discrimination parameters were estimated using MULTILOG. Command statements that specified the use of Samejima's (1969) two-parameter graded response model were used to estimate the item parameter estimates in MULTILOG. The item parameters were estimated separately for each group. One discrimination parameter and four difficulty parameters were estimated for each item. Estimates of θ were also calculated for the focal group for each comparison using MULTILOG. The initial estimates of all of the comparison groups can be seen in Appendix C. The likelihood χ^2 statistics for assessing the fit of Samejima's Graded Response Model were not interpretable due to the large samples used in this study (Koch, 1983). That is, all values were largely significant, indicating a poor fit, because the χ^2 statistic is very sensitive to sample size. Based on the factor analysis results, meeting the assumption of unidimensionality, the fit of the model will be assumed to be adequate for further analyses.

Linking Procedure

The FORTRAN program EQUATE 2.1 (Baker, 1993) was employed to match the estimated item parameters between each comparison group. For each comparison the group with the largest sample was used as the focal group. The non-manager (n = 3578) and nurse staff (n =

1086) groups were used as the focal groups and the manager (n = 650) and medical staff (n = 600) groups were used as the referent groups. The EQUATE 2.1 program yielded equating constants that are used in the DFITP4 program to transform the reference group parameters to the scale underlying the focal group. An iterative process was used to determine the EQUATE 2.1 constants if items were found to have DIF. The constants produced by EQUATE 2.1 and the iterations used to estimate the constants can be seen in Table 1.

Table1.

Equate Constants by Iteration.

	A	**K**
Manager/Non-manager		
Iteration 1	1.057	-.370
Iteration 2 (2 items removed - 10 and 13)	1.055	-.341
Medical Staff/Nurse Staff (Management Dimension)		
Iteration 1	.904	-.095
Medical Staff/Nurse Staff (Quality of Care and Service Dimension)		
Iteration 1	1.029	-.392

DFIT

The FORTRAN program DFITP4 (Raju, 1995) was used to estimate the CDIF and NCDIF indices. The program utilizes the equated item parameter estimates for both comparison groups and the θ estimates for the focal group. The summarized results are reported in Table 2. A glossary of relevant terms is also presented in Table 3 to give a brief summary of the different indices discussed in the following sections.

Table 2.

Items demonstrating DIF for each comparison group based on .096 criteria for NCDIF and DTF and an associated .01 χ^2 significance level.

	NCDIF	CDIF	DTF
Manager/Non-manager	Items 39 and 43	Item 33	Item 33
Medical Staff/Nurse Staff (Management Dimension)	____	____	____
Medical Staff/Nurse Staff (Quality of Care and Service Dimension)	____	Item 17	Item 17

Table 3.

Explanation of relevant terms.

Glossary of Important Terms	
Terms	**Explanation**
DIF	Differential Item Functioning, term used to describe the observation that an item displays different statistical properties between different groups. The presence of DIF indicates that an item measures different constructs for different individuals based on group membership.
DFIT	Differential Functioning of Items and Test, a new method for assessing DIF at both the item and test level.
NCDIF	Non-compensatory Differential Item Functioning, terms used in the DFIT framework to refer to normal DIF. This is a measure of DIF that has the assumption that no other items in the test have DIF.
CDIF	Compensatory Differential Item Functioning, term used in the DFIT framework to identify the impact of DIF at the test level. Compensatory DIF takes into account the interaction among items in the test. CDIF identifies items that minimizes the pattern of DIF accumulated across all items in the scale and not simply the item with the largest DIF.
DTF	Differential Test Functioning, terms used to indicate measurement differences between groups at the scale level. This measure gives an indication of the additive DIF impact that each item has on the scale. This measure, along with CDIF, helps to identify the least amount of items to remove to create a measure free from DIF. The items removed from the scale are not necessarily the items with the largest DIF. The goal of examining this index is to reduce the overall DIF of the test to zero. Items are removed based on how closely the value of DTF is reduced to zero.

Manager/Non-Manager Comparisons on the 'Management' Dimension

It was hypothesized that there would be significant DTF and NCDIF between the managers and non-managers on the 'Management' dimension of the attitude assessment survey. The current version of the DFITP4 program will only use 3000 estimates of θ for the focal group, so the number of θ estimates had to be reduced to use the program. The initial non-manager group had 3578 estimates of θ, so 578 estimates were not used in the analyses. Raju (personal communication, February 22, 1999) indicated that 3000 estimates of θ were sufficient for attaining stable estimates of the DFIT indices.

The results indicated that there were two items that demonstrated significant NCDIF; item 39 (NCDIF=.097; χ^2 = 31114 (P<.01)) and item 43 (NCDIF=.327; χ^2 = 85870 (P<.01))(see Appendix D for NCDIF values for all items). These items are: 39) My chief/supervisor clearly communicates the expectations for my work unit/practice team, and 43) I am involved in making decisions that affect my work/practice. These items were then removed from the sample of items and new equate constants were estimated (Table 1, iteration 2). Using these new equating estimates no new items were identified as having significant NCDIF so the analysis was stopped.

The analysis of DTF was also significant (DTF=.13488; χ^2 = 3180.49 (P>.01) according to the .096 cutoff and a .01 χ^2 significance level. Based on the mathematical equations used to derive the DFIT indices, item 33 was removed from the 'Management' dimension for the manager/non-manager comparison group because it had the largest impact on DTF. After the removal of this item the overall DTF (DTF=.04753; χ^2 = 3180.49 (P>.01)) became less than the designated cutoff of .096 and the χ^2 probability was greater than .01. The analysis of DTF indicates that only one item would need to be removed from the 'Management' dimension to make DTF non-significant for the manager/non-manager comparison groups.

Medical Staff/Nurse Staff Comparisons on the 'Management' Dimension

It was hypothesized that there would be significant DTF and NCDIF between the medical staff and the nurse staff on the 'Management' dimension of the attitude survey. The results of the medical staff/nurse staff 'Management' dimension comparisons indicated no items that demonstrated NCDIF (see Appendix D for NCDIF values for all items). DTF was also not found to be significant (DTF=.01368; χ^2 = 1137.16 (P>.01)). Thus, part of hypothesis 2 was not supported.

Medical Staff/Nurse Staff Comparisons on the 'Quality of Care and Service' Dimension

In the second part of hypothesis 2 it was hypothesized that there would be significant DTF and NCDIF between the medical staff and nurse staff on the 'Quality of Care and Service' dimension of the attitude survey. The results of the medical staff/nurse staff 'Quality of Care and Service' dimension comparisons indicated no items that demonstrated NCDIF (see Appendix D for NCDIF values for all items). DTF was found to be significant (DTF=.20701; χ^2 = 2012.02 (P<.01)), thus the item with the largest contribution to DTF was removed first and the estimate of DTF was re-estimated. The item that was removed was number 17. DTF became non-significant (DTF=.05022; χ^2 = 104.94 (P>.01)), based on the cutoff value of .096 and χ^2 value with a p<01. This indicates that the rest of the items demonstrating CDIF were comparably matched so as to cancel out the effects of DIF on DTF. Based on these results, the second part of hypothesis two was partially supported.

Discussion

The following discussion is organized by comparison groups and dimension used. Possible rationale is provided for items that demonstrated NCDIF. Second, these results are compared with the findings of the Lynch et al. (1998a) study. Third, the evaluation of DTF and CDIF is explored. Fourth, limitations of the study are addressed. Finally, implications and further research based on the results are explored.

Manager/Non-manager Comparisons

The hypothesis that the comparison of managers/non-managers on the 'Management' dimension will result in significant NCDIF and DTF was supported. Two items (39 and 43) were found to have significant NCDIF. These items are: 39) My chief/supervisor clearly communicates the expectations for my work unit/practice team, and 43) I am involved in making decisions that affect my work/practice. The boundary response functions (BRFs) for items 39 and 43 are presented in Figure 4 and 5, respectively. These figures represent the probability of responding to each of the five response categories at a particular level of θ. The item response categories are: strongly disagree (SD), disagree (D), neither agree nor disagree (N), agree (A), and strongly agree (SA). The BRFs of the focal group (non-managers) are black and the BRFs of the reference group (managers) are red. Figures 4 and 5 are read by starting at a certain value θ, as indicated on the x-axis, and following that point up until it crosses all possible BRFs. At each point where a BRF is crossed, the probability of responding to that response category for each group is located directly across on the y-axis. If there is a large difference between the two groups of interest then DIF is occurring. That is, at the same level of θ the groups have different probabilities of responding to the certain option. If these people truly have the same level of θ

