

5-2006

Assessing Reliability of Expert Ratings Among Judges Responding to a Survey Instrument Developed to Study the Long Term Efficacy of the ABET Engineering Criteria, EC2000

Tracy L. Litzinger

Embry-Riddle Aeronautical University - Daytona Beach

Follow this and additional works at: <https://commons.erau.edu/db-theses>



Part of the [Industrial and Organizational Psychology Commons](#)

Scholarly Commons Citation

Litzinger, Tracy L., "Assessing Reliability of Expert Ratings Among Judges Responding to a Survey Instrument Developed to Study the Long Term Efficacy of the ABET Engineering Criteria, EC2000" (2006). *Theses - Daytona Beach*. 123.

<https://commons.erau.edu/db-theses/123>

This thesis is brought to you for free and open access by Embry-Riddle Aeronautical University – Daytona Beach at ERAU Scholarly Commons. It has been accepted for inclusion in the Theses - Daytona Beach collection by an authorized administrator of ERAU Scholarly Commons. For more information, please contact commons@erau.edu.

Assessing Reliability of Expert Ratings Among Judges Responding to a Survey
Instrument Developed to Study the Long Term Efficacy of the ABET Engineering
Criteria, EC2000

by

Tracy L. Litzinger

A Thesis Submitted to the Department of Human Factors and Systems in Partial
Fulfillment of the Requirements for the Degree of Master of Science in Human Factors &
Systems

Embry-Riddle Aeronautical University
Daytona Beach, Florida
May 2006

UMI Number: EP32097

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.



UMI Microform EP32097
Copyright 2011 by ProQuest LLC
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

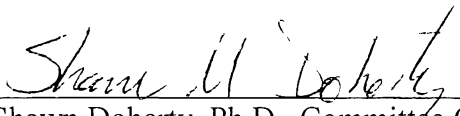
Assessing Reliability of Expert Ratings Among Judges Responding to a Survey
Instrument Developed to Study the Long Term Efficacy of the ABET Engineering
Criteria, EC2000


by

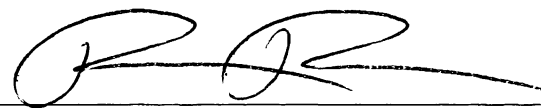
Tracy L. Litzinger

This thesis was prepared under the direction of the candidate's thesis committee chair, Shawn Doherty, Ph.D., Department of Human Factors & Systems, and has been approved by the members of the thesis committee. It was submitted to the Department of Human Factors & Systems and has been accepted in partial fulfillment of the requirements for the degree of Master of Science in Human Factors & Systems.

THESIS COMMITTEE:

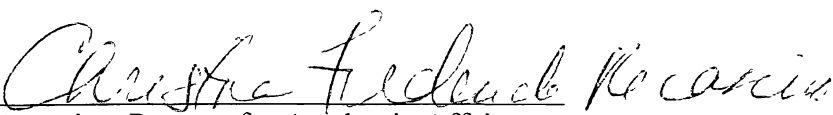

Shawn Doherty, Ph.D., Committee Chair


Elizabeth Blickensderfer, Ph.D., Member


Rosemarie Reynolds, Ph.D., Member


MS HFS Program Coordinator


Department Chair, Department of Human Factors & Systems


Associate Provost for Academic Affairs

Acknowledgements

The author would like to express deep gratitude to her family, most notably, her husband, Mark, and her father, Darwin, and friends for their “immeasurable” support and patience through all of these long years. She would especially like to dedicate this accomplishment to her children, Aubrey and Dane, and in memory of her mother, Ginger. She also wishes to thank her committee members, Dr. Shawn Doherty, Dr. Beth Blickensderfer, and Dr. Rosemarie Reynolds for their time, knowledge, guidance, and particularly, their warmth and tolerance. She also wishes to recognize the former members of her committee, Dr. John Wise, Dr. Steve Hall, and Ms. Maria Franco, as well as, the staff of the Proctor Library at Flagler College. Their expertise, advice, and most importantly, their encouragement were “inestimable” toward finally completing this epic thesis.

Abstract

In today's assessment processes, especially those evaluations that rely on humans to make subjective judgements, it is necessary to analyze the quality of their ratings. The psychometric issues associated with assessment provide the lens through which researchers interpret results and important decisions are made. Therefore, inter-rater agreement (IRA) and inter-rater reliability (IRR) are pre-requisites for rater-dependent data analysis. A survey instrument cannot provide "good" information if it is not reliable; in other words, reliability is central to the validation of an instrument. When judges cannot be shown to reliably rate a performance, item, or target, the question becomes why the judges' responses are different from one another. If the judges' ratings covary unreliably because the construct is poorly defined or the rating framework is defective, then the resultant scores will have questionable meaning. On the other hand, if the judges' ratings differ because they have a true difference in opinion, this is of importance to the researcher and may not necessarily diminish the validity of the scores. The intraclass correlation coefficient (*ICC*) is the most efficient method to assess these rater differences and identify the specific sources of inconsistency in measurement. This study examined how *ICCs* can be used to inform researchers of the extent in which legitimate differences of opinion may appear as a lack of reliability and/or agreement, demonstrating the need for analyzing survey data beyond standard descriptive statistics. Overall, both the IRA and IRR correlations, as calculated by *ICC*, ranged from .79 to .91 indicating high levels of agreement and consistency in the scoring among the judges' ratings. When group membership was accounted for the IRA values increased suggesting the common judges agreed more than those judges who varied in their perspectives.

