ED 457 183 TM 033 279

AUTHOR Wang, Ning; Wiser, Randall F.; Newman, Larry S.

TITLE Examining Reliability and Validity of Job Analysis Survey Data.

PUB DATE 1999-04-00

NOTE 49p.; An earlier version of this paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 20-22, 1999).

PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Generalizability Theory; Goodness of Fit; Item Response Theory; *Job Analysis; Rating Scales; *Reliability; *Sample Size; Surveys; *Validity

IDENTIFIERS FACETS Model; Rasch Model

ABSTRACT

Job analysis has played a fundamental role in developing and validating licensure and certification examinations, but research on what constitutes reliable and valid job analysis data is lacking. This paper examines the reliability and validity of job analysis survey results. Generalizability theory and the multi-facet Rasch item response theory (IRT) model (FACETS) are applied to investigate consistency and generalizability in task importance measures, suggest reliable sample size, justify the number and use of rating scales, and detect possible rating errors. By using random samples from job analysis data for two professions with divergent job activities, the study finds that a representative sample as small as 400 respondents produced reliable estimates of task importance to the same degree of generalizability as obtained from a larger sample of job analysis respondents. Analyses of rating scales suggest that the effectiveness of using differing numbers and types of rating scales depends on the nature of a profession. Limited rating ranges and fatigue effects are two types of erratic ratings identified in this study. Results indicate that FACETS' indices, such as rater severity, as well as infit and outfit statistics, are efficient and precise in detecting those rating errors. Appendixes contain charts of task importance measures in Rasch logits with transformed percentage weights for combinations of rating scales and data. (Contains 9 tables and 21 references.) (SLD)

Examining Reliability and Validity of Job Analysis Survey Data

Ning Wang

Randall F. Wiser

Larry S. Newman

Assessment Systems, Inc.
Three Bala Plaza West, Suite 300
Bala Cynwyd, PA 19004

Running Head: Examination of Job Analysis Survey Data

An earlier version of this paper was presented in the symposium "Establishing Validity of Licensure and Certification Examination: Job Analysis Methodologies and Developing Test Specifications" at the 1999 annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Corresponding concerning this article should be addressed to Ning Wang, Assessment Systems, Inc., Three Bala Plaza West, Suite 300, Bala Cynwyd, PA 19004.

Abstract

Historically, job analysis has played a fundamental role for developing and validating licensure and certification examinations. Still, research on what constitutes reliable and valid job analysis data is lacking. Consequently, few guidelines exist for collection and use of job analysis data in practice. This paper examines the reliability and validity of job analysis survey results. Generalizability theory and the multi-facet Rasch IRT model (FACETS) are applied to investigate consistency and generalizability in task importance measures, to suggest reliable sample size, to justify the number and use of rating scales, and to detect possible rating errors. By using random samples from job analysis data for two professions with divergent job activities, this study finds that a representative sample as small as 400 respondents produces reliable estimates of task importance to the same degree of generalizability as obtained from a larger sample of job analysis respondents. Analyses of rating scales suggest that the effectiveness of using differing numbers and types of rating scales depends on the nature of a profession. Limited rating ranges and fatigue effect are two types of erratic ratings identified in this study. Results indicate that FACETS' indices, such as rater severity as well as infit and outfit statistics, are efficient and precise in detecting those rating errors.

Examining Reliability and Validity of Job Analysis Survey Data

Examinations used for licensure and certification are designed to assess professional competence. According to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), validation of these examinations depends mainly on content-related evidence, with job analysis providing the primary basis for defining the test content domain. Often, such a job analysis is conducted on the work performed by people in a profession or occupation to document the tasks that are essential to practice (AERA, APA, NCME, 1999; Kane, 1982, 1986, 1997; Mehren, 1997; Raymond, 1995). To serve this purpose, a task survey questionnaire is commonly administered to practicing professionals; through the survey, relevant and important tasks that constitute job performance in the profession are rated on the basis of one or more rating scales (Knapp & Knapp, 1995). To adequately represent the major job characteristics, multiple rating scales are commonly selected to reflect separate aspects of the tasks, such as frequency of performance, criticality to public protection, and necessity at time of initial licensure. After collecting the survey data, task ratings are analyzed and a numerical measure of importance is computed for each task. Detailed test specifications are then developed using these task importance measures. The goal of such a job analysis is to obtain reliable and valid task measures for defining the test content domain.

Although job analysis has played a fundamental role for developing and validating licensure and certification examinations, issues regarding the reliability and validity of a job analysis result have scarcely been addressed. Significant gaps abound in the job analysis research base; consequently, few guidelines exist for collection and use of job analysis data in practice (Harvey, 1991; Nelson, 1994). The best type of rating scales to use, optimal and minimal number of rating scales, adequate sample size, and treatment of low response rates in task surveys are just some of the issues remaining to be investigated in job analysis.

It is well recognized that how rating scales are selected and survey data are collected have a significant impact on the interpretation and generalizability of the job analysis result (Harvey, 1991; Nelson, 1994). In practice, however, the choice of task rating scales has usually been guided by historic precedents, with some consideration given to minimizing possible overlap because of too many ratings per task, and to reducing survey takers' fatigue (Knapp & Knapp, 1995). Few studies have been conducted to provide a basis for justifying the adequacy and effectiveness in selecting and using job analysis rating scales (Sanchez & Levin, 1989; Sanchez & Fraser, 1992).

Survey sample representativeness, sufficient sample size, and low response rates are related issues for a job analysis. Traditionally, to ensure representativeness of survey respondents, job analysts distributed surveys to a large, or even unduly large number of practicing professionals. It is common for relatively low response rates to occur, especially in the areas of licensure and certification. It is a particular concern that low response rates would introduce systematic errors into job analysis data, thereby, reducing the validity of job analysis results. In general, however, there are limited investigations that consider what minimal sample size is required to assure a reliable and representative job analysis result.

Job analysis task surveys usually employ Likert-type rating scales. Even though these types of scales are widely used, there are problems associated with their interpretation due to rater errors, such as rater severity, central tendency, and restriction of range (Saal, Downey, & Lahey, 1980; Zegers, 1991). There has been limited investigation into possible approaches for detecting such rating errors to ensure reliable and valid job analysis data collection, use, and interpretation.

The purpose of this paper is to examine the reliability and validity of job analysis survey results by applying various measurement techniques, such as Generalizability theory, and the multi-facet Rasch IRT model. With the goal of providing evidence for valid and reliable job analyses, this study investigates whether survey ratings are consistent across different samples of raters, given the samples are representative of the survey population. The study also explores appropriate sample size as well as the minimal number of rating scales required to obtain reliable and valid job analysis results. In addition, this study attempts to provide a forum for discussing possible procedures for detecting rating errors that may occur in job analysis survey data. Through this investigation, the study explores a process for ensuring valid use and interpretation of job analysis data.

Facets of Job Analysis Survey Data

The objective in job analysis is to obtain measures of relative task importance. The meaning of these task importance measures provides the basis for assessing the reliability and validity of job analysis results (Messick, 1989); therefore, to examine reliability and validity of job analysis results, evidence needs to be collected on how the meaning of the task measures are derived. Consequently, it is useful to breakdown job analysis survey data into facets, so the influence of each facet on the task importance measures can be examined. If individual facets of task importance measures are valid and reliable, then evidence exists for the valid and reliable meaning of the task measures as a whole.

In a job analysis task survey, there are at least three major facets: Task Measure, Rater, and Rating Scale. The first facet, *Task Measure*, quantifies relative importance of tasks performed in the profession. It encompasses different aspects of a task, such as frequency of performance, criticality (i.e., importance for public protection), and need at entry-level. It is expected that tasks will vary in their importance measures based on the ratings of these perspectives. The goal of job analysis is to validly identify and reliably differentiate these tasks' importance measures.

The second facet is *Raters*, the job analysis task survey respondents. Raters' knowledge and experience in the profession are essential in determining the task measures. Respondents' unique reactions to the survey as well as their personal characteristics can also influence task ratings. Some raters may provide consistently low ratings across tasks, while others tend to rate tasks higher. How to distinguish between true diversity of ratings and erratic raters (e.g., severe or lenient raters, halo effect, limited ranges, etc.) is important for ensuring reliable and replicable job analysis results, as well as valid data interpretation.