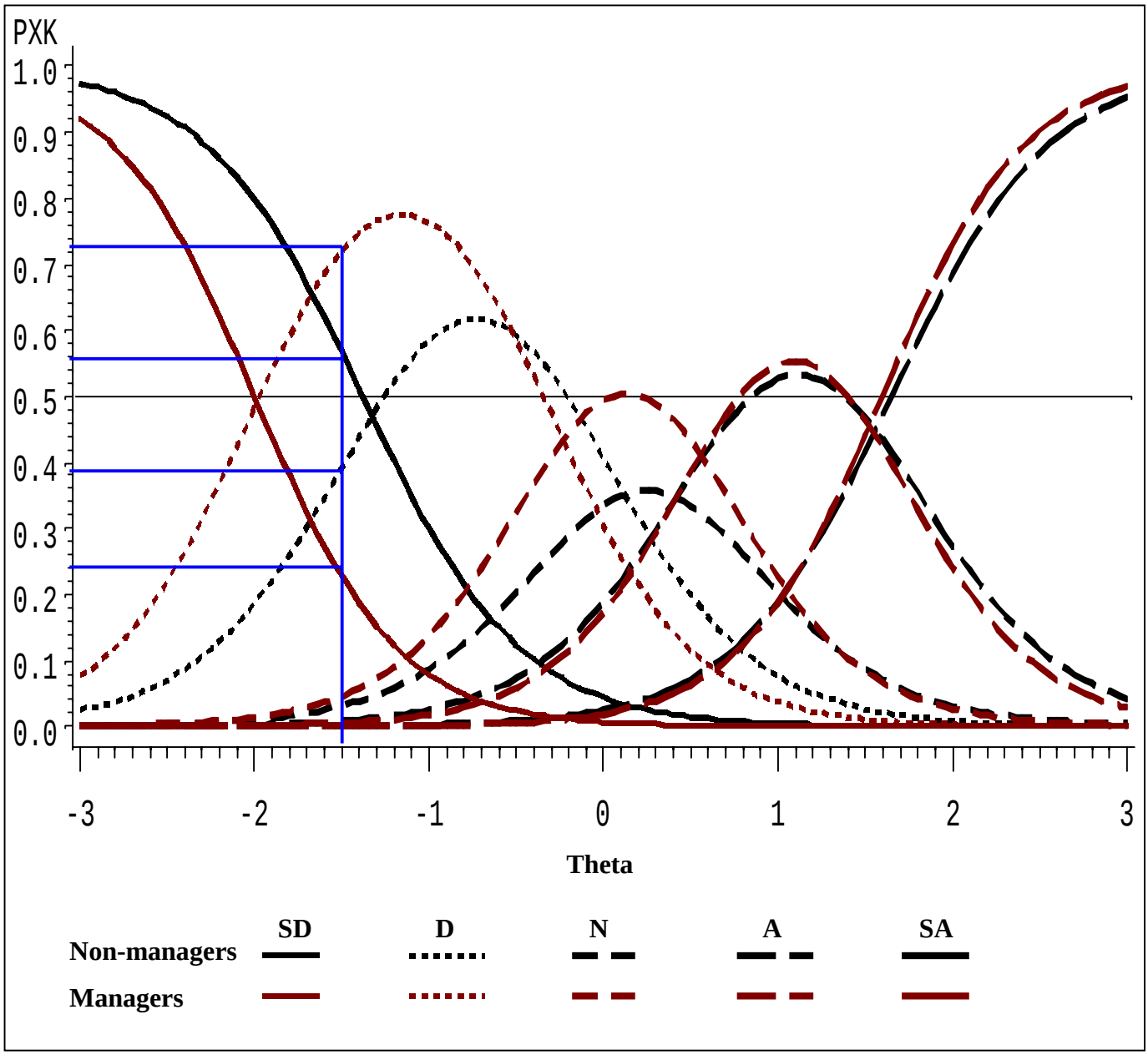

Figure 4. The BRFs for item 39. The black line represents the non-manager group and the red line represents the manager group. Each pattern corresponds to a BRF for each response category.

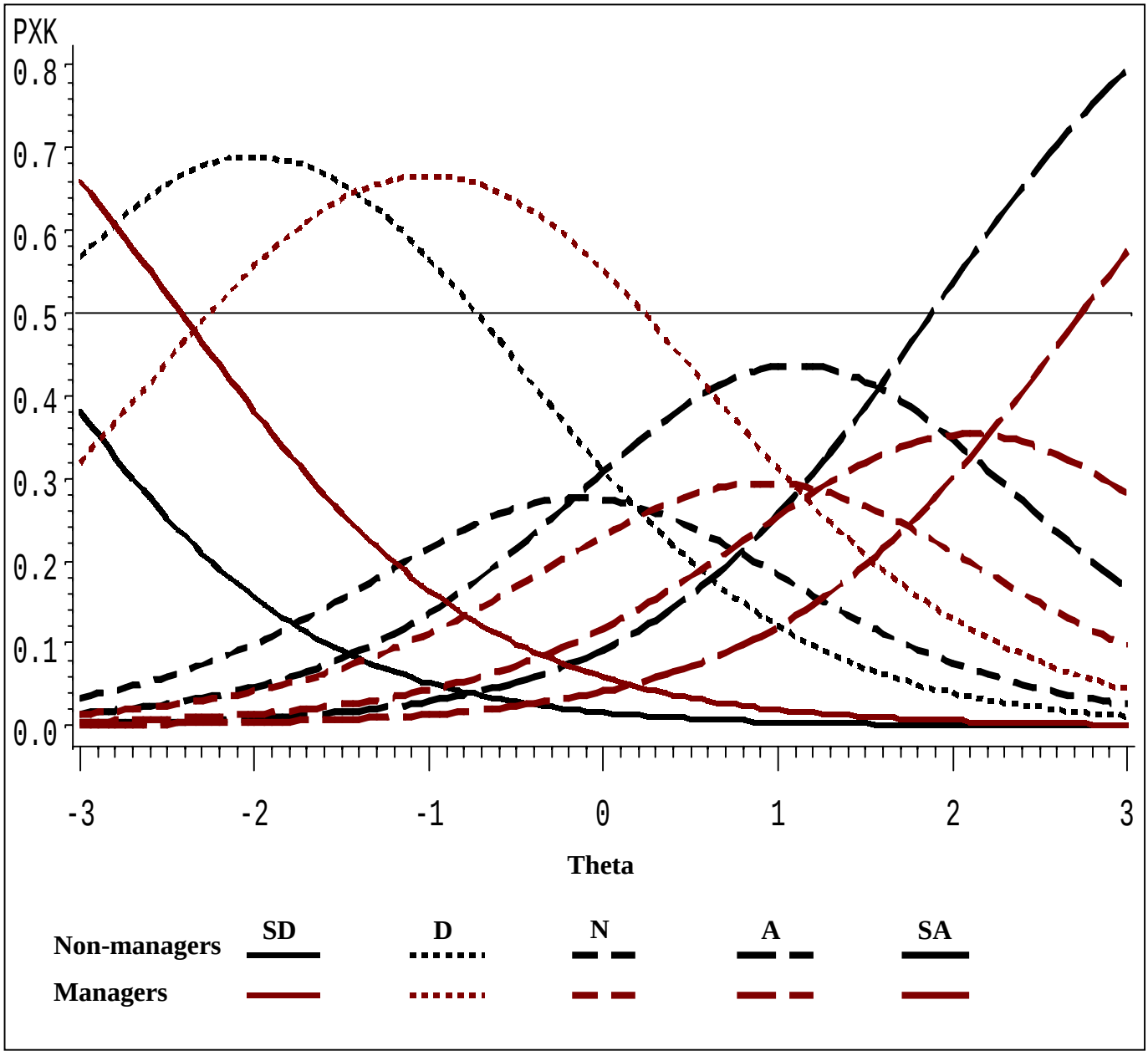

Figure 5. The BRFs for item 43. The black line represents the non-manager group and the red line represents the manager group. Each pattern corresponds to a BRF for each response category.

then something besides θ is accounting for the different probabilities of responding to the items. An example of DIF can be seen in Figure 4 by examining the blue lines.

This example is based on item 39, which is, 'My chief/supervisor clearly communicates the expectations for my work unit/practice team'. Figure 4 indicates that at θ of -1.5, more dissatisfied with management, the probability of a manager responding to the disagree option is .73. However, non-managers only have a .4 probability of responding to the option, even when they have the same satisfaction with management. Furthermore, the non-managers have an even higher probability (.56) of responding to the strongly disagree option at the same level of θ while the managers have a low probability of responding to the strongly disagree option (.24).

Item 39 was found to have significant NCDIF. As can be seen in Figure 4, DIF of item 39 is located within the lower levels of θ. That is, at lower levels θ, or less satisfied, the non-managers were more likely to indicate that they more strongly disagreed that their chief/supervisor clearly communicates work expectations. Managers, at the same level of θ, have a higher probability of responding to the disagree option than the strongly disagree option. The more dissatisfied the non-managers are with management, the more likely they are to indicate that their managers do not clearly communicate their job roles. The results from this item could be related to the fact that managers' jobs are more self-directed which requires them to define their own roles within the organization. Managers may not be as likely to indicate that they strongly disagree with this item because it is up to them to define their own roles. Non-managers, however, might be more sensitive to the clarity of their job roles. Non-managers may mark strongly disagree to this item if management is providing little direction because they feel more frustrated or powerless to deal with problems. Thus, non-managers may need more guidance with respect to their job roles. This might be an example of an item with greater

consequence or importance to the non-management group. The results from this item suggest that responses may be confounded with the type of job and roles associated with the job that the individuals hold, which leads to the conclusion that comparisons between dissatisfied non-managers and managers should be more closely examined.

These results also indicate that self-rating bias may be occurring. Managers may not have as high of a probability of answering the strongly disagree option because they are interpreting the question as a self-rating of how well they clarify job roles to their subordinates. The conclusion that comparisons between these groups should be more closely examined is also warranted if managers are interpreting the item as a self-rating.