Table of Contents

Acknowledgements.....	3
Abstract.....	4
List of Tables.....	6
List of Figures and Formulas.....	7
Introduction.....	9
Validity vs. Reliability: Their Interrelatedness.....	17
Survey Methodology.....	18
Errors of measurement.....	19
Survey construction.....	22
Rating scale.....	24
Rater error.....	26
Types of rater errors.....	27
Survey administration.....	28
Sources of Validity Evidence.....	29
Content validity evidence.....	29
Construct validity evidence.....	31
Assessment-criterion relationship validity evidence.....	34
Consequence validity evidence.....	35
Methods of Estimating Reliability.....	37
Classical Test Theory vs. Modern Test Theory: Approaches to Reliability	
Estimation.....	40
“G” theory.....	41
Scale-Dependent Metrics: Internal Consistency Reliability Coefficients.....	46
Rater-Dependent Metrics: Methods of Estimating Inter-rater Agreement and	
Reliability.....	48
Inter-rater Agreement (IRA).....	52
Inter-rater Reliability (IRR).....	60
Intraclass Correlation Coefficient (<i>ICC</i>).....	63
Increasing Reliability.....	74
Statement of Hypothesis.....	77
Method.....	78
Aim of Study.....	78
Participants.....	78
Measures.....	79
Procedure.....	80
Reliability Analysis.....	81
Results.....	89
Discussion.....	98
Recommendations.....	103
Summary.....	108
References.....	110
Appendix A.....	118
Appendix B.....	121
Appendix B1.....	126

Appendix B2.....	127
Appendix B3.....	128
Appendix B4.....	129
Appendix C.....	130
Appendix C1.....	131
Appendix D.....	132

List of Tables

Table 1 Conditions in which the outcomes of IRA and IRR may be lower or higher due to either rater idiosyncrasies and subjectivity or survey idiosyncrasies	16
Table 2 Description of seven independent sources of variance	43
Table 3 Hypothetical ratings of accurate empathy illustrating different levels of IRA and IRR for interval-scaled data	49
Table 4 Fully-crossed (judge X target) Shrout and Fleiss (1979) analysis of variance and mean square expectations design in which each judge rates each target once	69
Table 5 <i>ICC</i> formulas for cases of consistency and whether it is an individual rating or the mean of all the ratings	71
Table 6 <i>ICC</i> formulas for cases of absolute agreement and whether it is an individual rating or the mean of all the ratings	71
Table 7 ANOVA summary table for Case 1	84
Table 8 ANOVA summary table for Case 2	84
Table 9 ANOVA summary table for Case 3	85
Table 10 Computation of <i>ICC</i> for Case 1	85
Table 11 Computation of <i>ICC</i> for Case 2	85
Table 12 Computation of <i>ICC</i> for Case 3	85
Table 13 Computed confidence intervals and <i>F</i> values (<i>BMS/EMS</i>), significant at 05 level for Case 2	86
Table 14 Computed confidence intervals and <i>F</i> values (<i>BMS/EMS</i>) significant at 05 level for Case 3	86
Table 15 Diverse groups' rating of EC2000 student learning outcomes criteria	89
Table 16 ANOVA summary table for diverse group	90
Table 17 Computation of <i>ICC</i> for diverse group	90
Table 18 Computed <i>F</i> values for diverse group	91
Table 19 Industry groups' rating of EC2000 student learning outcomes criteria	91
Table 20 ANOVA summary table for industry group	92
Table 21 Computation of <i>ICC</i> for industry group	92
Table 22 Computed <i>F</i> values for industry group	93
Table 23 Education groups' rating of EC2000 student learning outcomes criteria	94
Table 24 ANOVA summary table for education group	94
Table 25 Computation of <i>ICC</i> for education group	94
Table 26 Computed <i>F</i> values for education group	95
Table 27 Summary of <i>ICC</i> computations for diverse, industry and education groups	95

List of Figures and Formulas

Figure 1 Sources of error in a hypothetical test.....	19
Formula 1 The standard deviation (s_x).....	21
Formula 2 The equation for variance (s^2_x).....	21
Formula 3 Lawshe's <i>CVR</i>	31
Formula 4 Cronbach's alpha (α).....	47
Formula 5 Cohen's kappa (k).....	53
Formula 6 $P(E)$ proportion of times the judges are expected to agree by chance.....	53
Formula 7 Lawlis and Lu's (1972) chi-square (χ^2).....	55
Formula 8 T -index.....	55
Formula 9 r_{wg}	57
Formula 10 $r_{wg(J)}$ multi-item agreement.....	59
Formula 11 Average Deviation (AD).....	59
Formula 12 Finn's (1970) r	61
Formula 13 Finn's (1970) r chi-square test for significance.....	62
Formula 14 The basic linear model for a fully-crossed Shrout and Fleiss (1979) analysis of variance design in which the same k judges rate all n targets.....	67
Formula 15 The basic linear model for an analysis of variance design for Case 1 in which the effects due to the judges, the interaction between judges and targets, and the random error are not separately estimable.....	68
Formula 16 The population value under Case 1.....	69
Formula 17 $ICC(1, 10)$	70
Formula 18 The population value under Case 2.....	70
Formula 19 $ICC(2A, 1)$	70
Formula 20 The population value under Case 3.....	70
Formula 21 $ICC(3, 1)$	70
Figure 2 Computed IRR values and confidence intervals.....	96
Figure 3 Computed IRA values and confidence intervals.....	97