The third facet, *Rating Scales*, represents separate aspects of the tasks. In this study, the rating scales include frequency (how often is a task performed), criticality (how important is a task for public protection), and need at entry-level (is this a task that someone must do when first licensed or certified). How each task is rated on these scales provide useful information for the meaning of task importance measures. In job analysis practice, it has been debated at length about the effectiveness of different rating scales. The necessity of using multiple rating scales, the best kind of rating scales to use, and the optimal number of rating scales to use, are issues of frequent concern for job analysts. Through analyzing data from job analyses for different professions, this study attempts to provide empirical evidence about the necessity of the scales.

The combination of these facets forms the job analysis survey structure. When task importance measures derived from this structure differs from sample to sample, the job analysis results are not replicable, generalizable, or valid. To ensure a reliable and valid job analysis, information related to each component of the structure should be carefully examined. Possible improvements suggested from the examination of previous job analyses in similar professions should be undertaken for future job analyses.

Methods

Job Analysis Survey Data

Two job analyses conducted by Assessment Systems, Inc. (ASI) provide the survey data used in this paper. *A National Analysis of the Occupational Tasks and Activities of Real Estate Professionals* was conducted in 1998 for the real estate licensing program of ASI. The survey for this job analysis was designed to identify tasks and activities that were most frequently performed, most critical for public protection, and most essential at entry level into the profession. Eighty-three tasks and activities compiled by a national committee of established real estate professionals and subject matter experts were rated on scales of frequency of performance, criticality for public protection, and need at time of licensure. The frequency scale was coded: 0=Never, 1=Rarely, 2=Sometimes, 3=Often. Criticality was coded: 0=Not Important, 1=Slightly Important, 2=Moderately Important, 3=Extremely Important. The need scale was coded: 0=Not required at all, 1=Not required at entry, 2=Required at entry, 3=Required at entry and further developed. Subject matter experts eliminated 16 tasks as unimportant after the survey had been completed, leaving 67 tasks that were ultimately used both in the job analysis and in this study. Both major groups of real estate professionals, sales and brokers, were sampled using the same survey. Nine sample regions were defined and targeted for the United States to avoid state specific variations in response rate, and to maintain a balanced return from all regions.

Demographic data used from the job analysis survey included information on job description, license type, gender, years of practice, and area of specialty. 16,351 surveys were mailed, and results of the job analysis were based on 1,420 respondents.

The *Job Analysis of Touch Therapies Practitioners* conducted in 1997 by ASI for the National Certification Board for Therapeutic Massage and Bodywork (NCBTMB) provide a second set of survey data to analyze. The purpose for this job analysis was to validate content for a new entry-level credentialling examination. The survey was composed of 342 tasks, knowledge statements, and professional standards that were to be rated for relevance to the practice of touch therapies. Focus groups of subject matter experts, representing the various types of touch therapies, reviewed the previous job analysis and made recommendations to the job analysis task force for additions, changes, and deletions of tasks, activities, and knowledge statements to be included in the survey. The final survey was approved by the NCBTMB. Respondents rated the elements of the survey for frequency (how often a task or activity was performed in practice), competence (how important the task was to practice), and entry level (how necessary was the task, activity, or standard for entry-level performance). These rating scales were coded as follows: Frequency (0=Never, 1=Seldom, 2=Often, 3=Almost always); Competence (0=Not necessary, 1=Slightly necessary, 2=Moderately necessary, 3=Very necessary); Entry Level (0=Not relevant, 1=Necessary, 2=Not necessary). From a mailing list of 72,368 people representing ten different organizations and credentialling groups under the NCBTMB, a stratified random sample of 20 percent from each group was selected to receive the survey. From the 14,917 surveys mailed, the job analysis was performed on 1,903 respondents.

Generalizability Analysis

In Generalizability theory (G-theory), a behavioral measurement is considered a sample from a universe of admissible observations described by one or more facets. The universe of observations for a measurement includes all the facets of the observation that can vary without altering the reliability or acceptability of the measurement. For example, if the choice of rating scales might effect task importance measures, then an adequate sample of scales must be included in the measurement procedure. Ideally, we would like to know if a task importance measure (the universe score) over all combinations of facets and conditions (i.e., all possible raters, all possible scales, and all possible occasions) reflects competent performance in a profession. By establishing a variance component for the universe score and variance components for the other facets that are inherent in an observed score, G-theory allows a true score (universe score) variance to be separated from error variances of a given measurement.

For job analysis, task importance measures are the universe scores to be estimated. Variability of task measures due to the design facets (rating scales and raters) can be estimated via G-theory so that variances due to each facet are identified. Therefore, errors due to unexpected factors possibly can be detected and adjusted or eliminated. Judgement of whether the variance due to each facet is expected or unexpected helps the investigation of job analysis rating validity.

In this study, a series of G-studies are conducted to examine reliability and validity of job analysis survey data. These include a two-facet (Task × Rater × Scale) random effects design for Real Estate and Body Therapy job analysis data, and studies on various subsets of the rating data. Variance components due to main effect and two-way interactions are examined. In each analysis, generalizability and dependability coefficients are calculated for decision studies to assess the reliability of task importance measures for different sample sizes and rating scales.

Multi-Facet Rasch Analysis

The multi-facet Rasch rating model [FACETS] (Linacre, 1989; Wright & Master, 1982) is also used in analyzing the job analysis survey data. The basic Rasch model is a one-parameter IRT logistic model for dichotomously scored responses, while FACETS extends this model to ordinal rating data. FACETS models the probability that a rater assigns a rating in category j rather than a rating in category j-1. In analyzing job analysis rating data, the probability (P_{nilx}) of rater n rating task i with a rating x (x ranging from 0 to m) on scale l (l = 1, 2, or 3 in this study) is modeled as

$$P_{nilx} = \{\exp\sum_{j=0}^{x}[B_n - (D_i + F_j + S_l)]\} \div \{\sum_{k=0}^{m}\exp\sum_{j=0}^{k}[B_n - (D_i + F_j + S_l)]\} ,$$

where B_n is the rater's propensity towards higher ratings (rater severity), D_i is the task's lack of propensity to obtain high ratings (task difficulty or measure), S_l represents the measure of scale l (scale difficulty), Fj is the marginal lack of propensity to obtain the *jth* rating on the rating scale l (difficulty being rated in category j rather than category j-1).

FACETS is a unidimensional model with a single proficiency parameter for the objective of measurement (task measure in job analysis), and a collection of other facets. In a job analysis, these other facets can be viewed as a series of rating opportunities that yield multiple ratings for each task. FACETS is appropriate if the intent is to sum ratings from the rating opportunities provided by the separate facets, to produce a total measure for the objective (Engelhard, 1994).

Through FACETS analysis, measures in a log-linear scale (units of logits) for each facet of task, rater, and rating scale are estimated separately. The ordering of task measures, raters, and rating scales on the logit scale provide a frame of reference for understanding relationships of the facets in the job analysis data. By maintaining the optimal property of IRT logistic models, FACETS makes it possible to separately observe estimated task measures from highest to lowest, estimated rater severity from most to least severe, and estimated scale difficulty from most to least difficult. Therefore, task measures can be obtained in terms of their relative importance. Also, outliers in terms of rater severity can be identified and further investigated. In addition, goodness-of-fit statistics are also estimated for individuals from a perspective of each facet, so that further diagnostics can be conducted to examine the quality of the rating data.

In this study, FACETS analyses are conducted for both the Real Estate and Body Therapy job analysis survey data, and for various subsets of the rating data. Task measures obtained from different rater samples and rating scales of the same survey are compared to examine the consistency of task measures. In each analysis, diagnostic information, such as goodness-of-fit statistics and rater severity are examined to detect possible rating errors.

Results and Discussions

Examination of Task Measure Consistency and Generalizability

Consistency and generalizability of the task importance measures are examined using Task by Rater by Scale ($t \times r \times s$) random effects generalizability and decision studies, as well as FACETS analyses. Through G-studies on all of the job analysis data for both Real Estate and Body Therapy, and on various subsets of the same job analysis data, variability due to each facet and their two-way interactions are compared. The $t \times r \times s$ design for each data set is also used to examine the extent to which generalizations of the task ratings from the selected sample of raters and scales to the larger domain of job activities in the profession are valid. To allow statistical tests on rank distributions of identical task measures obtained from different samples, FACETS analyses are conducted on all of the Real Estate job analysis data and on various subsets of the data.