Item 43 is 'I am involved in making decisions that affect my work/practice'. Figure 5, BRFs of item 43, demonstrates a situation were DIF occurs at all levels of θ. The pattern of DIF indicates that the non-management group is more likely to agree at higher levels of θ and less likely to disagree at lower levels of θ. That is, at all levels of θ the non-managers are more likely to respond in an agreeing manner. This is an interesting finding in that it would seem more reasonable that those with the most actual decision making authority would likely be relatively more satisfied with this aspect of management. However, the different roles that managers and non-managers engage in could explain the differences found in item 43. A possible reason for DIF of this item is that the nature and type of decisions that affect a person's job may be qualitatively different at these organizational levels. That is, decisions that affect managers' jobs are things such as budgets and staffing or major organizational policies, whereas decisions affecting non-managers' jobs are more related to work rules and standard operating procedures. Managers may be more sensitive or powerless to affect these major organizational decisions, while non-managers may understand that their roles exclude them from many of the decisions

being made within the organization. This particular organization had been going through a period of downsizing and budget cuts, which could have contributed to frustration felt by managers. While these events are likely to affect non-manager attitudes, it is less likely that it will be reflected in their attitudes towards their decision making power. A way to phrase the question to better equate the two groups might be: “I am involved as appropriate when decisions are made that affect how I conduct my own work.” In this way the focus of the question is more directly on the person's immediate job and work environment rather than on organizational issues.

The results based on the computation of CDIF and DTF also indicated that if item 33 was removed from the scale the additive affects of DIF would cancel out and the overall scale would not have significant DTF. A discussion of the implications of these results will be discussed below.

Medical Staff/Nurse Staff Comparisons on the 'Management' Dimension

The hypothesis that DIF is occurring within the 'Management' dimension between the medical staff and nurse staff was not supported. The DFIT analyses indicated that no items demonstrated NCDIF. Furthermore, DTF was also found to be non-significant, which exempts the scale from having possible significant CDIF. Although the hypothesis was not supported, the implications of the results can be considered beneficial to those practicing current techniques for analyzing and comparing attitude survey results. That is, comparisons made between medical staff and nurse staff on their feelings about management can accurately be assessed and compared. Mean differences that are present actually represent their differences in feelings about their management.

Medical Staff/Nurse Staff Comparisons on the 'Quality of Care and Service' Dimension

The hypothesis of DIF within the 'Quality of Care and Service' dimension was partially supported. Although no items were found that demonstrated significant NCDIF, DTF for the dimension was significant. This is a perfect illustration of one of the advantages of using the DFIT framework to analyze not only DIF at the item level, but also at the scale or dimension level. The results obtained in the present study indicate that no single item had particularly strong DIF, as measured by traditional measures of DIF, but the cumulative effects of all the items is significant. Further inspection of the DTF indices will be explained in the following sections.

Comparison of DIF Results with Lynch et al.'s (1998a) Study

The results of the Lynch et al. (1998a) study provided a very interesting path with which to take advantage of the newly developed DIF techniques. Lynch and her colleagues found one item that had significant DIF within each of the two dimensions examined. The two dimensions they examined were a 'Management' dimension and a 'Customer Orientation' dimension, which were compared between managers and non-managers. The item found to have significant DIF on the 'Management' dimension was 'My SBU/SSU management consistently communicates in a candid, frank, and open manner' (p. 9). The item that was found to have DIF in the 'Customer Orientation' dimension was '[company name] responds quickly to changes in the marketplace - customers and competitors' (p. 10).

There are several reasons why it is important to assess the generalizability of their findings. First, it is always important to examine the effects of a study within a different sample of people. Second, it is also important to take advantage of a different attitude survey. Finally, the use of a different method of DIF detection was critical in this instance.

The method used for identifying DIF in the Lynch et al. (1998) study was certainly less than rigorous, as mentioned in their paper (p. 12). Lynch et al. used a method referred to as the 'designated anchor' method. The concept of this method is to identify a few items that have equal means for both groups, assuming that these items do not have DIF because of the similar mean scores. These items are then used as the 'anchor' items. Then, the item under investigation for DIF is chosen. The two models are then pitted against one another. In the first model all parameter estimates are fixed to be equal across groups. In the second model the parameters for the item suggested to have DIF are freed to allow variation between the two groups while the 'anchor' items are still fixed to be equal. Then, the fit of the models is examined. The problem with this method is that it relies on the use of mean differences to identify DIF. As mentioned above, mean differences do not necessarily indicate DIF. Therefore, the current study employed a more rigorous examination of DIF in order to validate the findings of Lynch et al..

The current finding of this study suggest that the Lynch et al.'s (1998) results are generalizable to another sample and attitude survey using different techniques. While the analysis in this study found 2 out of 21 items to have significant DIF between the managers and non-managers, Lynch et al. found 2 out of 11 items to function differently for these groups. These results indicate that there is a likelihood of some consistent DIF between the managers and non-managers on attitude assessment surveys. Although the items found to have DIF were different between the two studies for the manager and non-manager, there is a concern that the likelihood of DIF is present between manager and non-managers on any attitude assessment scale.

DTF Results

The discussion of NCDIF is important because it helps to identify items that have significant DIF, even if their contribution to the overall test is cancelled out by other items in the scale. The evaluation of individual items allows researchers to make post hoc interpretations of DIF. These interpretations can be used to further understand attitude measurement within organizations. It is also important to note other important aspects of the DFIT indices. The results of the DTF indices are helpful in evaluating the extent to which each of the dimensions is comparable or can be made comparable.

The unique contribution of the DFIT framework lies within the detection of DTF and CDIF. While traditional measures of DIF have the assumption that no other items in the test have DIF, CDIF and DTF are evaluated assuming that the DIF of multiple items contributes to the overall DTF. The CDIF index helps to determine the amount of contribution, in conjunction with other items, a single item has on the overall DTF. This advantage allows for the evaluation and removal of items that contribute most substantially to the DTF index. In this study it was found that one item for the 'Management' dimension and one item for the 'Quality of Care and Service' dimension should be removed based on the CDIF index. An understanding of why an item is removed requires a more detailed description of the CDIF index. According to Raju (personal communication, February 20, 1999), CDIF is just a special case of NCDIF, a traditional measure of DIF. That is, CDIF is equal to NCDIF if no other items in the measure have DIF because CDIF is based on the covariance of the item with the total true score difference (refer to equation 11). When the value of DTF is above the cutoff, .096 in this study, the item that contributes most to the DTF index is removed. The item to be removed is not solely based on the CDIF index. Instead, 2*CDIF-NCDIF is computed for each item to

determine the impact of that item on the entire test. As can be seen from this formula, if all the other items in the measure do not have DIF, then CDIF is equal to NCDIF and the item that is removed would have the highest NCDIF or CDIF value. So, given the more likely situation in which there are multiple items with DIF, when an item is removed not only is its CDIF value removed but also the effect that the item has on the CDIF indices of other items are removed, CDIF-NCDIF. The DTF and CDIF indices help to determine the least amount of items that need to be removed to make a scale free from DTF. In order for this to be accomplished, the value of DTF must be reduced to zero. So, the item that is removed is the item that will reduce the value of DTF the closest to zero. Given this criteria for item removal, it is not always the item with the largest DIF that is removed from the scale because the item with the largest DIF, when removed, may cause the scale to become biased in the opposite direction. For example, if you evaluated the following series of numbers: -6, 6, -4, 4, 2, the number 2 would be removed to obtain a sum of zero.

The CDIF and NCDIF indices hold different value to the researcher. First, if the goal of the researcher is to create a measure that does not have overall DTF, then CDIF is appropriate to evaluate because of its properties described above. These indices can help in the construct validation of attitude assessment research. That is, more confidence can be placed in the interpretations of comparisons when a scale is free from DTF.

Second, if the goal of the researcher is to identify particular items that have the most significant DIF, as described in traditional measures of DIF, NCDIF is the most important index to evaluate. Although the first goal is important, it is the second goal that has received the most attention in the current study. While it is important to note that the removal of one item from the 'Management' and 'Quality of Care and Service' dimensions would have resulted in dimensions

free from DTF, the examination of specific items can contribute more to understanding the nature of DIF in attitude assessment surveys. DTF findings, however, guide further development of attitude scales that are comparable between groups. When between group comparisons is an important goal of attitude surveys such analyses could lead to the creation of stronger and more interpretable instruments.

The results indicate that both the 'Management' dimension and the 'Quality of Care and Service' dimension would have non-significant DTF if one item were removed. This is of interest and practical concern if the removal of one item in a scale has any impact on the findings of mean differences between the groups. A comparison of the mean differences between groups after the items were removed can be seen in Figures 8 and 9. These figures demonstrate that the removal of the significant CDIF items has little impact on the overall mean differences between the groups. These results help to validate comparisons that are made between groups at the dimension level. For example, managers are still more satisfied with management than non-managers when the dimension has non-significant DTF. Thus, the impact of DIF in this study lies more in a method for identifying, interpreting, and improving individual items rather than interpreting results at the dimension level.

Summary of DFIT Analyses

The DFIT framework provided useful indices from which to identify the impact of DIF at both the item and the test level. The results of the DFIT indices usually identify more items using the NCDIF index than the CDIF index (Flanagan, Raju & Haygwood, 1998). The cumulative nature of CDIF allows for the effects of one item to cancel out the effects of another item on DTF, thus reducing the expected number of items that are identified. The largest amount

Figure 6a

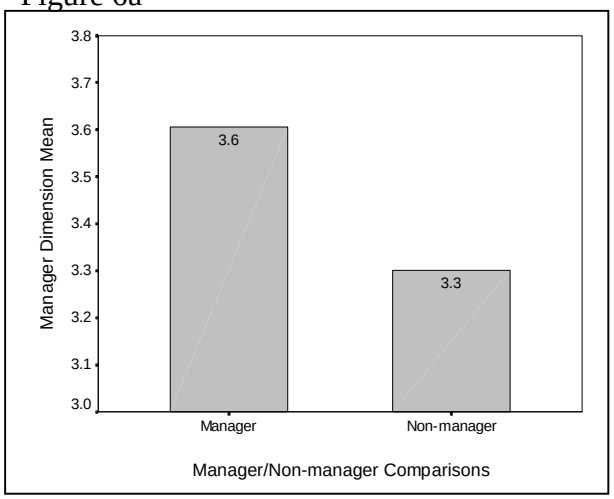

Figure 6a represents the mean differences between the managers and non-manager on the 'Management' dimension. The values were computed after item 33 was removed from the dimension.