Introduction

Today's society relies heavily on contemporary assessment tools that require humans to judge characteristics of an individual or another entity and assign it to a point on a defined scale according to a set of rules. This judgmental information is collected in the form of rater-dependent scales called rankings, comparisons, or one of the most popular types of measures, rating scales (Saal, Downey, & Lahey, 1980). For example, judges may be used to grade students writing an essay question, to score athletes' performance in the Olympics, or to evaluate the feasibility of a new product (Stemler, 2004). No longer is psychology the only field utilizing judgmental information, now other areas of social and interdisciplinary sciences such as education, health care, marketing, human factors engineering, and industrial psychology are investing considerable resources to conduct these assessments (Saal, Downey, & Lahey, 1980). Assessments in education appear in a variety of forms, for example, to depict student strengths and weaknesses, to revise a curriculum, or to evaluate a school program (Linn & Gronlund, 2000). In medicine, it is natural to ask how often two healthcare professionals examining the same patients agree on the diagnosis data (Gwet, 2001). Managers must do everything possible to retain customers, and marketing and survey research plays an important role in increasing customer satisfaction. Human factors engineering involves an iterative process to seek and incorporate feedback about human behavior, capabilities, limitations, and motivation to the design and usability of everyday things (Sanders & McCormick, 1993). Industrial and organizational psychologists use ratings in the context of personnel selection, performance appraisals, and organizational improvement (Saal, Downey, & Lahey, 1980). A society without testing may sound

tempting; however, the critical decisions that are based on the assessment effort and human judgement may have a significant impact on peoples' lives (Cohen & Swerdlik, 2002).

Given this demand, there is an increased need for ensuring the adequacy and usefulness of assessment methods and the valid and reliable interpretations made from them (Donald & Denison, 2001; Krueger, 1993). In education research, for instance, invalid or unreliable measures can lead to erroneous conclusions and incorrect educational decisions, which can adversely impact students and teachers. Consider instructor evaluations, a widely-used assessment tool in higher education. Results may differ as a function of the item being worded differently or the timing of administration. Ratings might also differ across groups of students or if the ratings were completed at home instead of in the classroom. Even environmental issues such as poor lighting or an extreme temperature in the classroom might impact ratings (Light, Singer, & Willett, 1990). Essentially, the utility of the instrument is diminished to the extent that factors other than teaching skill influence student ratings (Light, Singer, & Willett, 1990).

While the evaluations given by raters may be influenced by factors other than the one of primary interest, the differences in the ratings are more likely a function of the raters than anything else. For example, during a road test for a driver's license, the student may pass or fail based partly on who is sitting in the passenger's seat and not on their performance behind the wheel (Cohen & Swerdlik, 2002). Or, people differ in terms of political views, not because some are right and some are wrong, but because they are different (S.M. Hall, personal communication, June 18, 2004). That is, their inter-individual differences emerge as one rater responds in a noticeably different manner

on a measure relative to another rater. Such variances in the ratings result in part from rater idiosyncrasies and subjectivity (i.e., raters' true differences in perspective or opinion) or from survey idiosyncrasies (i.e., no two raters interpret the items in the same manner) (Brannick, 2003; Feldt & Brennan, 1989; Tinsley & Weiss, 1975). Traditional and new forms of assessment and their resultant scores are the lens through which researchers, policy makers, managers, and legislators indirectly observe people.

Therefore, evidence regarding expanded concepts of validity and reliability is imperative to assure that the results provide valuable information to make important decisions (Linn, Baker, & Dunbar, 1992). Not surprisingly, researchers have devoted much time and effort to the science of refining measurement techniques called psychometrics (Saal, Downey, & Lahey, 1980). Psychometrics are mainly concerned with empirical data generated from an assessment tool and the statistical analysis and standardization of the results (i.e., validity and reliability) (Psychometrics, 2005).

The best way to confirm the psychometric quality of a measuring device is to examine the psychometric properties of the scores themselves. If researchers want to yield meaningful data, they cannot avoid establishing the integrity of their data (Thompson, 2003a). Prior to this evaluative work, the single most important first step toward producing useful and reliable findings is careful research design (i.e., survey methodology) (Light, Singer & Willett, 1995). The qualities that constitute a valuable assessment form include obvious elements such as clarity of directions, logical arrangement of questions and response categories, and convenience of administration, scoring, and interpretation, but the most important quality of the measuring device is its validity. An instrument with high validity provides the information the decision maker

needs to improve decisions (Cronbach, 1970). Validity will inform them most about their desired interest (i.e., construct); therefore, a quality instrument is defined as a set of items designed to elicit or describe a behavior in a specified construct (Fowler, 1993; Joint Committee, 1999). Equally important is presenting these items in such a way that they will elicit a similar response from test-takers who have similar perspectives (Murphy & Davidshofer, 2001). Assessing rating quality and estimating the reliability of raters' scores impacts the ultimate utility (i.e., validity) of the data and the instrument itself. Basically, consistency of scores is important in determining whether or not an instrument can provide "good" information (Murphy & Davidshofer, 2001).

Relative to the weight that has been given to assessment information, researchers must move beyond simply recounting a property of the measure to investigating the psychometric implications (i.e., validity and reliability) of the results and the interpretations, and use of those results. Although determining the psychometric quality of a measurement is an essential statistical method, little research is reported in the literature and the training available regarding psychometric properties is even more limited in graduate and doctoral programs (Coleman, VanAken, & Shen, 2002; Thompson, 2003a). For this reason, most survey researchers tend to create simple descriptive statistics to communicate their results and interpretations, which often fail to capture what is really occurring within the survey instrument itself and its results. There is much more to be learned from survey interpretation rather than just generating statistics such as frequency counts, central tendencies (i.e., mean, median, and mode), standard deviations, tables and graphs. In practice, consider the manner in which survey results are often reported. Frequency distributions are probably the most common and

are a usual first step to organize the data using simple indexes (e.g., 25% of the respondents rated Item 4 as 3 or higher). Means probably best represent an entire group of scores and provide certain assumptions about the distribution of those scores (e.g., the average number or mean of correct answers given on a 10-word spelling test for 25 4th graders is 8.04). While standard deviations may help to fully understand the distribution of scores, all it tells the survey user is on average how much each score in a set of scores varies from the mean (Salkind, 2000).