Generalizability studies for Real Estate job analysis. To examine task measure consistency, 1,420 raters from the Real Estate job analysis data are divided into three random groups. The first random group consists of 472 raters, the second group has 452 raters, and the third random group consists of 496 raters. A series of three-way t × r × s ANOVAs are conducted on the entire data set, the three random groups, and a complete data set, which consist of 457 raters who responded to all 67 tasks on each of the three rating scales. Table 1 provides the random effects ANOVA estimates from the generalizability studies for the five data sets.

Insert Table 1 about here

As can be seen in Table 1, the results are similar for the five analyses. Across the five data sets, the variability due to tasks account for a large percentage (≈20%) of the total variance, whereas the variability due to raters account for approximately 10% of the total variance, and the variability due to scales account for the least amount (≈2%) of the total variance. The variance component for t × r, which represents the differential rating of raters across tasks, account for the largest percentage (≈30%, except for the error term) of the total variability across the four analyses. Due to insufficient computing memory, the variance component for t × r for the entire data set can not be estimated. The t × s component, which accounts for the differential rating of tasks across scales, is relatively small (<5%). The variance component for the r × s interaction, which represents the differential rating of raters across scales, also accounts for a small percentage of the total variability (≈8%).

For the decision studies, a random effects design is used. The decision studies incorporate 1, 2, or 3 scales and either 400 or 1,400 raters. 400 raters are selected to reflect the smallest sample size used in the analyses of this study, and 1,400 raters are chosen to reflect the sample size used for the Real Estate job analysis. A different number of scales are used to examine how much the generalizability and dependability coefficients are improved by using more scales.

Table 2 shows the generalizability (ρ^2) and dependability (ϕ) coefficients for decision studies on three random groups and the complete data set. The generalizability coefficients are for relative decisions in which the reliability of the rank order of task measures is of interest. The dependability coefficients are for absolute decisions in which the reliability of the absolute level of task measures is of interest.

Insert Table 2 about here

It can be seen in Table 2, that the reliability of task ratings in terms of either rank orders or absolute values is relatively high across all four data sets, even with a sample size of 400 raters, rating on only one scale. Results from both Tables 1 and 2 indicate task ratings are very stable across different samples, variances from using different rating scales are very small, and increasing the number of scales from one to three does not greatly increase the reliability of the task ratings. In addition, when two rating scales are used, reliability coefficients are mostly greater than 0.90.

FACETS analyses for Real Estate job analysis data. To further investigate if task measures are consistent across different samples, FACETS analyses are conducted for six data sets (i.e., three random groups, a complete data set, the total Real Estate job analysis data, and rating data consisting only of real estate brokers). The three random groups and complete data set are similarly representative of real estate professionals. The results from these groups should be consistent, if the use of a smaller sample size has no effect on the job analysis results. The rating data for brokers is included as a group, since brokers only represent one sub-population of the real estate professionals. A difference in ratings of this subgroup from the total group is expected, thereby providing evidence of discriminant validity for the job analysis results.

To effectively communicate the results across analyses, the 67 task measures obtained from each of the six FACETS analyses are transformed into percentage weights. Task measures in logits and their transformed percentage weights for each of the six data sets are provided in Appendix A.

To determine if the same rank order of task weights is obtained from different samples, a correlation of the task weights between each of the five data sets (i.e., the three random groups, the complete data set, and the brokers) and the total job analysis data set is calculated. The Wilcoxon non-parametric signed ranks test for two related samples is also performed for each pair of data sets to examine whether the pairs of task weights have the same rank distributions. There are 66 degrees of freedom for each test (the number of tasks minus 1). The correlation coefficients and results from the statistical tests are shown in Table 3.

Insert Table 3 about here

Results in Table 3 indicate no significant difference in task measures obtained from random sub-groups of the job analysis data, or from the data set of complete responses, when compared to the total job analysis data. These results further confirm the G-studies' finding that consistent task measures can be obtained with a representative sample size as small as 400 raters. The result from the broker subgroup of the real estate professionals significantly differed from the total group of professionals ($p<0.05$). Because a single subgroup of brokers can not represent the entire population of the profession, this result provides one type of evidence for discriminant validity of the job analysis.

Generalizability studies for Body Therapy job analysis data. To examine the influence of sample size on the consistency of job analysis task survey results, another job analysis data set from a totally different profession, Body Therapy, is analyzed using the same design of generalizability studies as for the Real Estate job analysis data. Of the 1,903 respondents to the task survey, 1,046 raters provide complete responses to each of 342 tasks on the three rating scales. Three generalizability studies of the Task × Rater × Scale random effects ANOVA design are conducted based on these 1,046 respondents. The first study includes ratings from all 1,046 raters. The second study analyzes task ratings from a randomly selected set of 521 raters from the 1,046 raters (Rgrp1). The third study analyzes task ratings from another independent random set of 461 raters out of the 1,046 raters (Rgrp2). Table 4 provides the variance component estimates from the generalizability studies on the three data sets.

Insert Table 4 about here

Results in Table 4 confirm the findings from the Real Estate data analyses. That is, the distributions of the variability due to each component are stable across the three analyses, even though the one sample size is as small as 421 raters. For this job analysis, the variability due to tasks account for about 25% of the total variance, the variability due to raters account for a small percentage of the total variance (≈6%), and the variability due to scales account for approximately 13% of the total variance. The variance component due to t × r account for the second largest percentage of the total variability (≈24%), whereas the variance component due to t × s account for about 10% of the total variation. Variability due to r × s account for the smallest percentage of the total variance (≈3%). Since the variance components due to raters is small and the variance components due to scales is relatively large, the interpretation is that raters are using the scales in the same fashion. The use of the scales across raters is even more stable in the Body Therapist job analysis than in the Real Estate analysis, although the variability due to r × s in the Real Estate job analysis data is also relatively small.

The variance due to raters is smaller in the Body Therapist job analysis than in the Real Estate job analysis. This fact can be related to the different nature of the two professions. The job activities are more likely to be technique oriented for the body therapists, whereas real estate professionals are working in broad geographic areas, where people's socio-economic status, educational levels, and other demographic background are quite varied. The more socially oriented nature of the real estate profession is a possible factor contributing to a larger variability in the rater effect. It is also interesting to note that the variance component due to scales is much larger for the body therapy profession, in comparison to the almost negligible variance component from scales in the real estate job analysis. Again, a possible cause may be the different nature of the professions. For job performance of body therapists, public protection should be more emphasized than in real estate. Therefore, while a task may not be frequently performed, it should be weighted more heavily if it is important for public protection and needed for entry level of licensing. As a result, to accommodate different perspectives of a task in body therapy, more scales are needed for their job analysis, as compared to real estate.

For decision studies, a random effects design is used. The decision studies incorporated 1, 2, or 3 scales and either 400, 1,000, or 1,900 raters. 400 raters are selected to reflect the smallest sample size used in the analyses of this study. 1,000 raters are chosen to reflect the sample size from which complete responses to all 342 tasks are collected on each of the three rating scales. 1,900 raters are selected to reflect the sample size used in all of the Body Therapy job analysis. Different numbers of scales are used to examine how much the reliability is improved by using more rating scales.

Table 5 shows the generalizability (ρ^2) and dependability (ϕ) coefficients for decision studies of the two random groups and the complete data set. It can be seen in Table 5, that the reliability of task ratings in terms of either rank orders or absolute values is very stable across the three analyses. Table 5 further confirms the finding from the real estate data analysis, that increasing sample size from 400 raters to 1,000 or 1,900 does not improve the reliability of the job analysis results. Unlike the Real Estate job analysis, in the Body Therapy job analysis, increasing the number of rating scales greatly increases the reliability of the results, in terms of both rank orders and absolute values of task measures.