Figure 6b

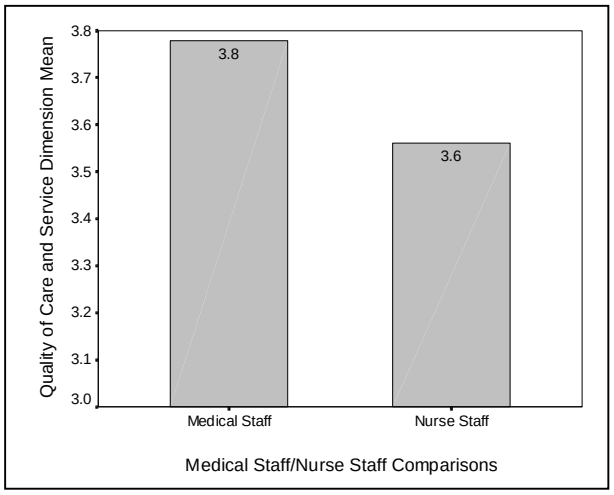

Figure 6b represents the mean differences between the medical staff and nurse staff on the 'Quality of Care and Service' dimension. The values were computed after item 17 was removed from the dimension.

Figure 6. Comparison of mean differences between groups after items that were identified as having significant CDIF were removed from the dimension. These means are representative of dimensions with non-significant DTF.

of DIF was found for the manager/non-manager comparisons on the 'Management' dimension. The results of the analyses indicated that 2 out of the 21 items in that dimension demonstrated significant NCDIF. If dimension level DIF is explored, one item would be removed in order to remove the effects of DIF on the entire dimension. This was also found to be the case for the medical staff/nurse staff comparison on the 'Quality of Care and Service' dimension. This

comparison demonstrated a situation in which items that demonstrated CDIF were identified to have DIF. The overall results of the current study indicate that DIF is not a serious problem when comparisons were made between medical staff and nurse staff.

Limitations

The first limitation plagues all DIF studies. The complex nature of detecting DIF only allows for a limited amount of comparisons to be made. While researching and practicing attitude assessment it is evident that many comparisons are made between groups within the organization. While manager/non-manager comparisons have been found to demonstrate DIF between different scales within an attitude assessment, this is the only study that has evaluated these differences within other comparison groups. The medical staff/nurse staff comparison was made in the current study because of the highly salient differences between these groups, but certainly other comparisons could have been made. For example, comparisons are often made between hourly and salaried employees, work groups (often defined by geographic areas), and other positions within the organization of interest. Further studies could also follow the path in which Collins (1996) conducted his research. That is, traditional DIF comparisons could be made using such variables as racial groups, gender, and age.

Examining the abundance of differences present is only possible if many studies are conducted or the evaluation DIF becomes less cumbersome. The techniques currently available to analyze DIF require complex calibrations that are time consuming. Also, there are numerous methodological variations that need to be considered when evaluating DIF. For example, a single method for assessing DIF has not been agreed upon. As indicated in the literature review presented above, each step of the analyses encompasses a decision to be made about the 'right' method. The right method depends not only on the structure of the survey, polytomous or

dichotomous, but also on the theory that guides your research, IRT or CTT. Depending on the theory used, the results of the DIF study may be different. It is likely that future programs will be created to make the detection of DIF less cumbersome. It is possible that these techniques will become common methods for validating the comparisons made between groups on attitude assessments within organizations.

The second limitation of the current study is due to the particular method used to assess DIF. The DFIT framework, although gaining in popularity, has yet to receive enough rigorous empirical testing. This problem was pronounced when making the decision of an appropriate cutoff score to use to identify DIF. Although recent literature has used a .024 cutoff for NCDIF and DTF, the most current suggestion for a valuable cutoff is .096 for a five-point scale. The new, more stringent cutoff values are based solely on studies conducted with dichotomous data (e.g., Raju, 1995). The extrapolation to polytomous data should be verified through empirical work.

Third, the power to explain why DIF is occurring is fairly weak. Although suggestions for why DIF might be occurring in specific items were given, these are post-hoc explanations without a confirmation of the exact reason for the differences. The only consideration in dealing with this limitation is that the identification of significant DIF is important by itself. The identification of DIF will hopefully lead researchers and practitioners to consider ways to develop attitude surveys that do not demonstrate DIF. Some suggestions are indicated in the section below.

Implications and Future Research

The implications of this study are important to those researching, developing, and conducting attitude assessment surveys. The first implication is that comparisons between groups must be done carefully. Some groups may exhibit DIF, and subsequently comparison might be misinterpreted. As indicated above, the current study only provided information about a few possible comparisons. Future research could be conducted on any number of these other comparison groups and with other dimensions of attitude surveys.

Further research needs to be done to determine how robust attitude measurement scales are to the impact of DIF. It is a matter for further research whether or not findings of DIF are simply applicable to a particular setting (i.e., when looking at medical staff/nurse staff comparisons with this survey), in this organizations, or whether these findings are indicative of a generalizable phenomenon. An example of this would be the consistent findings of DIF between the manager/non-manager groups. These findings indicate that these groups may be answering some questions based on different interpretations. This finding may well be generally applicable to a wide variety of settings where this type of comparison might be made.

DIF indicates that results should not be compared between groups. Future interpretation of attitude assessment data may need to focus on specific interpretation of the items by group, as indicated by the results from the manager/non-manager comparisons. Other comparisons need to be examined to determine which groups can be compared. However, the results indicated that when items were removed from the dimensions to make DTF non-significant the mean differences were similar to the mean differences of the groups when these were included. This information is valuable to the interpretation at the scale level. For example, the statement that

managers are more satisfied with management than non-managers can be considered a valid comparison. That is, DIF alone does not explain away the mean differences between groups.

With the development of more accessible methodologies, investigation of DIF may become a standard practice when evaluating measurement scales in a specific research setting. In much the same way that a distribution of scores may be checked for normality or the internal consistency reliability of a test calculated, an index of DIF is a useful tool in understanding the specific measurement properties of an instrument. Problem items can be clearly identified. Although possible reasons can be theorized for why the DIF is occurring, such as self-rating bias or job roles, it is nearly impossible to identify the exact causes of the DIF. But identifying the specific cause of DIF is not necessarily as important as the knowledge that DIF exists. The knowledge that DIF is present indicates that the dimension, or certain items in the dimension, could be written more specifically to reduce DIF.

With the introduction of the new DTF indices, attitude scale development could incorporate the removal or modification of items that contribute the most to the overall DTF of the dimensions. This can help to provide evidence of construct validity for a scale by ensuring that the statistical properties of the attitude assessment scale are the same for different groups. This will help to ensure that appropriate comparisons at the dimension level are made between groups.

Conclusion

The results of the Lynch et al (1998a), Collins (1996) and the current study consistently show that only a few items, if any, demonstrate DIF in any one dimension. These results are very promising for the robustness of attitude assessment practice within organizations. Given that organizations readily compare groups it is important to find that that there is some support for these comparisons. The examination of the differential item functioning at the dimension level indicated that mean differences commonly found between certain groups within an organization are not accounted for by DIF. So, dimension level comparisons would be appropriate. The individual item analyses indicate that future research should be conducted to evaluate specific items before comparisons are made based on these items. At a minimum the technique proves highly useful in identification and improvement of specific items that may be subject to differential interpretation.

References

Angoff, W. H. (1993). Perspective on Differential Item Functioning Methodology. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 3-23). Hillsdale, NJ: Lawrence Earlbaum.

Baker, F. B. (1995). EQUATE 2.1: Computer Program for Equating Two Metrics in Item Response Theory [Computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Bernstein, E. J. (1981). Employee Attitude Surveys: Perception vs. Reality. Personnel Journal, 300-305.

Brief, A. P. (1998). Attitudes In and Around Organizations. Prentice Hall; New York; New York.

Camilli, G., Wang, M. and Fesq, J. (1995). The Effects of Dimensionality on Equating the Law School Admissions Test. Journal of Educational Management, 32, 1, 79-96.

Campbell, D. T. (1950). The incorrect assessment of social attitudes. The British Journal of Social Psychology, 47, 15-39.

Canter, R. R. (1948). Psychologists in Industry. Personnel Psychology, 1, 145-161.

Cattell, R. B., Maxwell, E. F., Light, B. H. & Unger, M. P. (1949). The Objective Measurement of Attitudes. The British Journal of Psychology, 40(2), 81-90.

Cohen, A. S. & Kim, S. (1998). An Investigation of Linking Methods Under the Graded Response Model. Applied Psychological Measurement, 22(2), 116-130.

Cohen, A. S. & Kim, S. (1993). A Comparison of Lord's $\chi 2$ and Raju's Area Measures in Detection of DIF. Applied Psychological Measurement, 17(1), 39-52.

Cohen, A. S., Kim, S. & Baker (1993). Detection of Differential Item Functioning in the Graded Response Model. Applied Psychological Measurement, 17, 335-350.

Cole, R. J. (1940). A Survey of Employee Attitudes. Public Opinion Quarterly, 4, 497-506.

Collins, W. C. (1996). An Empirical Investigation of the DFIT Framework for Measuring DTF and DIF in a Polytomous Satisfaction Scale. Unpublished Ph.D. dissertation, Georgia Institute of Technology.

Crocker, L. & Algina, J. (1986). Introduction to Classical and Modern Test Theory. Fort Worth, TX: Harcourt Brace Jovanovich.