Instead, the more advanced researcher will look for patterns of variability to measure if certain respondents systematically differ in their scores from one another and then quantify this agreement (or lack thereof) and consistency (or lack thereof). If two judges cannot be shown to reliably rate an individual's observed behavior, what does that reveal about the particular scores, the scorers themselves, the individual, and the ultimate utility of the instrument? Since all forms of subjective assessment are susceptible to variability due to such factors as the raters, the rating scale, the sampling of the content or topic area, the tasks or items, and the test-taking environment, knowledge of inter-rater agreement (IRA) and inter-rater reliability (IRR) are critical in establishing best practices for evaluation of subjective data (Coleman, VanAken, & Shen, 2002; Linn, Baker, & Dunbar, 1992; Stemler, 2004). Tinsley and Weiss (1975) offered a framework that recommended whenever rating scales are applied and before the ratings can be accepted, evidence of both IRA and IRR of the ratings is required. "When both inter-rater reliability and agreement are low, the ratings are of no value and should not be used for research or applied purposes", proposed Tinsley and Weiss (1975, p. 360). For the most part, the survey researchers that choose to investigate rater variability in survey responses

and obtain a low IRA and/or IRR correlation follow this procedure by revising or discarding their instrument. They conclude that it is a function of poor survey construction and not the “true” ideological differences among the judges’ perspectives. However, the reality is a lack of IRA and IRR in response to specific items by various groups of judges is still meaningful and may indicate legitimate differences of opinion and not an artifact of survey idiosyncrasies (S. M. Hall, personal communication, August 25, 2004).

In addition to the limited knowledge about reliability analysis, researchers often use the terms IRA and IRR interchangeably as if they mean the same thing (Goodwin, 2001). IRA refers to the extent that judges tend to make exactly the same judgments about a rated subject. On a numerical scale, the judges assign exactly the same values when rating the same item (Tinsley & Weiss, 1975). “Whereas inter-rater reliability is the extent to which the raters order the participants’ performances, behaviors, or essays in the same way,” defined Goodwin (2001). There are several statistical methods for computing IRA and IRR and researchers should also be aware that different approaches to analyzing results may shape different implications for how rating scores across multiple judges are estimated (Stemler, 2004). Although there is no single, preferred approach to assessing rater reliability, the methods of generalizability theory (“G” theory) are best suited for scores derived from such measures (Joint Committee, 1999). The intraclass correlation coefficient (*ICC*) is an extension of “G” theory. Unlike traditional reliability techniques, *ICC* uses a familiar statistical method, the repeated measures analysis of variance (ANOVA), to compare the different ratings of multiple judges and compartmentalize the variability due to the raters and what is being rated, called targets

(i.e., items). This level of detail provides obligatory information about the properties of scores that are central to the validation of an instrument (Murphy & Davidshofer, 2001). Research that relies on subjective ratings must establish that the collected data meets some psychometric criteria for both reliability and validity. Therefore, it is the responsibility of the researcher to argue that the data represents a single construct of interest and there is a reasonable level of agreement¹ among the raters. By analyzing judgmental data using *ICCs* the extent to which the judges' ratings reliably covary (i.e., reliability) and are in absolute agreement can be determined. High levels of agreement indicate the various judges share a similar meaning of the construct being measured, regardless of whether or not they apply the construct similarly across targets. High levels of reliability indicate a consistent application of criteria across targets, even if each judge applies the criteria differently (i.e., some judges may be more lenient than others, but their resulting ratings covary). On the other hand, an analysis of data collected from distinct groups of judges who collectively hold legitimate differences in opinions, as a function of their group membership, may produce a low IRA correlation. As a result, if group membership is considered in the analysis the IRA value should be higher. Still, common judges responding to a well-developed survey with a well-defined construct may truly disagree with one another (i.e., low IRA), but in conjunction with a high IRR value indicates a consistent within-group application of criteria. It is important to note that a lack of variability within the judges' ratings (i.e., ratings that are highly similar) will produce a lower IRR correlation causing the scores to appear unreliable even though they are accurately reflecting raters' true scores. This condition may occur particularly if

¹When continuous rating scales are used "agreement" reflects similarity across raters in the resulting rankings of the targets.

group membership is considered in the interpretation of the data due to the homogenous nature of a group of common judges. Nevertheless, the degree of agreement or consistency that exists between the judges is what makes validity possible (Cohen & Swerdlik, 2002; Lahey, Downey, & Saal, 1983; Linn & Gronlund, 2000). Table 1 describes those conditions in which the outcomes of IRA and IRR may be low or high due to either rater idiosyncrasies and subjectivity or survey idiosyncrasies.

	Low IRA	High IRA	Low IRR	High IRR
Rater Idiosyncrasies				
Raters possess true differences in perspective or opinion	✓		✓	✓
Raters truly agree about the items of interest (i.e., homogeneity)		✓	✓	✓
Raters inconsistent in their application of shared criteria	✓		✓	
Raters consistent in their application of shared criteria		✓		✓
Systematic rater errors (i.e., restriction of range)	✓		✓	
Survey Idiosyncrasies				
Items are ambiguous or have multiple meanings	✓		✓	
Items are clear and concise		✓		✓
Inconsistent administration across multiple administrations	✓		✓	

Table 1 shows those conditions in which the outcomes of IRA and IRR may be low or high due to either rater idiosyncrasies and subjectivity or survey idiosyncrasies

The main purpose of this study is to examine the extent to which legitimate differences in opinion among groups of raters can appear as lower IRA and IRR correlations. The data provided for this study was collected from expert raters via an

authentic survey instrument designed to measure the value and impact of Engineering Criteria 2000 (EC2000). EC2000 are adopted outcomes-based criteria used by the Accreditation Board of Engineering and Technology (ABET) to accredit programs in applied science, computing, engineering, and technology at colleges and universities across the nation. Appendix A contains a brief background in respect to ABET and EC2000. A sample of ABET representatives as well as non-ABET members from the engineering community were presented a set of EC2000 related questions specific to their profession. The survey instrument and its resulting data were used primarily as a real-world vehicle, maintaining important characteristics of the judges and the task, to propose a framework that can and should be applied outside the realm of EC2000.