Insert Table 5 about here

Evaluation of Scale Necessity

Generalizability studies for Real Estate job analysis. To evaluate the necessity of scales in the Real Estate job analysis, data from each combination of two rating scales are analyzed using data from raters who responded to all 67 tasks. Data from 486 such raters is analyzed for the combination of frequency and criticality scales. Data from 523 of these respondents are used for the combination of frequency and need-at-entry scales. For the combination of criticality and need-at-entry scales, data from 470 complete raters are analyzed. Three-way t × r × s ANOVAs are performed for the combination of two scales. Table 6 provides the estimates of variance components resulting from the three-way random effects ANOVAs.

Insert Table 6 about here

Even though the absolute percentages of variance components due to each effect are not as similar as those found in Table 1, a similar trend in the variability distributions still can be observed across the three analyses. The variability due to scales is negligible. In particular, almost no variation is found between frequency and need-at-entry scales. The variance component due to the t × s interaction, which represents differential rating of tasks across scales, is also very small. Again, the variability due to the r × s interaction, which represents differential rating of raters across scales, is the next smallest component. The rater component accounts for about 10% of the total variance. The variance component due to tasks accounts for a relatively large percentage of the total variation (18% to 28%) depending on the combinations. Except the error term, the t × r interaction, representing differential rating of raters across tasks, accounts for the largest percentage of the total variance (25% to 30%).

For the decision studies, a random effects design is used. The decision studies incorporate 400 raters and 1, 2, or 3 scales. 400 raters are used because results from the examination of task measure consistency indicate that a sample size of 400 provided highly reliable results. Different numbers of scales are selected to examine how much the generalizability and dependability coefficients are improved by using more scales. Table 7 shows the generalizability (ρ^2) and dependability (ϕ) coefficients for the decision studies.

Insert Table 7 about here

Findings from Tables 6 and 7 indicate that three rating scales may not be necessary for obtaining reliable job analysis results for the real estate profession. Two scales such as criticality and need may be sufficient, if the survey questionnaire is carefully constructed.

FACETS analyses for Real Estate job analysis data. To further investigate if three scales are necessary, FACETS analyses are conducted on rating data from each single scale and each combination of two scales. To effectively communicate results across analyses, the 67 task measures obtained from each of the six FACETS analyses are transformed into percentage weights. For each of the six data sets, task measures in logits and their transformed percentage weights are provided in Appendix B.

To determine if the same rank order of task weights is obtained from different data sets, a correlation is done between the task weights of each of the six data sets (i.e., three single rating scales and three combinations of two rating scales) and the task weights obtained from the total job analysis data set. The Wilcoxon non-parametric signed ranks test for two related samples is conducted for each pair of data sets to examine whether each set of task weights has the same rank distributions as those obtained from the entire job analysis data set. There are 66 degrees of freedom for each test (the number of tasks minus 1). The correlation coefficients and the results from the statistical tests are shown in Table 8.

Insert Table 8 about here

Results in Table 8 indicate there is no significant difference between task weights obtained from subsets of the rating data and task weights obtained from the all of the job analysis data. These results further confirm that three rating scales are not necessary for a job analysis in this profession.

Detection of Rating Errors

Job analysis task surveys usually employ Likert-type rating scales. Given that diversity of responses to a certain degree is desired to reflect opinions obtained from professionals with different work experiences and job activities, it is necessary to identify erratic ratings to ensure that consistent and valid task measures are obtained. The errors in interpreting Likert-type rating scales in the context of performance assessment are commonly identified as rater severity, central tendency, halo effect, and restriction of range (Saal, Downey, & Lahey, 1980; Zegers, 1991). Other rater effects such as fatigue can also threaten the validity of job analysis results (Knapp & Knapp, 1995). Some questions then, in job analysis, are what kind of rating errors can occur and should be identified in the survey data? For what circumstances do those errors effect the accuracy of task measures and data interpretation? If errors exist, how should the job analysis and the subsequent analysis of rating data be designed to eliminate those errors?

In this study, rater errors are detected using indices obtained from FACETS analyses, such as rater severity, outfit, and infit statistics. Descriptive statistics such as means, variances, and frequencies of ratings are also examined to detect rater errors. Using real job analysis data for Real Estate and Body Therapy, this study attempts to provide a forum to discuss possible answers to these questions, and to provide suggestions for job analysis practices.

Limited rating range. According to our experience, the most commonly committed rating error in a job analysis survey is that raters restricted their ratings on all tasks to just one category on the rating scales. Since the purpose of a valid and effective job analysis is to identify relative importance of the tasks, failure to distinguish relevant rankings among the tasks is considered to be an erratic rating. Restricted ratings for the Real Estate job analysis data can be one of the extreme categories (0 or 3), or a middle category on the scales (1 or 2). Several indices from FACETS analysis are promising for identifying this type of rater error.

One index from FACETS analysis used for investigating limited rating range is *Rater Severity*. A rater severity value in logits is obtained for each rater through analyzing ratings on all three scales in all of the Real Estate job analysis data. The range of rater severity values for the 1,420 raters in this analysis is from –5.49 to 4.28. Most of the values fall between –2.00 and 2.00. For raters with relatively low logit values and relatively high logit values, means, variances, and frequencies of their ratings across tasks are calculated for each of the three rating scales. The results indicate that raters with a severity value lower than –1.00 give a rating of 3 on at least two rating scales for all or most of the tasks. Eighty-four raters with this kind of erratic rating are identified out of the 1,420 people. Raters with a severity value higher than 2.50 provide a rating of 0 on at least two rating scales for all or most of the tasks. Three people with this kind of erratic rating are identified out of the 1,420 real estate job analysis raters.

FACETS analysis using the 1,420 raters is also conducted for ratings on each scale. The rater severity ranges from –4.82 to 3.63 for the frequency scale, -4.48 to 5.17 for the importance scale, and –4.75 to 6.79 for the need scale. Examination of relatively lenient and severe raters yields similar results to those found from FACETS analysis using ratings on all three scales. Raters with higher severity values give a rating of 0 on the corresponding scale for all or most of the tasks, and raters with lower severity values give a rating of 3 on the respective scale for all or most of the tasks.

The second index from FACETS analysis used for investigating limited rating range is *Fit Statistics*, including both *Outfit* and *Infit* statistics. The infit statistic is an information weighted mean-square residual difference between observed and expected values, which focuses on the accumulation of central, inlying, deviations from expectation. The outfit statistic is the usual unweighted mean-square residual, which is particularly sensitive to outlying deviations from expectation. The expected value for the mean-square is 1.0 with a range from 0 to infinity. The region for acceptable fit is usually recommended to be greater than 0.6 and less than 1.5 (Linacre, 1989; Lunz et al, 1990; Stone, 1997).

As expected, the outfit statistics from FACETS analyses using ratings on all three scales and on each rating scale individually, indicate that raters with muted outfit values (as low as 0.2 or 0.3) tend to give a rating of 3 for all or most of the tasks. Raters with noisy outfit values (higher than 1.5) tend to give a rating of 0 across the tasks. As expected for the infit statistics, the results show that raters with muted infit values tend to give a rating of 2 (one of the middle categories) across the tasks, and people with noisy infit values tend to rate on one extreme category for one scale(s) and on another extreme category for the other scale(s).

Fatigue effect. Body Therapy job analysis data are analyzed to investigate the fatigue effect in rating job analysis tasks. The 342 tasks are divided into three blocks according to the administered tasks' sequence number in the survey. The first block consists of the first 114 tasks in the survey, the second block contains the second 114 tasks, and the third block consists of the last 114 tasks in the survey. Frequencies of each rater's ratings on each rating scale are calculated separately for each block and for all tasks. Table 9 provides the number of raters who rate only one category on each scale within each block, and who rate only one category for all 342 tasks (i.e., a rater's rating variance for those tasks on the scale is 0).

Insert Table 9 about here

From Table 9, it can be seen that as more tasks are being rated, more raters tended to select just one rating category on the competence and need-at-entry scales. Also, more raters select only one category on the second and the third scale (i.e., competence and need-at-entry) than on the first scale (i.e., frequency). A possible interpretation of these results may be a fatigue effect in raters, simply because rating such a large number of tasks (342) on three scales in a single survey is overwhelming and tiresome to complete or to give appropriate attention and judgement.