Dansereau, F., Graen, G., & Haga, W. J. (1975). A Vertical Dyad Linkage Approach To Leadership Within Formal Organizations. Organizational Behavior and Human Performance,13,46-78.

Day, D. D. (1940). Methodological Problems in Attitude Research. Journal of Social Psychology, 14, 165-179.

Dorans, N. J. & Holland, P. W. (1993). DIF Detection and Description. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 35-66). Hilsdale, NJ: Lawrence Earlbaum.

Drasgow, F. and Hulin, C. L. (1990). Item Response Theory. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of Industrial Organizational Psychology (Vol. 1, pp. 577-636). Palo Alto, CA: Consulting Psychologists Press, Inc.

Drasgow, F. & Parsons, C. K. (1983). Application of Unidimensional Item Response Theory Models to Multidimensional Data. Applied Psychological Measurement, 7, 189-199.

Dutka, S. & Frankel, L. R. (1993). Measurement Errors in Organizational Surveys. American Behavioral Scientist, 36(4), 472-484.

Edwards, A. L. & Kilpatrick, F. R. (1948). A Technique for the Construction of Attitude Scales. Journal of Applied Psychology, 32, 374-384.

Edwards, J. E. & Thomas, M. D. (1993). The Organizational Survey Process. American Behavioral Scientist, 36(4), 419-442.

Farh, J., Dobbins, G. H. & Cheng, B. (1991). Cultural Relativity in Action: A Comparison of Self-Ratings Made by Chines and U.S. Workers. Personnel Psychology, 44, 129-147.

Festinger, L. (1957). A Theory of Cognitive Dissonance. CA: Stanford University.

Flanagan, W. J., Raju, N. S. & Haygood, J. M. (1998). Impression Management, Measurement Equivalence, and Personality Factors: Can IRT be Used to Determine the Impact of Faking. Paper presented at the 13th annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Gallup, G. (1988). Employee Research: From Nice to Know to Need to Know. Personnel Journal, 67, 42-43.

George, J. M. & Jones, G. R. (1996). The experience of Work and Turnover Intentions: Interactive Effects of Value Attainment, Job Satisfaction, and Positive Mood. Journal of Applied Psychology, 81(3), 318-325.

Greenwald, A. G. & Banaji, M. R. (1995). Implicit Social Cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102, 4-27.

Hambleton, R. K. (1989). Principles and Selected Applications of Item Response Theory. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 147 – 200). New York: Macmillan.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Newbury Park: Sage.

Harris, M. M. & Schaubroeck, J. (1988). A Meta-Analysis of Self-Supervisor, Self-Peer, and Peer-Supervisor Ratings, Personnel Psychology, 41, 43-62.

Harvey, R.J. and Thomas, L. A. (1996) Using Item Response Theory To Score the Myers-Briggs Type Indicator: Rational and Research Findings. Journal of Psychological Type, 37, 16 -60.

Houston, W. M., Raymond, M. R. & Svec, J. C. (1991). Adjustment for Rater Effects in Performance Assessment. Applied Psychological Measurement, 15(4), 409-421.

Illgen D. R. & Hollenbeck J. R. (1991). The Structure of Work. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of Industrial Organizational Psychology (Vol. 2, p. 165-208). Palo Alto, CA: Consulting Psychologists Press, Inc.

Irwin, J. W. (1945). Sampling Workers' Opinions. Dun's Review, 52, 32-40.

Jacoby, S. M. (1988). Employee attitude surveys in historical perspective. Industrial Relations, 27(1), 74-93.

Katz, D. (1941). The Public Opinion Polls and the 1940 Election. Public Opinion Quarterly, 5, 52-78.

Katz, D. (1942). Do Interviewers Bias Poll Results? Public Opinion Quarterly, 6, 248-268.

Koch, W. R. (1983). Likert Scaling Using the Graded Response Latent Trait Model. Applied Psychological Measurement, 7(1), 15-32.

Kornhauser, A. W. (1930). The Study of Work Feelings'. Personnel Journal, 8, 348-351.

Kornhauser, A. W. (1946). Are Public Opinion Polls Fair to Organized Labor? Public Opinion Quarterly, 10, 484-509.

Kornhauser, A. W. (1947). Industrial Psychology as Management Technique and as Social Science. The American Psychologist, 2, 224-229.

Koslowsky, M. Sagie, A., Krausz, M. & Singer, A. D. (1997). Correlates of employee lateness: Some Theoretical Concerns. Journal of Applied Psychology, 82(1), 79-88.

Laffite, L. J., Raju, N. S., Scott, J. C. & Fasolo, P. M. (1998). Examination of the Measurement Equivalence of a 360 Feedback Assessment with Confirmatory Factor Analysis and Item Response Theory. Paper presented at the 13th annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. New Jersey: Earlbaum,

Lynch, A., Barnes-Farrell, & Kulikowich, J. (1998a). Do Organizational Survey Items Function Differently for Managers and Non-managers? Paper presented at the 13th annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Lynch, A., Barnes-Farrell, & Kulikowich, J. (1998b). Using Samejima's Graded Response Model of Employee Attitude Survey Items. Paper presented at the 13th annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Maurer, T. J., Raju, N. S. & Collins, W. C. (1998). Peer and Subordinate Performance Appraisal Measurement Equivalence. Journal of Applied Psychology, 81(5), 693-702.

Marks, M. L. (1982). Conducting an Employee Attitude Survey. Personnel Journal, 684-693.

McGuire, W. J. (1976). The Concepts of Attitudes and Their Relations to Behaviors. In H. W. Sinaiko & A. B. Broedling (Eds.), Perspectives on Attitude Assessment: Surveys and Their Alternatives (pp. 7-38). Champaign, Illinois: Pendleton.

Mckinley, R. & Mills, C. N. (1985). A Comparison of Several Goodness-of-Fit Statistics. Applied Psychological Measurement, 9(1), 49-57.

McMurry, R. N. (1932). Management's Reactions to Employee Opinion Polls. Journal of Applied Psychology, 30, 212-219.

McNemar, Q. (1946). Opinion-Attitude Methodology. Psychological Bulletin, 43(4), 289-374.

Millsap, R. E. & Everson, H. T. (1993). Methodology Review: Statistical Procedures for Assessing Measurement Bias. Applied Psychological Measurement, 17, 297-334.

Mobely, W. H., Griffeth, R. W. Hand, H. H. & Mezlino, B. M. (1979). Review and Conceptual Analysis of the Employee Turnover Process. Psychological Bulletin, 86, 493-522.

Moretti, D. M. (1986). The Prediction of Employee Counterproductivity Through Attitude Assessment. Journal of Business and Psychology, 1(2), 134-147.

Park, D. & Lautenschlager, G. L. (1990). Improving IRT Item Bias Detection With Iterative Linking and Ability Scale Purification. Applied Psychological Measurement, 14(2), 163-173.

Petty, R. E. (1995). Attitude Change. In A. Tesser (Ed.), Advanced Social Psychology (pp. 195-255). New York: McGraw-Hill.

Potenza, M. T. & Dorans, N. J. (1995). DIF Assessment for Polytomously Scored Items: A Framework for Classification and Evaluation. Applied Psychological Measurement, 19(1), 23-37.

Putnam, M. L. (1930). Improving employee relations: a plan which uses data from employees. Personnel Journal, 8, 314-325.

Raju, N. (1995). DFITPUA: A Computer Program for Analyzing Differential Item and Test Functioning [Computer program]. Atlanta: Georgia Institute of Technology.

Raju, N., Van der Linden, W., & Fleer, P. (1995). An IRT-Based Internal Measure of Test Bias With Applications for Differential Item Functioning. Applied Psychological Measurement, 19, 353-368.

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multi-Factor Tests: Results and Implications. Journal of Educational Statistics, 4, 207-230.

Rosenfeld, P., Edward, J., & Thomas, M. D. (1993). Improving organizational surveys. American Behavioral Scientist, 36(4), 414-418.

Rothwell, W. J. (1983). Conducting an Employee Attitude Survey. Personnel Journal, 308-311.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. Psychometrika Monograph Supplement, 34 (4, Pt. 2).

Scarpello, V. & Vandenberg, R. J. (1991) Some Issues to Consider When Surveying Employee Opinions. In J. W. Jones, B. S. Steffy & D. W. Bray (Eds.), Applying Psychology in

Business: The Handbook of Managers and Human Resource Professionals (pp. 623-639). Lexington, MA: Lexington Books.

Shiemann, W. A. (1991) Using Employee Surveys to Increase Organizational Effectiveness. In J. W. Jones, B. S. Steffy & D. W. Bray (Eds.), Applying Psychology in Business: The Handbook of Managers and Human Resource Professionals (pp. 623-639). Lexington, MA: Lexington Books.

Sherif, M. & Cantril, H. (1945). The Psychology of 'Attitudes' Part I. The Psychological Review, 52(6), 295-319.

Sherif, M. & Cantril, H. (1946). The Psychology of 'Attitudes' Part II. The Psychological Review, 53(1), 1-24.

Sijtsma, K. (1998). Methodology Review: Nonparametric IRT Approaches to the Analysis of Dichotomous Item Scores. Applied Psychological Measurement, 22(1), 3-31.

Stagner, R. (1950). Psychological Aspects of Industrial Conflict II. Personnel Psychology, 3, 1-15.

Steers, R. M. & Mowday, R. T. (1981). Employee Turnover and Post-Decision Accommodation Processes. Research in Organizational Behavior, 3, 235-281.

Thissen, D. (1986). MULTILOG [Computer program]. Mooresville, IN: Scientific Software.