Validity vs. Reliability: Their Interrelatedness

Validity, in terms of psychometrics, is used to describe the function of what the scores on a measure mean, or the meaningfulness of the scores. The instrument itself is not being validated, but the inferences and conclusions reached on the basis of the scores. Also, built in to a judgment of validity is a judgment of usefulness (Cohen & Swerdlik, 2002; Murphy & Davidshofer, 2001). Validity, which is based on appraisals of relevance and utility, is unlike reliability, which is essentially a technical issue. Furthermore, validity is a matter of degree, not all or none. Therefore, validity refers to the degree to which “empirical” evidence justifies or nullifies the adequacy and appropriateness of what the scores mean. It is an evolving property which can be enhanced or reduced by new modes of assessment, new findings, or changing social conditions (Messick, 1989).

Whereas validity is concerned with the appropriateness of the interpretations made from the results, reliability refers to the consistency of those results. Reliability is a

necessary condition and precursor to validity, meaning an instrument can be no more valid than it is reliable. One cannot ask the question, “Does it measure what it purports to measure?” because, without reliability, there is no “it”. Reliability provides the consistency that makes validity possible; hence, an assessment tool that produces totally inconsistent results cannot provide valid information about whatever is being measured. The scores must be shown to be reasonably consistent over different conditions, different samples, or different raters of the same performance or the instrument will have little utility. Therefore, it would be beneficial to spend as much effort on improving an instrument as assessing its validity (Cronbach, 1970; Linn & Gronlund, 2000; Sawilosky, 2000).

Survey Methodology

A “good” survey, one that obtains meaningful and valid results, is not created by chance; it is usually the result of careful design (i.e., concept, construction, selection of items, and administration) (Fowler, 1993). Associated with a good instrument is a scoring procedure (i.e., scoring rubrics and rater training) that allows researchers to accurately quantify and interpret behavior (Joint Committee, 1999). At the same time, these elements, test construction, test administration, scoring, and interpretation, also contribute to the sources of measurement error reflected in an individual’s score. Figure 1 illustrates the different components of error found in a hypothetical measure (Cohen & Swerdlik, 2002, p. 140). Essentially, better survey design is one of the least costly ways to minimize the effects of measurement error and improve survey results (Fowler, 1993).

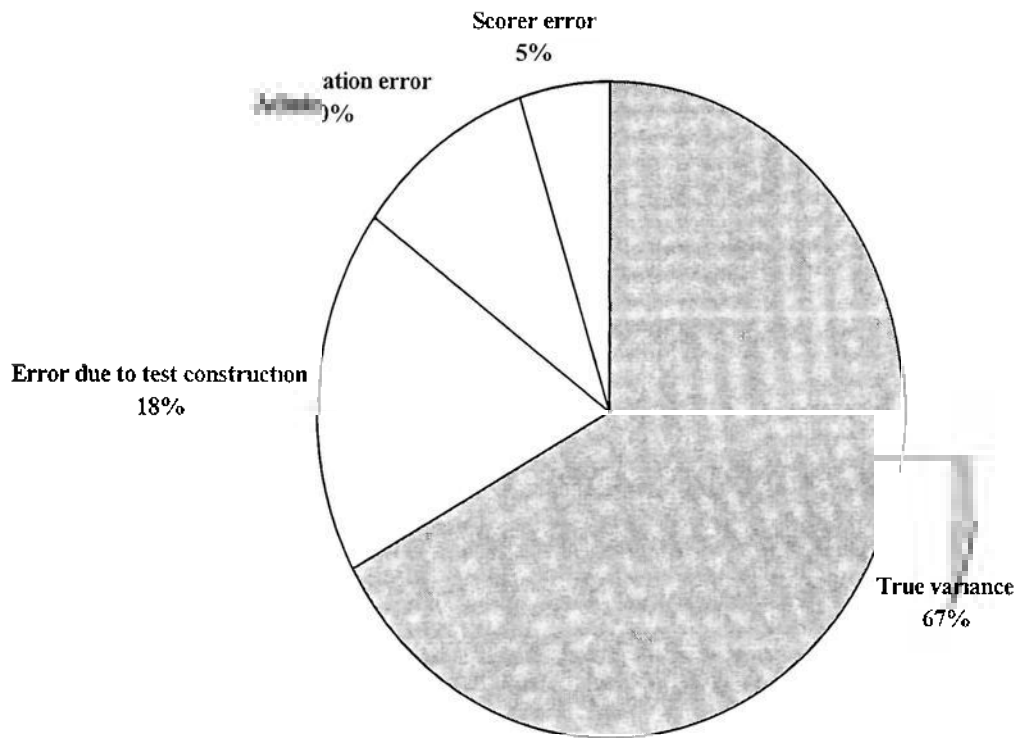


Figure 1 Sources of error in a hypothetical test

Errors of measurement. In simplest terms, “error” refers to the factors of the observed score on a measure that is unrelated to the construct of interest. An individual’s observed score will always reflect at least a small amount of error. For example, on an ability test, an individual’s score captures their “true” score on the ability being measured, as well as error. It can make the individual’s ability appear lower or higher than their true or actual ability or if the errors have little effect on their ability scores, then they should accurately reflect an individual’s ability (Cohen & Swerdlik, 2002, Joint Committee, 1999). This variation in scores may be due to fluctuations in memory, attention, motivation, or fatigue; changes in health, experience, or environment; or misinterpretations of the wording of an instrument or especially subjective judgements