Conclusions

According to professional standards, federal regulations, and legal precedent, job analyses are considered essential to the development and validation of licensure and certification examinations. Still, research on what constitutes reliable and valid job analysis data is lacking. In addition, guidelines and procedures for ensuring valid use and interpretation of job analysis results remains to be investigated. This paper illustrates procedures that can be used to examine the reliability and validity of job analysis results. G-theory and the multi-facet Rasch IRT model are applied to job analysis data and results to investigate consistency and generalizability in task importance measures, to suggest reliable survey sample sizes, to justify the number and use of rating scales, and to detect possible rating errors.

By using random samples from job analysis data for two professions with divergent job activities, this study finds that a representative sample as small as 400 raters produces reliable task importance measures to the same degree as obtained from a large sample (more than 1,000) used in the actual job analyses. For the Real Estate job analysis, a representative sample of 450 raters produces statistically equivalent task measures for 67 tasks, as compared to a total of 1,420 raters for the actual job analysis. The decision studies also reveal that dependability coefficients for a sample size of 400 raters are larger than 0.90 when using either two or three rating scales. These coefficients are the same as produced from a sample size of 1,400. Discriminant validity is shown from an analysis of only the broker subgroup of real estate professionals, which indicates that ratings from such a non-representative sample do not produce the same task measures as those obtained from a representative sample. For the Body Therapy job analysis, results from varying sample size are similar to the Real Estate results: increasing sample sizes from 400 raters to 1,000 or 1,900 does not alter the reliability of the task measures. Additionally, using three rating scales with a small sample size yields a generalizability coefficient as high as 0.87 and a dependability coefficient of 0.75 for the task measures. All of these results suggest survey respondents numbering as small as 400 can be used to obtain reliable estimates of task importance in a job analysis. If a sample as small as 400 people fully represents the survey population, the degree of the generalizability of the job analysis results from this small sample is the same as that obtained from a larger sample of job analysis respondents.

Analyses of rating scales suggest the effectiveness of using differing numbers and types of rating scales depend on the nature of a profession. For a profession like real estate where harm to the public may have a different interpretation than for body therapy, results from two rating scales such as criticality and need at entry level produce the same reliable and valid task measures as using more rating scales. Even with one rating scale, the real estate job analysis produces very similar task measures as when using three scales. These results suggest that using not just one, but different rating scales for different tasks may be more efficient in conducting a job analysis for this type of profession. If information for some tasks in a profession concern the necessity at the time of licensing, then only one scale of need-at-entry is required for collecting task ratings. Otherwise, a frequency or importance of task rating may be adequate. Generalizability analyses find that the variance components due to scales in the body therapy job analysis data are much larger than that obtained from the real estate data; and increasing the number of rating scales greatly increases the reliability of the task ratings. These findings suggest that, for a technique-oriented profession like body therapy, three rating scales are necessary to obtain reliable and valid task measures.

The use of G-theory and the multi-facet Rasch model have been shown to be useful tools for examining the reliability and validity of job analysis results. Both methods are able to provide replicable information that can be used to determine task measure consistency, reliable sample size, and the optimal number of rating scales to use.

G-theory analyses also help to investigate validity of the job analysis measuring systems by revealing main effect (i.e., task, rater, and rating scale) variations and variability due to differential ratings of raters across scales and tasks. The findings from these analyses coincide with the conceptual understanding of job activities for each profession. For instance, the study finds for both professions, that variability due to tasks accounts for a larger percentage of the total variance as compared to the other two main effects of raters and scales. This is expected for job analysis results, because the goal of a job analysis is to validly identify and reliably differentiate tasks' importance measures. The study also finds that variation due to raters for body therapy is smaller than in real estate. A possible explanation for this finding is that training for body therapists is more stringent and standardized than for real estate salespersons.

FACETS analyses yield promising results for detecting rating errors. Limited rating ranges and fatigue effect are two types of erratic ratings that are identified for job analysis data. Results indicate that the rater severity index is a precise statistic in detecting extreme ratings across tasks. Also, outfit statistics are useful in examining ratings on the extreme categories, and infit statistics are helpful in detecting ratings in both the middle and extreme categories. Although lengthy presentations of descriptive statistics, such as means, variances, and frequencies across raters and tasks, may also detect these types of errors, the use of FACETS' indices, such as severity and fit statistics, is very efficient and much less time-consuming. A benefit of FACETS is that this diagnostic information is generated in conjunction with the task measures analysis. To fully answer questions on the kind of rating errors occurring and which errors should be identified in job analysis survey rating data, more research should be undertaken to conceptualize and categorize possible rating errors, as well as to investigate methods for detecting those errors.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Engelhard, Jr. G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. Journal of Educational Measurement, 31(2), 93-112.

Engelhard, Jr. G. (1996). Evaluating rater accuracy in performance assessments. Journal of Educational Measurement, 33(1), 56-70.

Harvey, (1991). Job analysis. In M. Dunnette and L. Hough (Eds.), Handbook of industrial and organizational psychology (2nd Ed). Palo Alto: Consulting Psychologists Press.

Kane, M. (1982). The validity of licensure examinations. American Psychologist, 6, 161-171.

Kane, M. (1986). The future of testing for licensure and certification examinations. In B. Plake & J. Will (Eds.), The future of testing (pp. 145-181). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kane, M. (1997). Model-based practice analysis and test specifications. Applied Measurement in Education, 10(1), 5-18.

Kane, M., Kingsbury, C., Colton, D., & Estes., C. (1989). Combining data on Criticality and frequency in developing test plans for licensure and certification examinations. Journal of Educational Measurement, 26(1), 17-27.

Knapp, J. & Knapp, L. (1995). Practice analysis: Building the foundation for validity. In J. C. Impara (Eds.), Licensure Testing: Purposes, Procedures, and Practices (pp. 93-116). Lincoln, NE: Buros Institute of Mental Measurements.

Linacre, J. M. (1989). Many-faceted Rasch measurement. Chicago: MESA.

Lunz, M., Stahl, J., & James, K. (1989). Content validity revisited: Transforming job analysis data into test specifications. Evaluation & the Health Professions, 12(2), 192-206.

Lunz, M. & Schumacker, R. E. (1997). Scoring and analysis of performance examinations: A comparison of methods and interpretations. Journal of Outcome Measurement, 1(3), 219-238.

Mehren, W. A. (1997). Validating licensing and certification test score interpretations and decisions: A response. Applied Measurement in Education, 10(1), 97-104.

Messick, (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-104). New York: American Council on Education and Macmillan.

Nelson, D. S. (1994). Job analysis for licensure and certification exams: Science or politics? Educational Measurement: Issues and Practices, 13(3), 29-15.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88(2), 413-428.

Sanchez, J. I. & Fraser, S. L. (1992). On the choice of scales for tasks analysis. Journal of Applied Psychology, 77(4), 545-553.

Sanchez, J. I. & Levine, E. L. (1989). Determining important tasks within jobs: A policy-capturing approach. Journal of Applied Psychology, 74(2), 336-342.

Stone, G. E. (1997). Taming tasks analysis with FACETS. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Wright, B. D. & Masters, G. N. (1982). Rating scale analysis: Rasch Measurement. Chicago: MESA.

Zegers, F. E. (1991). Coefficients of interrater agreement. Applied Psychological Measurement, 15(4), 321-333.

Table 1

Variance Estimates for Task × Rater × Scale G-studies, Using Random Effects Design (Three Random Groups, Complete Data Set, and Total Data Set from the Real Estate Job Analysis)

	Grp1	%	Grp2	%	Grp3	%	Comp	%	Total	%
Task	0.23	18.7	0.31	23.7	0.25	20.3	0.26	20.8	0.27	21.0
Rater	0.12	9.8	0.12	9.2	0.13	10.6	0.13	10.1	0.11	8.8
Scale	0.01	0.8	0.01	0.8	0.02	1.6	0.02	1.2	0.03	2.1
t × r	0.36	29.3	0.35	26.7	0.33	26.8	0.35	27.4	*	*
t × s	0.04	3.3	0.04	3.1	0.04	3.3	0.04	3.3	0.07	5.3
r × s	0.10	8.1	0.10	7.6	0.11	8.9	0.10	8.2	0.11	8.8
r × t × s	0.37	30.1	0.38	29.0	0.35	28.5	0.37	29.0	0.69	54.0

* Due to insufficient computing memory, data set was too large to estimate the effect.