Thissen, D., Steinberg, L., Pyszczynski, T. & Greenberg, J. (1993). An Item Response Theory for Personality and Attitude Scales: Item Analysis Using Restricted Factor Analysis. Applied Psychological Measurement, 7(2), 211-226.

Uhrbrock, R. S. (1934). Attitude of 4430 Employees. Journal of Social Psychology, 5, 365-377.

Viteles, M. S. (1953). Motivation and Morale in Industry. Norton & Co; New York.

William, D. L. (1979). How to Develop and Conduct Successful Employee Attitude Surveys. Chicago, Illinois: Dartnell.

Yu, J. & Murphy, K. R. (1993). Modesty Bias in Self-Ratings of Performance: A test of the Cultural Relativity Hypothesis. Personnel Psychology, 46, 357-363.

Yukl, G. (1998). Leadership in Organizations (4^{th} Ed.). Englewood Cliffs, NJ: Prentice-Hall.

Appendix A

Table A1. Quality of Care and Service Dimension

QUALITY OF CARE AND SERVICE	
5.	**My commitment to quality of care and service**
6.	I am proud of the quality of care we provide to our patients.
7.	I am proud of the quality of service we provide to our patients and other customers.
8.	I would recommend ---- to my family and friends as a place to receive healthcare.
10.	**Others' commitment to quality of care and service**
11.	People in *my* department treat our patients and other customers well.
12.	People in *other* departments with whom I interact treat patients and other customers well.
13.	In general, people at ---- are competent in performing their jobs.
14.	---- is committed to continuously improving patients' and other customers' satisfaction with healthcare outcomes.
15.	My chief/supervisor is committed to improving customer satisfaction.
16.	**Quality Improvement Efforts**
17.	----'s efforts to improve the quality of care are resulting in improvements.
18.	----'s efforts to improve the quality of service are resulting in improvements.
19.	---- provides me with opportunities to learn quality concepts and methods at GHC.
20.	I am participating in efforts to improve quality.
21.	My chief/supervisor supports or participates in efforts to improve quality.
22.	My chief/supervisor implements the findings from efforts to improve quality.

Table A2. Management Dimension

MANAGEMENT	
26.	**Career Development**
27.	---- provides me with opportunities to upgrade the skills and abilities I need to do my work/practice.
28.	My career is progressing about as I hoped it would.
31.	**Communication**
32.	My chief/supervisor communicates *frequently* with me.
33.	My chief/supervisor communicates *effectively* with me.
34.	I can express my opinions and ideas honestly to my chief/supervisor.
35.	My chief/supervisor respects my values and opinions.
36.	My chief/supervisor is well-informed about the major problems that face me on the job or in my practice.
37.	I know where to go within ---- to get information that addresses my questions and concerns.
38.	I receive the information I need to do my work well in a timely manner.
39.	My chief/supervisor clearly communicates the expectations for my work unit/practice team.
40.	**Decision Making**
41.	People on our team have the authority to make *decisions* that affect our work/practice.
42.	I believe decisions at ---- are made at the appropriate levels.
43.	I am involved in making decisions that affect my work/practice.
44.	I would like to be more involved in making decisions that affect my work/practice.
45.	My chief/supervisor uses data to make decisions.
46.	**Performance Feedback**
47.	My chief/supervisor demonstrates concern about employees and their problems.
48.	My chief/supervisor appropriately clarifies to me what I need to do to get my work done.
49.	I know the basis on which I am evaluated.
50.	I get enough feedback about my performance to know if I am meeting my chief/supervisor's expectations.
51.	**Receptiveness to Innovation in the Workplace**
52.	Innovation is rewarded at ----.
53.	I am encouraged to try new ideas and express different points of view at ----.

Appendix B

Descriptive Statistics and Factor Analysis of Management Dimension

Table B1. Factor Analysis Results

Component	Eigenvalues	% of Variance	Cumulative %	Item	Component Loadings	
					1	2
1	9.853	46.919	46.919	Q33	.846	-.258
2	1.735	8.264	55.183	Q35	.843	-.214
3	.998	4.751	59.933	Q47	.801	-.232
4	.942	4.487	64.420	Q34	.790	-.243
5	.852	4.057	68.477	Q36	.788	-.197
6	.763	3.635	72.112	Q39	.784	-.233
7	.670	3.191	75.303	Q32	.780	-.246
8	.636	3.031	78.334	Q48	.750	-.247
9	.583	2.777	81.111	Q50	.738	-.236
10	.499	2.376	83.487	Q53	.686	.372
11	.447	2.127	85.614	Q49	.678	-.157
12	.418	1.990	87.604	Q43	.666	.362
13	.390	1.857	89.461	Q45	.645	-.137
14	.376	1.788	91.249	Q38	.625	.131
15	.354	1.684	92.933	Q41	.623	.381
16	.321	1.527	94.461	Q52	.587	.451
17	.307	1.464	95.924	Q42	.574	.417
18	.258	1.228	97.153	Q37	.545	.113
19	.242	1.150	98.303	Q28	.540	.363
20	.191	.910	99.212	Q27	.539	.323
21	.165	.788	100.000	Q44	-.293	

Table B2. Descriptive Statistics of Management Dimension for Manager Subgroup

	N	Min	Max	Mean	Std. Dev.
Q27	647	1.00	5.00	3.7728	.9778
Q28	642	1.00	5.00	3.3879	1.1147
Q32	636	1.00	5.00	3.6855	1.1344
Q33	637	1.00	5.00	3.7268	1.1007
Q34	636	1.00	5.00	4.0126	1.0408
Q35	623	1.00	5.00	4.0819	.9779
Q36	633	1.00	5.00	3.7536	1.0715
Q37	634	1.00	5.00	3.7871	.9238
Q38	635	1.00	5.00	3.4394	1.0206
Q39	636	1.00	5.00	3.4921	1.0817
Q41	637	1.00	5.00	3.4239	1.1489
Q42	630	1.00	5.00	2.6841	1.0833
Q43	637	1.00	5.00	3.6907	1.0048
Q44	631	1.00	5.00	3.8796	.9075
Q45	621	1.00	5.00	3.9823	.9494
Q47	640	1.00	5.00	3.9922	.9762
Q48	625	1.00	5.00	3.5344	1.0435
Q49	634	1.00	5.00	3.6420	1.0678
Q50	634	1.00	5.00	3.5647	1.0898
Q52	636	1.00	5.00	2.6887	1.0943
Q53	646	1.00	5.00	3.2724	1.1179
Dimension	640	1.20	5.00	3.5945	.6532
Dimension (no Q33)	640	1.30	5.00	3.6053	.6459

Table B3. Descriptive Statistics of Management Dimension for Non-manager Subgroup

	N	Min	Max	Mean	Std. Dev.
Q27	3537	1.00	5.00	3.3613	1.1082
Q28	3516	1.00	5.00	2.9406	1.1141
Q32	3496	1.00	5.00	3.2832	1.2599
Q33	3485	1.00	5.00	3.4565	1.2081
Q34	3456	1.00	5.00	3.7049	1.1622
Q35	3347	1.00	5.00	3.6271	1.1411
Q36	3435	1.00	5.00	3.4780	1.1966
Q37	3425	1.00	5.00	3.4070	1.0263
Q38	3484	1.00	5.00	3.3335	1.0407
Q39	3460	1.00	5.00	3.3708	1.1234
Q41	3442	1.00	5.00	2.9033	1.1920
Q42	3311	1.00	5.00	2.4086	1.0891
Q43	3468	1.00	5.00	2.8544	1.1518
Q44	3464	1.00	5.00	3.9824	.8227
Q45	2990	1.00	5.00	3.8157	.9486
Q47	3482	1.00	5.00	3.6904	1.1358
Q48	3453	1.00	5.00	3.3860	1.0498
Q49	3390	1.00	5.00	3.5088	1.0915
Q50	3440	1.00	5.00	3.3584	1.1604
Q52	3309	1.00	5.00	2.4455	1.0469
Q53	3502	1.00	5.00	2.8361	1.1284
Dimension	3533	1.10	5.00	3.2975	.7107
Dimension (no Q33)	3531	1.20	5.00	3.3001	.7005

Descriptive Statistics and Factor Analysis of Quality of Care and Service Dimension

Table B4. Factor Analysis Results

Component	Eigenvalues	% of Variance	Cumulative %	Item	Component Loadings		
					1	2	3
1	5.301	37.861	37.861	Q17	.700	-.144	-.329
2	1.947	13.904	51.765	Q8	.697	-.359	-.112
3	1.111	7.934	59.699	Q6	.693	-.385	2.475E-02
4	.955	6.824	66.523	Q18	.648	-8.7E-02	-.402
5	.778	5.554	72.077	Q7	.647	-.426	8.682E-02
6	.657	4.693	76.770	Q22	.635	.562	.161
7	.570	4.072	80.841	Q14	.622	-.403	-.172
8	.527	3.767	84.608	Q21	.614	.601	.143
9	.435	3.110	87.718	Q15	.604	.520	.147
10	.420	3.003	90.721	Q13	.589	-.153	.290
11	.404	2.885	93.606	Q19	.586	.341	-.277
12	.360	2.573	96.179	Q12	.547	-.222	.437
13	.317	2.266	98.445	Q11	.499	-4.37E-02	.496
14	.218	1.555	100.000	Q20	.485	.398	-.353

Table B5. Descriptive Statistics of Quality of Care and Service Dimension for the Nurse Staff Subgroup