made by humans, in other words, "...those rooted within the examinees and those external to them" (Joint Committee, 1999, p.26). Survey developers would prefer to base their conclusions on an ideal, error-free value; however, an observed score is actually what all tests can produce, because the attributes sampled are limited and the test cannot be repeated exhaustively until all errors balance out or until an average score is obtained that more closely approaches their true score (Cronbach, 1977; 1990). The difference between an individual's observed score and their true score on a measure is called *error*. It can be expressed as the observed score equals the true score plus error represented as $X = T + e$. Information about measurement error allows researchers to predict the range of fluctuations likely to occur in an individual's score due to factors that are not considered to be a part of the construct of interest and it is essential to the study of reliability (i.e., the more consistent test scores are from one measurement to another or one rater to another, the less error there is and the higher the reliability). Moreover, this consistency of test scores is important in determining whether a test can provide good measurement (Murphy & Davidshofer, 2001).

Variance is a useful statistic to describe variability, dispersion, or spread of scores. Although variance is used as a practical measure of variability in many statistical formulas, it is not often applied directly to a set of data because it is based on squared deviation scores. The standard deviation is computed as an average distance from the mean. The larger the standard deviation, the more spread the values are, and more different they are from one another. The formula for computing the standard deviation is as follows:

$$s_x = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}} \quad [1]$$

where, s_x is the standard deviation, \sum is the sigma or find the sum of what follows, X is each individual score, \bar{X} is the mean of all the scores, and N is the sample size, or for any variance computation this is also referred to as the degrees of freedom (df), the number of scores minus one. Therefore, the equation for variance (s_x^2) is:

$$s_x^2 = \frac{\sum (X - \bar{X})^2}{N - 1} \quad [2]$$

As a measure of variability, the standard deviation and the variance are similar, however, the standard deviation is stated in the original units that it was derived from and the variance is stated in units that are squared. The much more interpretable measure is, for example, on average, each person in a distribution of 10 different people is within 3.7 years of the group mean of 25 years rather than the average difference between each person is 13.67 years squared from the mean (Salkind, 2000).

In classical test theory, reliability is defined as the ratio of the true score variance to the total score variance and total score variance in an observed distribution of scores equals the sum of the true variance (variance from true differences) plus the error variance (variance from irrelevant sources of variability). The greater the proportion of the total variance attributed to true variance, the more reliable the test (Cohen & Swerdlik, 2002). The value of a reliability estimate, expressed as r , will always range between 0 and 1. If a measure is perfectly reliable, there is no error in the measurement and everything observed is a true score, then the reliability of the scores will be equal to 1. If a measure is perfectly unreliable, the measure is entirely error and there is no true

score, the reliability of the scores will be equal to 0. For instance, a score reliability (r^2) of .5 means that about 25% of the variance of the observed score is attributable to error. A score reliability (r^2) of .8 means the variability is about 64% true ability and 36% error (Trochim, 1999).

However, there is a flaw in this concept of reliability; if a group of respondents to a normally reliable instrument are homogenous (i.e., similar) then variability in their true scores is attenuated making the measuring instrument appear unreliable although it is providing precise and accurate measurements. Likewise, the reliability of a measuring instrument can be inflated by sampling only “extreme” raters. In other words, artificially enhancing the variability of the people within a group will make r bigger. Still, the role of homogeneity in attenuating r does not make it a useless parameter; instead it tells the researcher about the disposition of the group of respondents. It might indicate that there are “true” differences in perspective among the group. Of course, one can never eliminate all the effects of measurement error, but by investigating measurement error and applying good design its potential influence can be minimized and can provide a better answer to the usefulness of a measure (Cronbach, 1990; Joint Committee, 1999; Linn & Gronlund, 2000; Murphy & Davidshofer, 2001).

Survey construction. Survey objectives and methods are diverse. Some have far-reaching uses, while others meet very specific needs (Fink & Kosecoff, 1985). The development of a new measure may be in response to a new research frontier or to address a practical problem. Many instruments are designed to measure a hypothetical construct (i.e., anxiety, intelligence) that are measurable only to the extent that they are linked to observable behavior. Therefore, it is evident that the first step to designing a

survey is to clearly define the topic of study. Behaviors related to the construct that are not defined properly may cause results to be confusing and difficult to interpret, which may lead to measurement error (Bordens & Abbott, 1999; Cohen & Swerdlik, 2002).

Another source of error variance during survey construction is item sampling. This step involves generating an *item pool*, or a reservoir of items to be drawn from or discarded on the final version that comprehensively samples the content domain. A pool of items may be developed by writing a large number of items from personal experience, interviews with experts in targeted industries, or through research literature (Cohen & Swerdlik, 2002). Some survey developers use the technique of drawing on one particular theory and translating the ideas of that theory into questionnaire items (Murphy & Davidshofer, 2001). According to Fowler (1993), a good question has the following properties: 1) the question-and-answer process is entirely scripted so that the questions, as written, fully prepare the respondent to answer, 2) the questions mean the same thing to every respondent, and 3) the kinds of answers that elicit appropriate responses to the questions are communicated consistently to every respondent. However, two separate assessment tools measuring the same skill would differ in the way the items are worded and the exact content sampled. The extent that the respondent's score is affected by these extraneous factors is a source of error variance (Cohen & Swerdlik, 2002).

In addition to item construction, there are some common issues that survey developers need to take into account when writing items such as the item length, the vocabulary used in the item (i.e., simple, neutral, universally understood words), implying the "correct" answer, and the presence of sexist, racial, or offensive language (Murphy & Davidshofer, 2001). Also, the various formats available to respond to an

item (i.e., constructed-response format, selected-response format, ranking, and rating scales) and their related advantages and disadvantages present a challenge for any survey author (Cohen & Swerdlik, 2002).