Table 2

Generalizability and Dependability Coefficients for Rater × Task × Scale Decision Studies (Three Random Groups, Complete Data Set, and Total Data Set from the Real Estate Job Analysis)

Coefficient	n_r	n_s	Grp1	Grp2	Grp3	Comp
ρ^2	400	1	0.86	0.87	0.85	0.86
		2	0.92	0.93	0.91	0.92
		3	0.94	0.95	0.94	0.95
	1400	1	0.86	0.87	0.85	0.86
		2	0.92	0.93	0.92	0.93
		3	0.95	0.95	0.94	0.95
ϕ	400	1	0.82	0.84	0.80	0.82
		2	0.90	0.91	0.88	0.90
		3	0.93	0.94	0.92	0.93
	1400	1	0.82	0.84	0.80	0.82
		2	0.90	0.91	0.89	0.90
		3	0.93	0.94	0.92	0.93

Table 3

Correlation Coefficients and Statistical Tests of Task Weights between the Total Data Set and Three Random Groups, Complete Data Set, and Only Brokers from the Real Estate Job Analysis

Data paired with the total group	Sample Size	Correlation Coefficient	Test z-statistic	p-value (two-tailed)
Group 1	452	0.999	0.456	0.648
Group 2	472	0.999	0.162	0.901
Group 3	496	0.999	0.381	0.871
Complete responses	457	0.998	0.125	0.703
Brokers	233	0.837	2.268	0.023

Table 4

Variance Estimates for Task × Rater × Scale G-studies, Using Random Effects Design (Two Random Groups and Complete Data Set from the Body Therapy Job Analysis)

	Rgrp1	%	Rgrp2	%	Complete	%
Task	0.35	24.8	0.34	24.5	0.34	24.2
Rater	0.08	6.0	0.09	6.0	0.09	6.2
Scale	0.19	13.6	0.19	13.6	0.19	13.2
t × r	0.33	23.6	0.33	23.4	0.34	24.0
t × s	0.15	10.4	0.15	10.5	0.15	10.6
r × s	0.05	3.2	0.05	3.4	0.04	3.1
r × t × s	0.26	18.4	0.26	18.6	0.26	18.6

Table 5

Generalizability and Dependability Coefficients for Rater × Task × Scale Decision Studies (Two Random Groups and Complete Data Set from the Body Therapy Job Analysis)

Coefficient	n_r	n_s	Rgrp1	Rgrp2	Complete
ρ^2	400	1	0.70	0.70	0.69
		2	0.82	0.82	0.82
		3	0.87	0.87	0.87
	1000	1	0.70	0.70	0.70
		2	0.82	0.82	0.82
		3	0.88	0.87	0.87
	1900	1	0.70	0.70	0.70
		2	0.83	0.82	0.82
		3	0.88	0.87	0.87
ϕ	400	1	0.51	0.50	0.50
		2	0.67	0.67	0.67
		3	0.75	0.75	0.75
	1000	1	0.51	0.50	0.50
		2	0.67	0.67	0.67
		3	0.75	0.75	0.75
	1900	1	0.51	0.50	0.50
		2	0.67	0.67	0.67
		3	0.76	0.75	0.75

Table 6

Variance Estimates for Task × Rater × Scale G-studies, Using Random Effects Design (Complete Data Set for Each Combination of Two Rating Scales from the Real Estate Job Analysis)

	Freq & Crit	%	Freq & Need	%	Crit & Need	%
Task	0.23	18.4	0.39	27.7	0.22	18.1
Rater	0.12	9.7	0.11	8.2	0.15	12.7
Scale	0.02	1.7	0	0	0.03	2.1
t × r	0.31	24.8	0.38	27.1	0.35	29.4
t × s	0.08	6.6	0.02	1.5	0.03	2.2
r × s	0.09	7.0	0.12	8.3	0.10	8.3
r × t × s	0.40	31.8	0.38	27.2	0.33	27.3

Table 7

Generalizability and Dependability Coefficients for Rater × Task × Scale Decision Studies (Complete Data Set for Each of the Single Rating Scales and the Combination of Two Rating Scales from the Real Estate Job Analysis)

Coefficient	n_r	n_s	Freq & Crit	Freq & Need	Crit & Need
ρ^2	400	1	0.73	0.94	0.89
		2	0.84	0.97	0.94
		3	0.89	0.97	0.96
ϕ	400	1	0.68	0.94	0.80
		2	0.81	0.97	0.89
		3	0.86	0.97	0.92

Table 8

Correlation Coefficients and Statistical Tests of Task Weights between the Total Data Set and Complete Data Sets of the Three Single Rating Scales and the Combinations of Two Rating Scales from the Real Estate Job Analysis

Data paired with the total group	Sample Size	Correlation Coefficient	Test z-statistic	p-value (two-tailed)
Frequency	815	0.977	0.175	0.861
Criticality	499	0.938	0.319	0.750
Need-at-entry	545	0.992	0.181	0.856
Freq & Crit	486	0.997	0.469	0.639
Freq & Need	523	0.994	0.250	0.803
Crit & Need	470	0.986	0.468	0.639

Table 9

Numbers of Respondents Rating Only One Category within Each Scale in Three Equal Tasks Blocks and All Tasks from the Body Therapy Job Analysis

Scale	Block 1	Block 2	Block 3	Total
Frequency	7	6	5	2
Competence	17	34	30	8
Need-at-entry	30	48	46	20

Appendix A

Task Importance Measures in Rasch Logits and their Transformed Percentage Weights for the Total Real Estate Job Analysis Data Set, Complete Data Set, the Three Random Groups, and Brokers

	Total (# r =1420)		Complete (# r = 457)		Group 1 (# r = 472)		Group 2 (# r = 452)		Group 3 (# r = 496)		Brokers (# r = 233)	
Task #	Rasch	Percentage	Rasch	Percentage	Rasch	Percentage	Rasch	Percentage	Rasch	Percentage	Rasch	Percentage
3	0.68	1.4	0.72	1.42	0.72	1.38	0.64	1.42	0.69	1.4	0.86	1.02
7	0.8	1.54	0.78	1.5	0.9	1.59	0.75	1.54	0.74	1.47	0.99	1.19
8	1.53	2.41	1.56	2.49	1.61	2.43	1.49	2.36	1.5	2.41	1.79	2.28
9	1.69	2.61	1.67	2.63	1.79	2.64	1.66	2.54	1.64	2.58	1.92	2.46
10	0.32	0.97	0.37	0.98	0.37	0.97	0.28	1.02	0.32	0.95	0.73	0.84
11	1.76	2.69	1.79	2.78	1.91	2.78	1.69	2.58	1.7	2.65	1.95	2.5
12	1.06	1.85	1.06	1.85	1.16	1.9	1.02	1.84	1.01	1.8	1.45	1.82
13	1.47	2.34	1.58	2.52	1.55	2.36	1.47	2.34	1.41	2.29	1.77	2.25
14	1.43	2.3	1.43	2.33	1.47	2.26	1.43	2.29	1.41	2.29	1.6	2.02
15	1.76	2.69	1.7	2.67	1.78	2.63	1.74	2.63	1.76	2.73	1.9	2.43
16	1.17	1.98	1.12	1.93	1.18	1.92	1.16	1.99	1.19	2.02	1.41	1.76
17	1.76	2.69	1.71	2.68	1.79	2.64	1.73	2.62	1.77	2.74	1.94	2.48
18	0.61	1.31	0.59	1.26	0.69	1.35	0.54	1.31	0.6	1.29	0.72	0.83
19	1.5	2.38	1.46	2.36	1.52	2.32	1.51	2.38	1.47	2.37	1.84	2.35
20	1.22	2.04	1.23	2.07	1.24	1.99	1.19	2.03	1.24	2.08	1.41	1.76
21	1.53	2.41	1.56	2.49	1.56	2.37	1.5	2.37	1.55	2.47	1.79	2.28
22	1.43	2.3	1.46	2.36	1.47	2.26	1.38	2.24	1.44	2.33	1.7	2.16
23	1.32	2.16	1.32	2.19	1.38	2.16	1.24	2.08	1.34	2.21	1.51	1.9
24	1.25	2.08	1.27	2.12	1.29	2.05	1.21	2.05	1.26	2.11	1.55	1.95
25	1.48	2.36	1.47	2.38	1.51	2.31	1.41	2.27	1.51	2.42	1.8	2.29
26	1.72	2.64	1.69	2.66	1.8	2.65	1.68	2.57	1.71	2.67	1.87	2.39
27	1.05	1.84	1	1.78	1.11	1.84	0.97	1.78	1.08	1.89	1.2	1.48
28	1.22	2.04	1.17	1.99	1.25	2	1.21	2.05	1.2	2.03	1.37	1.71
29	1.14	1.95	1.1	1.91	1.23	1.98	1.1	1.93	1.11	1.92	1.33	1.66
30	1.15	1.96	1.1	1.91	1.22	1.97	1.12	1.95	1.11	1.92	1.31	1.63
31	0.24	0.87	0.17	0.72	0.26	0.84	0.21	0.95	0.25	0.86	0.18	0.09
32	0.91	1.67	0.89	1.64	1.04	1.76	0.84	1.64	0.88	1.64	1.12	1.37
33	1.27	2.1	1.25	2.1	1.3	2.06	1.21	2.05	1.29	2.15	1.48	1.86
34	1.13	1.94	1.12	1.93	1.16	1.9	1.08	1.91	1.16	1.98	1.29	1.6
35	0.97	1.75	0.96	1.73	1	1.71	0.92	1.73	1	1.79	1.19	1.47
36	0.29	0.93	0.34	0.94	0.35	0.95	0.24	0.98	0.3	0.92	0.66	0.75
37	0.4	1.06	0.36	0.97	0.42	1.03	0.41	1.17	0.4	1.04	0.42	0.42
38	1.11	1.91	1.11	1.92	1.14	1.87	1.08	1.91	1.11	1.92	1.34	1.67
39	1.11	1.91	1.16	1.98	1.12	1.85	1.14	1.97	1.08	1.89	1.53	1.93
40	1.56	2.45	1.58	2.52	1.62	2.44	1.51	2.38	1.55	2.47	1.76	2.24