	N	Min	Max	Mean	Std. Dev.
Q6	1081	1.00	5.00	3.5976	1.1482
Q7	1075	1.00	5.00	3.3349	1.1633
Q8	1076	1.00	5.00	3.3467	1.1378
Q11	1081	1.00	5.00	4.2969	.8129
Q12	1044	1.00	5.00	3.6533	.9455
Q13	1077	1.00	5.00	4.2943	.7650
Q14	1060	1.00	5.00	3.0849	1.2037
Q15	1024	1.00	5.00	3.8867	.9405
Q17	1036	1.00	5.00	2.5869	1.1221
Q18	1033	1.00	5.00	2.4908	1.1146
Q19	1067	1.00	5.00	3.1884	1.0499
Q20	1069	1.00	5.00	3.6978	.9105
Q21	979	1.00	5.00	3.8396	.9178
Q22	917	1.00	5.00	3.6238	.9842
Dimension	1085	1.40	5.00	3.4958	.6501
Dimension (no Q17)	1086	1.50	5.00	3.5619	.6307

Table B6. Descriptive Statistics of Management Dimension for the Nurse Staff Subgroup

	N	Min	Max	Mean	Std. Dev.
Q27	1076	1.00	5.00	3.3894	1.0911
Q28	1070	1.00	5.00	2.9383	1.1237
Q32	1061	1.00	5.00	3.1074	1.2723
Q33	1051	1.00	5.00	3.3987	1.2039
Q34	1043	1.00	5.00	3.5954	1.1866
Q35	989	1.00	5.00	3.5157	1.1589
Q36	1026	1.00	5.00	3.3460	1.2408
Q37	1036	1.00	5.00	3.3803	1.0569
Q38	1054	1.00	5.00	3.2970	1.0550
Q39	1044	1.00	5.00	3.3506	1.1094
Q41	1039	1.00	5.00	2.7680	1.2105
Q42	1017	1.00	5.00	2.1701	1.0441
Q43	1048	1.00	5.00	2.6613	1.1517
Q44	1052	1.00	5.00	4.0247	.8157
Q45	874	1.00	5.00	3.7906	.9335
Q47	1050	1.00	5.00	3.5724	1.1882
Q48	1040	1.00	5.00	3.2740	1.0804
Q49	1027	1.00	5.00	3.5219	1.0861
Q50	1037	1.00	5.00	3.3259	1.1448
Q52	1019	1.00	5.00	2.3121	1.0072
Q53	1058	1.00	5.00	2.6503	1.1295
Dimension	1070	1.20	5.00	3.2095	.7067

Table B7. Descriptive Statistics of Quality of Care and Service Dimension for Medical Staff Subgroup

	N	Min	Max	Mean	Std. Dev.
Q6	598	1.00	5.00	3.9900	.9615
Q7	600	1.00	5.00	3.5167	1.1278
Q8	598	1.00	5.00	3.7592	1.0283
Q11	596	1.00	5.00	4.2584	.8160
Q12	585	1.00	5.00	3.7436	.8292
Q13	593	1.00	5.00	4.2327	.7883
Q14	595	1.00	5.00	3.5983	1.0707
Q15	576	1.00	5.00	4.0243	.8482
Q17	576	1.00	5.00	3.2049	1.1158
Q18	575	1.00	5.00	2.8557	1.1469
Q19	586	1.00	5.00	3.5051	.9523
Q20	588	1.00	5.00	3.9337	.8176
Q21	560	1.00	5.00	3.9982	.8147
Q22	533	1.00	5.00	3.7223	.8979
Dimension	599	1.20	5.00	3.7371	.6194
Dimension (no Q17)	600	1.20	5.00	3.7785	.6029

Table B8. Descriptive Statistics of Management Dimension for Medical Staff Subgroup

	N	Min	Max	Mean	Std. Dev.
Q27	594	1.00	5.00	3.7609	.9311
Q28	593	1.00	5.00	3.1922	1.1301
Q32	584	1.00	5.00	3.4264	1.1632
Q33	581	1.00	5.00	3.5869	1.0882
Q34	579	1.00	5.00	3.9102	1.0063
Q35	569	1.00	5.00	3.8436	.9957
Q36	575	1.00	5.00	3.6835	1.0563
Q37	576	1.00	5.00	3.3576	1.0336
Q38	580	1.00	5.00	3.4103	.9696
Q39	576	1.00	5.00	3.4392	1.0466
Q41	583	1.00	5.00	2.6364	1.1657
Q42	566	1.00	5.00	2.2633	1.0372
Q43	585	1.00	5.00	2.9402	1.1519
Q44	587	1.00	5.00	4.0017	.8476
Q45	533	1.00	5.00	3.8499	.9278
Q47	580	1.00	5.00	3.7966	.9746
Q48	566	1.00	5.00	3.3958	.9698
Q49	569	1.00	5.00	3.6397	1.0063
Q50	576	1.00	5.00	3.5122	1.0713
Q52	568	1.00	5.00	2.4754	1.0386
Q53	591	1.00	5.00	2.8376	1.1353
Dimension	590	1.20	4.90	3.3800	.6240

Appendix C

Table C1. Non-Manager Item Parameter Estimates for 'Management' Dimension

	a-parameter	b1-parameter	b2-parameter	b3-parameter	b4-parameter
Item27	.683E+00	-.358E+01	.446E+00	.184E+01	.391E+01
Item28	.696E+00	-.469E+01	-.897E+00	.876E+00	.314E+01
Item32	.232E+01	-.138E+01	.855E-01	.585E+00	.165E+01
Item33	.306E+01	-.122E+01	.287E+00	.813E+00	.164E+01
Item34	.245E+01	-.981E+00	.635E+00	.112E+01	.191E+01
Item35	.299E+01	-.106E+01	.404E+00	.111E+01	.186E+01
Item36	.236E+01	-.129E+01	.298E+00	.872E+00	.182E+01
Item37	.106E+01	-.281E+01	.347E+00	.150E+01	.315E+01
Item38	.129E+01	-.259E+01	.180E+00	.117E+01	.280E+01
Item39	.223E+01	-.172E+01	.164E+00	.927E+00	.192E+01
Item41	.885E+00	-.377E+01	-.576E+00	.434E+00	.229E+01
Item42	.856E+00	-.501E+01	-.196E+01	-.409E+00	.161E+01
Item43	.121E+01	-.340E+01	-.603E+00	.332E+00	.188E+01
Item44	.226E+00	-.467E+01	.517E+01	.118E+02	.182E+02
Item45	.117E+01	-.158E+01	.899E+00	.229E+01	.348E+01
Item47	.244E+01	-.104E+01	.606E+00	.116E+01	.202E+01
Item48	.190E+01	-.194E+01	.124E+00	.113E+01	.223E+01
Item49	.137E+01	-.199E+01	.567E+00	.134E+01	.249E+01
Item50	.169E+01	-.182E+01	.229E+00	.910E+00	.207E+01
Item52	.829E+00	-.545E+01	-.221E+01	-.208E+00	.182E+01
Item53	.979E+00	-.392E+01	-.885E+00	.483E+00	.214E+01

Table C2. Manager Item Parameter Estimates for 'Management' Dimension

	a-parameter	b1-parameter	b2-parameter	b3-parameter	b4-parameter
Item27	.771E+00	-.227E+01	.116E+01	.258E+01	.464E+01
Item28	.852E+00	-.275E+01	.946E-02	.149E+01	.319E+01
Item32	.214E+01	-.117E+01	.403E+00	.931E+00	.201E+01
Item33	.312E+01	-.104E+01	.363E+00	.988E+00	.179E+01
Item34	.225E+01	-.680E+00	.817E+00	.145E+01	.210E+01
Item35	.274E+01	-.646E+00	.871E+00	.150E+01	.213E+01
Item36	.226E+01	-.113E+01	.412E+00	.119E+01	.202E+01
Item37	.895E+00	-.221E+01	.105E+01	.269E+01	.406E+01
Item38	.121E+01	-.255E+01	.172E+00	.126E+01	.285E+01
Item39	.260E+01	-.154E+01	.492E-01	.902E+00	.186E+01
Item41	.887E+00	-.268E+01	.338E+00	.129E+01	.296E+01
Item42	.907E+00	-.440E+01	-.150E+01	-.110E+00	.214E+01
Item43	.121E+01	-.194E+01	.718E+00	.172E+01	.294E+01
Item44	.214E+00	-.513E+01	.362E+01	.111E+02	.167E+02
Item45	.116E+01	-.120E+01	.109E+01	.232E+01	.342E+01
Item47	.225E+01	-.883E+00	.872E+00	.148E+01	.230E+01
Item48	.235E+01	-.162E+01	.139E+00	.970E+00	.208E+01
Item49	.179E+01	-.156E+01	.443E+00	.109E+01	.222E+01
Item50	.203E+01	-.156E+01	.233E+00	.989E+00	.199E+01
Item52	.848E+00	-.432E+01	-.178E+01	.221E+00	.201E+01
Item53	.107E+01	-.286E+01	-.743E-01	.931E+00	.255E+01