Rating scale. Rating scales are used extensively in psychological assessment, educational assessment, and other professional settings to record a graded response to a question. Basically, the respondent indicates the strength of a particular trait, characteristic, or attribute through a group of words, statements, or symbols meaningfully categorized along a continuum. A score can be calculated when numbers or other indices are assigned to different amounts of the trait, characteristic, or attribute being measured. For example, a score reveals that an individual is thought to have more or less of the characteristic measured by the test and the higher or lower the score, the more or less of the characteristic they are thought to possess. Since developers scale an instrument to optimally fit their conceptualized measurement of the target trait there is no one method of scaling or a preferred type of scale (Cohen & Swerdlik, 2002). For instance, the Likert-type scale (1932, as cited in Cohen & Swerdlik, 2002) is used extensively in psychology, particularly in attitude or opinion measurement, and is usually reliable. Each item presents a respondent with as few as three or as many as 10 labeled alternative responses, usually indicating degrees of magnitude along a continuum. Respondents then are asked to consider the labels, consider their own attitude or opinion or someone else's, and place themselves or others in the proper categories. A 10-point scale allows a wide range of choice without overburdening the respondents, especially if they avoid the extreme values (Cohen & Swerdlik, 2002). However, scales with five to seven points are used more frequently since using more scale values adds no more precision (Landy &

Farr, 1980). Weights are usually assigned to the different categories, a 1 for endorsement of items at one extreme, through 5 for endorsement of items at the other extreme. For example, if the response “bad” is assigned the value 1, “poor” the value 2, “fair” the value 3, “good” the value 4, and “excellent” the value 5, the higher the score the more the response is indicative of excellence. However, the difference between the attitudes of respondents who scored 2 and 3 on the scale is not necessarily the same as the difference between their score values of 3 and 4 or the attitudes of other respondents with the same scores. All that is known for sure is that 3 is greater than 2 and “good” is greater than “fair” (Cohen & Swedlik, 2002). *Ordinal scales*, like these, therefore, do not imply equal intervals between the numbers. Only true measures of distance are possible between the numbers for *interval scales*; the Fahrenheit and Celsius scales have these properties. For example, a temperature of 40 degrees is higher than a temperature of 30 degrees, and an increase from 20 to 40 degrees is twice as much as an increase from 30 to 40 degrees (Murphy & Davidshofer, 2001). Interval scales reach a level of measurement in which it is possible to take the average of a set of scores and obtain a meaningful result, however, such scales are difficult to construct (Cohen & Swedlik, 2002; Murphy & Davidshofer, 2001). Since most psychological research does not meet the formal requirements for interval level measurement and since Likert-type scales only yield ordered data, researchers treat such measures as if they had full interval-scale properties maintaining there is some relative measure of equal distance between the ratings (Bordens & Abbott, 1999; Questionnaires, n.d. What are questionnaires and when can they be used?; Suskie, 1992). Tinsley and Weiss (1975) suggested interval-scale statistics could still be

applied to ordinal data without distortion as long as the assumption of equal intervals is not greatly inappropriate.

The 1970s and 1980s saw a significant upsurge of research on rating techniques and a standardization of definitions and procedures (Anastasi, 1988). Landy and Farr (1980) reviewed 30 years of extensive research literature on scale formats and concluded that researchers have probably gone as far as they can in improving rating formats. Essentially, current rating scale formats have relatively minor impact on the outcomes of ratings. Although respondents may have a preference for various physical arrangements (i.e., high and low anchors, graphic numbering system), these preferences seem to have little effect on rating behavior. Invariably, research has increasingly focused on the judgment processes that underlie rating scale formats particularly in performance evaluation and subsequent rater-training programs. So today, the question of who evaluates has become more important than the type of scale used (Murphy & Davidshofer, 2001).

Rater error Computer scanning and scoring programs have virtually eliminated any error variance previously due to the scoring processes. However, not all instruments are on paper and require a number 2 pencil. Some measures are anything but an objective, reliable, computer-scoring instrument. In these cases, especially where subjectivity is involved, the scoring system or the rater becomes the source of error variance (Cohen & Swerdlik, 2002). Whenever performances, items, or targets are judgementally scored, it is reasonable to ask whether another equally qualified judge would assign the same score (Linn & Gronlund, 2000). The degree of agreement and the consistency that exists between two or more scorers or judges is variously referred to as

inter-rater agreement (IRA) and inter-rater reliability (IRR), respectively (Cohen & Swerdlik, 2002). On this basis, an extensive theory of score reliability has been developed to account for the variability in raters' scores. The goal of estimating reliability is to determine how much variability in scores is due to errors in measurement and how much is due to variability in true scores. Subsequently, systematic errors can be found in many rating situations (Murphy & Davidshofer, 2001).

Types of rater errors. *Restrictions of range rating errors* occur when raters tend to rate all individuals at approximately the same position on the scale. More specifically, *generosity or leniency bias*, as the name implies, reflect raters' tendencies to use the high end of the scale only or in the case of *severity bias*, which occur less frequently, the lower end of the scaled is favored. A third type of response bias by some judges is *central tendency errors* in which the rater avoids both extremes of the scale and tends to rate everyone as average (Linn & Gronlund, 2000). *Logical errors* occur when the rater assumes there is a more direct relationship among traits than there actually is. For example, a teacher may tend to overrate achievement of a student who is gifted because they expect achievement and giftedness to go together or a teacher that underrates them on social abilities because they believe gifted students have poor social skills (Linn & Gronlund, 2000). *Halo errors* reflect raters' tendencies to allow their overall evaluation of an individual to affect the evaluation of each specific aspect of that person's performance. For instance, if a rater has a favorable attitude toward the person being rated there will be a tendency to give high ratings on all traits, but if the rater's attitude is unfavorable, the ratings will be low. It is unlikely that most people are very good, very bad, or even average in all aspects of their performance. Therefore, a judge who rates

individuals as uniformly good or bad is not providing accurate ratings. For instance, a high or low rating might reflect the personal outlook of the rater rather than the actual performance or personal characteristic of the individual being rated. Or, the ratings of different individuals may be so close together that they fail to discriminate between the targets themselves or their strengths and weaknesses on different traits (Linn & Gronlund, 2000; Murphy & Davidshofer, 2001).