41	1.37	2.22	1.36	2.24	1.38	2.16	1.34	2.19	1.4	2.28	1.73	2.2
42	0.24	0.87	0.28	0.86	0.28	0.86	0.2	0.94	0.25	0.86	0.69	0.79
43	1.69	2.61	1.67	2.63	1.73	2.57	1.7	2.59	1.66	2.6	2.02	2.59
44	0.42	1.09	0.45	1.08	0.43	1.04	0.38	1.13	0.43	1.08	0.94	1.13
45	0.74	1.47	0.73	1.44	0.79	1.46	0.76	1.55	0.68	1.39	0.99	1.19
46	0.86	1.61	0.83	1.56	0.92	1.62	0.84	1.64	0.84	1.59	1.03	1.25
47	0.94	1.71	0.91	1.66	1	1.71	0.91	1.72	0.92	1.69	1.31	1.63
48	0.95	1.72	0.96	1.73	0.96	1.66	0.93	1.74	0.96	1.74	1.22	1.51
49	0.99	1.77	0.98	1.75	1.01	1.72	0.97	1.78	0.98	1.76	1.3	1.62
50	0.77	1.51	0.78	1.5	0.83	1.51	0.75	1.54	0.72	1.44	1.14	1.4
51	1.06	1.85	1	1.78	1.1	1.83	1	1.82	1.09	1.9	1.32	1.64
53	0.36	1.02	0.44	1.07	0.44	1.05	0.3	1.05	0.36	0.99	1.63	2.06
54	1.4	2.26	1.4	2.29	1.49	2.29	1.34	2.19	1.39	2.27	1.7	2.16
55	0.24	0.87	0.35	0.95	0.39	0.99	0.12	0.85	0.24	0.85	1.51	1.9
56	0.89	1.65	0.9	1.65	0.98	1.69	0.8	1.6	0.91	1.68	2.09	2.69
57	0.01	0.6	0.09	0.62	0.04	0.58	-0.08	0.63	0.06	0.62	0.54	0.58
58	-0.17	0.38	-0.07	0.42	-0.16	0.35	-0.23	0.46	-0.11	0.41	0.46	0.47
59	-0.33	0.19	-0.27	0.16	-0.29	0.19	-0.42	0.25	-0.27	0.22	0.22	0.15
60	-0.26	0.27	-0.2	0.25	-0.22	0.28	-0.36	0.32	-0.19	0.31	0.28	0.23
61	-0.33	0.19	-0.28	0.15	-0.32	0.16	-0.41	0.26	-0.26	0.23	0.19	0.11
62	-0.32	0.2	-0.27	0.16	-0.29	0.19	-0.43	0.24	-0.24	0.25	0.27	0.22
63	-0.27	0.26	-0.22	0.23	-0.22	0.28	-0.38	0.3	-0.21	0.29	0.35	0.32
64	-0.35	0.17	-0.29	0.14	-0.33	0.15	-0.46	0.21	-0.27	0.22	0.18	0.09
66	-0.16	0.39	-0.08	0.41	-0.03	0.5	-0.27	0.42	-0.18	0.33	0.97	1.17
67	-0.11	0.45	-0.02	0.48	-0.01	0.52	-0.21	0.48	-0.12	0.4	1.07	1.3
68	0.05	0.64	0.14	0.69	0.15	0.71	-0.09	0.62	0.08	0.65	1.28	1.59
69	-0.1	0.47	-0.01	0.5	0	0.53	-0.22	0.47	-0.08	0.45	1.06	1.29
70	0.07	0.67	0.16	0.71	0.18	0.75	-0.02	0.69	0.07	0.64	1.42	1.78
71	-0.27	0.26	-0.23	0.22	-0.19	0.31	-0.38	0.3	-0.24	0.25	0.64	0.72
72	-0.23	0.31	-0.18	0.28	-0.15	0.36	-0.33	0.35	-0.22	0.28	0.96	1.15
73	-0.06	0.51	0.02	0.53	0	0.53	-0.14	0.56	-0.04	0.5	0.88	1.04
74	-0.32	0.2	-0.26	0.18	-0.18	0.32	-0.44	0.23	-0.32	0.15	0.98	1.18

Appendix B

Task Importance Measures in Rasch Logits and their Transformed Percentage Weights for Complete Data Sets of the Three Single Rating Scales and the Combinations of Two Rating Scales from the Real Estate Job Analysis Data