Table C3. Nurse Staff Item Parameter Estimates for 'Management' Dimension

	a-parameter	b1-parameter	b2-parameter	b3-parameter	b4-parameter
Item27	.762E+00	-.313E+01	.531E+00	.185E+01	.384E+01
Item28	.671E+00	-.486E+01	-.785E+00	.994E+00	.322E+01
Item32	.238E+01	-.145E+01	-.360E-01	.427E+00	.162E+01
Item33	.295E+01	-.128E+01	.341E+00	.873E+00	.166E+01
Item34	.241E+01	-.105E+01	.606E+00	.105E+01	.189E+01
Item35	.295E+01	-.113E+01	.375E+00	.106E+01	.190E+01
Item36	.255E+01	-.131E+01	.234E+00	.711E+00	.170E+01
Item37	.119E+01	-.256E+01	.422E+00	.141E+01	.279E+01
Item38	.127E+01	-.255E+01	.223E+00	.117E+01	.288E+01
Item39	.231E+01	-.168E+01	.222E+00	.940E+00	.204E+01
Item41	.844E+00	-.386E+01	-.774E+00	.150E+00	.221E+01
Item42	.823E+00	-.548E+01	-.255E+01	-.877E+00	.121E+01
Item43	.115E+01	-.362E+01	-.869E+00	.681E-01	.175E+01
Item44	.217E+00	-.442E+01	.615E+01	.129E+02	.188E+02
Item45	.114E+01	-.164E+01	.883E+00	.249E+01	.363E+01
Item47	.257E+01	-.101E+01	.521E+00	.104E+01	.194E+01
Item48	.194E+01	-.206E+01	.117E+00	.991E+00	.212E+01
Item49	.138E+01	-.196E+01	.776E+00	.144E+01	.255E+01
Item50	.168E+01	-.188E+01	.312E+00	.945E+00	.217E+01
Item52	.801E+00	-.564E+01	-.267E+01	-.509E+00	.176E+01
Item53	.953E+00	-.415E+01	-.117E+01	.174E+00	.196E+01

Table C4. Medical Staff Item Parameter Estimates for 'Management' Dimension

	a-parameter	b1-parameter	b2-parameter	b3-parameter	b4-parameter
Item27	.701E+00	-.271E+01	.147E+01	.308E+01	.519E+01
Item28	.632E+00	-.417E+01	-.208E+00	.136E+01	.385E+01
Item32	.238E+01	-.139E+01	.123E+00	.691E+00	.182E+01
Item33	.325E+01	-.123E+01	.258E+00	.931E+00	.183E+01
Item34	.251E+01	-.891E+00	.782E+00	.141E+01	.220E+01
Item35	.312E+01	-.955E+00	.536E+00	.144E+01	.209E+01
Item36	.228E+01	-.120E+01	.405E+00	.117E+01	.223E+01
Item37	.101E+01	-.301E+01	.125E+00	.139E+01	.306E+01
Item38	.101E+01	-.332E+01	.309E+00	.155E+01	.347E+01
Item39	.185E+01	-.190E+01	.912E-01	.117E+01	.206E+01
Item41	.668E+00	-.583E+01	-.155E+01	-.186E+00	.220E+01
Item42	.653E+00	-.778E+01	-.293E+01	-.108E+01	.160E+01
Item43	.107E+01	-.349E+01	-.553E+00	.397E+00	.211E+01
Item44	.222E+00	-.428E+01	.525E+01	.119E+02	.188E+02
Item45	.123E+01	-.162E+01	.981E+00	.217E+01	.320E+01
Item47	.226E+01	-.120E+01	.675E+00	.140E+01	.245E+01
Item48	.173E+01	-.218E+01	-.756E-01	.131E+01	.235E+01
Item49	.128E+01	-.202E+01	.695E+00	.152E+01	.288E+01
Item50	.139E+01	-.204E+01	.415E+00	.127E+01	.244E+01
Item52	.648E+00	-.632E+01	-.279E+01	-.274E+00	.226E+01
Item53	.892E+00	-.389E+01	-.113E+01	.479E+00	.211E+01

Table C5. Nurse Staff Item Parameter Estimates for 'Quality of Care and Service' Dimension

	a-parameter	b1-parameter	b2-parameter	b3-parameter	b4-parameter
Item6	.218E+01	-.107E+01	.490E+00	.917E+00	.226E+01
Item7	.225E+01	-.143E+01	.163E+00	.627E+00	.208E+01
Item8	.206E+01	-.149E+01	.726E-01	.804E+00	.204E+01
Item11	.713E+00	-.320E+00	.320E+01	.452E+01	.689E+01
Item12	.103E+01	-.215E+01	.838E+00	.196E+01	.431E+01
Item13	.484E+00	-.618E+00	.473E+01	.704E+01	.106E+02
Item14	.247E+01	-.169E+01	-.169E+00	.435E+00	.149E+01
Item15	.970E+00	-.151E+01	.148E+01	.272E+01	.397E+01
Item17	.267E+01	-.214E+01	-.863E+00	-.122E+00	.118E+01
Item18	.222E+01	-.241E+01	-.986E+00	-.226E+00	.111E+01
Item19	.102E+01	-.325E+01	-.143E+00	.110E+01	.295E+01
Item20	.695E+00	-.298E+01	.137E+01	.310E+01	.540E+01
Item21	.933E+00	-.174E+01	.134E+01	.278E+01	.429E+01
Item22	.911E+00	-.215E+01	.572E+00	.242E+01	.395E+01

Table C6. Medical Staff Item Parameter Estimates for 'Quality of Care and Service' Dimension

	a-parameter	b1-parameter	b2-parameter	b3-parameter	b4-parameter
Item6	.256E+01	-.691E+00	.983E+00	.150E+01	.272E+01
Item7	.234E+01	-.122E+01	.303E+00	.870E+00	.218E+01
Item8	.246E+01	-.922E+00	.490E+00	.137E+01	.240E+01
Item11	.117E+01	-.414E+00	.209E+01	.295E+01	.493E+01
Item12	.170E+01	-.170E+01	.783E+00	.186E+01	.360E+01
Item13	.101E+01	-.621E+00	.249E+01	.339E+01	.521E+01
Item14	.206E+01	-.118E+01	.279E+00	.117E+01	.250E+01
Item15	.736E+00	-.167E+01	.249E+01	.400E+01	.523E+01
Item17	.194E+01	-.180E+01	-.225E+00	.677E+00	.200E+01
Item18	.190E+01	-.210E+01	-.629E+00	.233E+00	.158E+01
Item19	.101E+01	-.268E+01	.399E+00	.202E+01	.354E+01
Item20	.680E+00	-.227E+01	.223E+01	.411E+01	.642E+01
Item21	.773E+00	-.180E+01	.215E+01	.400E+01	.538E+01
Item22	.842E+00	-.224E+01	.753E+00	.308E+01	.473E+01

Appendix D

NCDIF Results by Comparison

Management Dimension	Manager/ Non-manager	Doctors/Nurses (Management Dimension)	Quality of Care and Service	Doctors/Nurses (Quality of Care and Service Dimension)
Item27	.049 (.0000)	.077 (.0000)	Item6	.016 (.0000)
Item28	.049 (.0000)	.022 (.0000)	Item7	.028 (.0000)
Item32	.006 (.2221)	.005 (.0000)	Item8	.011 (.0000)
Item33	.030 (.0000)	.017 (.0000)	Item11	.066 (.0000)
Item34	.004 (.0000)	.007 (.0000)	Item12	.022 (.0000)
Item35	.009 (.0000)	.007 (.0000)	Item13	.072 (.0000)
Item36	.014 (.0000)	.019 (.0000)	Item14	.063 (.0000)
Item37	.049 (.0000)	.029 (.0000)	Item15	.014 (.2465)
Item38	.027 (.0000)	.005 (.0000)	Item17	.095 (.0000)
Item39	.097 (.0000)	.016 (.0000)	Item18	.005 (.0000)
Item41	.076 (.0000)	.064 (.0000)	Item19	.022 (.0000)
Item42	.001 (.0000)	.003 (.0000)	Item20	.018 (.0000)
Item43	.327 (.0000)	.015 (.0000)	Item21	.010 (.0072)
Item44	.023 (.0000)	.003 (.0000)	Item22	.007 (.0000)
Item45	.004 (.0000)	.020 (.0000)		
Item47	.007 (.0000)	.006 (.0000)		
Item48	.064 (.0000)	.006 (.0000)		
Item49	.070 (.0000)	.008 (.0000)		
Item50	.036 (.0000)	.001 (.0000)		
Item52	.001 (.0000)	.005 (.0000)		
Item53	.027 (.0000)	.003 (.0000)		

Carl Swander

104 Camelot Court
Blacksburg, VA 24060
(540) 961-3818
cswander@vt.edu

Education

1999 M.S. **Industrial/Organizational Psychology,** Virginia Polytechnic Institute and State University, Blacksburg, VA.
Thesis: *Assessing the Differential Functioning of Items and Tests of a Polytomous Employee Attitude Survey*

1997 B.S. **Psychology, minor in General Business,** Washington State University, Pullman, WA.

Related Experience

Consulting

Assistant Industrial/Organizational Psychologist
Ergometrics, Seattle, WA, 5/95-8/95, 5/96-8/96, 5/97-8/97, 5/98-8/98.
Responsible for statistical analysis and customer relations. Helped Industrial/Organizational Psychologists with projects involving job analysis and selection procedures. Analyzed and organized data pertaining to employee satisfaction, validation, and 360 performance appraisals.

Teaching

Undergraduate Advisor
Virginia Tech, Blacksburg, VA 8/98 – 5/99.
Advised approximately 400 undergraduate students for a major in Interdisciplinary Studies. Designed, implemented, and analyzed surveys to assess student satisfaction.

Graduate Teaching Assistant
Virginia Tech, Blacksburg, VA 8/97 – 5/98.
Currently teaching 3 sections with approximately 30 students in each class. Responsible for organizing lectures, quizzes and grading

Research Experience

Thesis Research
Robert J. Harvey (Chair). Analyzed and evaluate attitude survey data using item response theory and differential item functioning methods for polytomous data.

Independent Research
Robert J. Harvey (Chair). Designed and implemented a computerized adaptive version of a well known critical thinking appraisal.

Research Assistant
Washington State University, 1/95-5/95, 8/96-12/96
Conducted group decision-making experiments with introductory psychology students.