	Freq alone (# r = 815)		Crit alone (# r = 499)		Need alone (# r = 545)		Freq & Crit (# r = 486)		Freq & Need (# r = 523)		Crit & Need (# r = 470)	
Task #	Rasch	Percentage	Rasch	Percentage	Rasch	Percentage	Rasch	Percentage	Rasch	Percentage	Rasch	Percentage
3	0.99	1.72	0.72	0.74	0.67	1.38	0.76	1.41	0.72	1.59	0.73	1.09
7	1.01	1.74	1.15	1.25	0.64	1.35	0.96	1.64	0.72	1.59	0.89	1.3
8	2.05	2.43	2.16	2.44	1.4	2.09	1.85	2.65	1.42	2.23	1.65	2.27
9	1.75	2.23	2.8	3.19	1.64	2.33	1.97	2.78	1.47	2.28	1.97	2.68
10	-0.03	1.04	0.96	1.03	0.38	1.09	0.32	0.91	0.13	1.06	0.68	1.03
11	1.79	2.26	2.48	2.81	1.93	2.61	1.87	2.67	1.63	2.42	2.07	2.81
12	0.95	1.7	1.56	1.73	1.24	1.94	1.08	1.77	0.95	1.8	1.36	1.9
13	1.29	1.92	2.16	2.44	1.78	2.47	1.48	2.23	1.33	2.15	1.88	2.56
14	1.45	2.03	2.06	2.32	1.57	2.26	1.53	2.28	1.32	2.14	1.72	2.36
15	1.72	2.21	2.62	2.98	1.89	2.57	1.89	2.69	1.59	2.39	2.1	2.84
16	0.87	1.64	1.94	2.18	1.49	2.18	1.16	1.86	1.01	1.86	1.64	2.26
17	1.79	2.26	2.51	2.85	1.92	2.6	1.89	2.69	1.62	2.42	2.07	2.81
18	0.3	1.26	1.06	1.14	0.91	1.61	0.55	1.17	0.49	1.39	0.98	1.41
19	1.7	2.2	2.05	2.31	1.55	2.24	1.66	2.43	1.41	2.22	1.71	2.34
20	1.43	2.02	1.62	1.8	1.25	1.95	1.34	2.07	1.16	2	1.39	1.94
21	1.45	2.03	2.37	2.68	1.67	2.36	1.64	2.41	1.36	2.18	1.88	2.56
22	1.68	2.18	1.93	2.17	1.45	2.14	1.59	2.35	1.35	2.17	1.61	2.22
23	1.4	2	1.82	2.04	1.42	2.11	1.42	2.16	1.23	2.06	1.55	2.14
24	1.2	1.86	1.84	2.06	1.4	2.09	1.32	2.05	1.14	1.98	1.55	2.14
25	1.61	2.14	2.08	2.34	1.54	2.23	1.63	2.4	1.37	2.19	1.71	2.34
26	1.73	2.22	2.73	3.11	1.76	2.45	1.93	2.74	1.52	2.32	2.04	2.77
27	1.24	1.89	1.31	1.44	1.14	1.84	1.12	1.82	1.04	1.89	1.21	1.7
28	1.35	1.96	1.67	1.86	1.26	1.96	1.34	2.07	1.14	1.98	1.42	1.97
29	1.42	2.01	1.44	1.59	1.16	1.86	1.26	1.98	1.11	1.95	1.27	1.78
30	1.44	2.02	1.42	1.57	1.19	1.89	1.25	1.97	1.14	1.98	1.28	1.79
31	0.27	1.24	0.21	0.14	0.29	1.01	0.22	0.8	0.25	1.17	0.33	0.58
32	1.04	1.76	1.17	1.27	0.98	1.68	0.98	1.66	0.89	1.75	1.07	1.53
33	1.7	2.2	1.63	1.81	1.22	1.92	1.45	2.19	1.22	2.05	1.38	1.92
34	1.51	2.07	1.4	1.54	1.11	1.81	1.28	2	1.12	1.96	1.23	1.73
35	0.99	1.72	1.39	1.53	1.04	1.74	1.04	1.73	0.89	1.75	1.19	1.68
36	-0.04	1.03	0.88	0.93	0.35	1.07	0.29	0.88	0.11	1.04	0.64	0.98
37	0.59	1.45	0.36	0.32	0.39	1.1	0.44	1.05	0.44	1.34	0.44	0.72
38	1.58	2.12	1.35	1.48	1.05	1.75	1.27	1.99	1.1	1.94	1.18	1.67
39	1.04	1.76	1.75	1.95	1.17	1.87	1.2	1.91	0.96	1.81	1.39	1.94
40	1.55	2.1	2.39	2.71	1.65	2.34	1.7	2.48	1.39	2.21	1.87	2.55

41	1.44	2.02	1.97	2.21	1.46	2.15	1.49	2.24	1.26	2.09	1.63	2.24
42	-0.07	1.01	0.73	0.75	0.34	1.06	0.23	0.81	0.09	1.02	0.57	0.89
43	1.65	2.16	2.71	3.08	1.77	2.46	1.87	2.67	1.49	2.3	2.04	2.77
44	0	1.06	1.15	1.25	0.58	1.29	0.39	0.99	0.21	1.13	0.86	1.26
45	0.6	1.46	1.21	1.32	0.83	1.54	0.77	1.42	0.62	1.5	1.01	1.45
46	0.86	1.63	1.16	1.26	0.99	1.69	0.89	1.56	0.81	1.68	1.07	1.53
47	0.87	1.64	1.41	1.55	1.05	1.75	0.99	1.67	0.83	1.7	1.21	1.7
48	0.8	1.59	1.5	1.66	1.11	1.81	0.98	1.66	0.82	1.69	1.27	1.78
49	0.97	1.71	1.49	1.65	1.04	1.74	1.07	1.76	0.88	1.74	1.23	1.73
50	0.66	1.5	1.29	1.41	0.79	1.5	0.83	1.49	0.63	1.51	1.02	1.46
51	1	1.73	1.74	1.94	1.08	1.78	1.17	1.88	0.91	1.77	1.34	1.87
53	-0.01	1.05	1.12	1.21	0.41	1.12	0.38	0.98	0.15	1.08	0.76	1.13
54	1.25	1.9	2.27	2.57	1.55	2.24	1.48	2.23	1.22	2.05	1.78	2.43
55	-0.28	0.87	1.29	1.41	0.26	0.98	0.26	0.84	-0.05	0.89	0.73	1.09
56	0.66	1.5	1.7	1.9	0.99	1.69	0.95	1.63	0.7	1.58	1.27	1.78
57	-0.33	0.84	0.64	0.65	-0.06	0.66	0.04	0.59	-0.21	0.75	0.34	0.59
58	-0.73	0.57	0.66	0.67	-0.2	0.53	-0.16	0.37	-0.46	0.52	0.28	0.51
59	-0.91	0.45	0.37	0.33	-0.39	0.34	-0.33	0.17	-0.61	0.38	0.08	0.26
60	-0.8	0.52	0.5	0.48	-0.34	0.39	-0.25	0.26	-0.54	0.45	0.15	0.35
61	-0.84	0.5	0.33	0.28	-0.44	0.29	-0.32	0.18	-0.6	0.39	0.03	0.19
62	-0.86	0.48	0.39	0.35	-0.42	0.31	-0.31	0.2	-0.6	0.39	0.07	0.25
63	-0.83	0.5	0.53	0.52	-0.37	0.36	-0.25	0.26	-0.57	0.42	0.15	0.35
64	-1.07	0.34	0.51	0.5	-0.43	0.3	-0.35	0.15	-0.7	0.3	0.12	0.31
66	-0.56	0.68	0.47	0.45	-0.23	0.5	-0.15	0.38	-0.39	0.58	0.18	0.39
67	-0.54	0.7	0.68	0.7	-0.24	0.49	-0.07	0.47	-0.38	0.59	0.26	0.49
68	-0.4	0.79	1.01	1.08	-0.06	0.66	0.09	0.65	-0.24	0.72	0.47	0.76
69	-0.45	0.76	0.56	0.55	-0.22	0.51	-0.06	0.48	-0.32	0.65	0.23	0.45
70	-0.37	0.81	1.08	1.17	-0.02	0.7	0.12	0.68	-0.22	0.74	0.51	0.81
71	-0.67	0.61	0.42	0.39	-0.49	0.24	-0.22	0.3	-0.53	0.46	0.05	0.22
72	-0.45	0.76	0.23	0.17	-0.48	0.25	-0.17	0.36	-0.41	0.56	-0.03	0.12
73	-0.44	0.76	0.69	0.71	-0.18	0.55	-0.02	0.53	-0.31	0.66	0.29	0.53
74	-0.59	0.66	0.13	0.05	-0.52	0.21	-0.27	0.24	-0.5	0.48	-0.09	0.04

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM033279

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Examining Reliability & Validity of Job Analysis Survey Data

Author(s): Ning Wang, Randall Wiser, and Larry Newman

Corporate Source: Paper presented at the 1999 annual meeting of the National Council on Measurement in Education, Montreal, Canada

Publication Date: April, 1999

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection ~~subscribers only~~

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educationel Resources Information Center (ERIC) nonexclusive permission to reproduce end disseminete this document es indicated ebove. Reproduction from the ERIC microfiche or electronic medie by persons other then ERIC employees end its system contractors requires permission from the copyright holder. Exception is mede for non-profit reproduction by libraries end other service egencies to setisfy informetion needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:

Printed Name/Position/Title: Ning Wang / Senior Psychometrician

Organization/Address: Assessment Systems, Inc.
3 Bala Plaza West
Bala Cynwyd, PA 19004

Telephone: (610)617-5008

FAX: (610)617-1335

E-Mail Address: Ning-Wang@asisvcs.net

Date: 8/6/01

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com