Survey administration Measurement error can also occur during administration of even a well-developed survey. People vary from day to day in terms of attention or motivation, mood, attitude, or effects of drugs, or their degree of anxiety, emotional stress, or physical health. Human nature creates individual variability when data are collected over one occasion and not another, or on more than one, the same, or different tests. Another source of error variance that can occur during administration is from the test-taking environment itself. Examples of these unfavorable influences include temperature, noise level, lighting, an annoying fly, or a broken pencil. The presence or absence of an examiner in the room and whether or not they follow the procedures set for a particular instrument are also potential sources of error variance. Essentially, any event or nuance during or across administrations that may alter individual responses is categorized as measurement error. Respondents may intentionally distort their responses, guess at items, carelessly answer items, or mistakenly mark their answer sheet for whatever reason, all of which will also contribute to measurement error (Cohen & Swerdlik, 2002).

Unfortunately, the effects of these conditions discussed here cannot be removed from observed scores and even more importantly, cannot be overcome by statistical

analysis. Errors of measurement are generally viewed as random and unpredictable, and therefore, “they constitute a source of construct-irrelevant variance and thus may detract from validity” (Joint Committee, 1999, p. 26). Consequently, information about measurement error and the difference or change in scores is exactly what reliability pertains to and is fundamental to the proper evaluation and use of an instrument (Joint Committee, 1999).

Sources of Validity Evidence

Previously, it was thought that different types of validity were appropriate for different purposes and sometimes incompatible with one another, but today validity is viewed as a unitary concept based on various kinds of evidences. Researchers now agree there are four major interrelated considerations or strategies for validating inferences made on the basis of test scores all with the same basic goal of understanding the meaning and consequences of scores—content, construct, assessment-criterion relationship, and consequence validity evidence (Linn & Gronlund, 2000). In essence, the actual validity of survey information begins with careful scrutiny of the instrument related to its intended use and desired inferences (Qualls & Moss, 1996).

Content validity evidence. Content refers to themes, wording, format, etc. of the items, tasks, or questions on an assessment tool (Joint Committee, 1999). *Face validity* provides a superficial idea of content considerations and adequacy of sampling tasks by looking at the survey items. If all the items appear to measure what the test is supposed to measure, then there is some evidence of content validity (Murphy & Davidshofer, 2001). However, face validity should not be considered a substitute for more rigorous evaluation, either logical or empirical, of content domain and sampling adequacy (Linn &

Gronlund, 2000). Content domain represents the total set of behaviors that could be used to measure a specific attribute of the individuals to be tested (Murphy & Davidshofer, 2001). Linn and Gronlund (2000) defined the essence of content consideration in validation, "...to determine the extent to which a set of assessment tasks provides a relevant and representative sample of the domain of tasks about which interpretations of assessment results are made" (p. 78). Assessing content validity is difficult to implement because it lacks an exact statistical measure. Instead, content validity is a judgment regarding the degree to which an instrument provides an adequate sample of the tasks for a specific content domain. Murphy and Davidshofer (2001) provided a method for systematic evaluation of the content validity of measurement. First, the content must be described. Second, the areas of the content domain that are measured by each item must be determined. And the final step in assessing content validity is to compare the content and structure of the measure with the structure of the content domain. For example, if test items are concerned with only a small portion of the domain, the measure will have little evidence of content validity. Likewise, a measure that appears to provide a representative sample of the major parts of a content domain will be evidence for high levels of content validity. Another technique frequently used in determining content area is observation. Experts or judges may rate the degree to which the content of the tool is a representative sample of the universe of behavior the tool was designed to sample. Lawshe (1975) quantified content validity as a method for gauging agreement among raters or judges regarding how essential a particular item is. For each item, the raters state whether it is essential, useful but not essential, or not necessary. If more than half the raters indicate that an item is essential, that item has at least some content validity.

As more of the raters agree that a particular item is essential, evidence of content validity increases. Lawshe developed a formula to calculate this index of validity called *content validity ratio*, or *CVR* (Cohen & Swerdlik, 2002). Lawshe's CVR is represented at the item level as:

$$CVR = \frac{k_e - K/2}{K/2} \quad [3]$$

where k_e is the number of raters indicating the item essential and K is the total number of raters in the panel (Lindell & Brandt, 1999). A study of content validity evidence cannot assure the validity of measurement, however. Although a researcher may be able to establish that the domain is well understood, it is a representative sample from the domain, and the items are the best type of items sampled, the wording, for instance, may still be confusing or the response scales may be improper (Murphy & Davidshofer, 2001).

Construct validity evidence. A good measure of a specific construct is complex because constructs are often unobservable and based on an assumed, theoretical construction used to explain some aspect of behavior (Linn & Gronlund, 2000; Murphy & Davidshofer, 2001). Constructs represent abstract ideas of some regularity in nature and they can be summarized in a real, observable group of related events, phenomena, behavior, or objects. Murphy and Davidshofer (2001) provided an excellent description and example of a construct featuring gravity. When apples fall to the ground the construct gravity is used to explain and predict the apples' behavior. It is impossible to see gravity; the only thing seen is the apple falling. Still, gravity is measured and theories are developed about this abstract force. A construct such as gravity is related to a number

of concrete objects and events. Once all is learned about gravity, a wider variety of phenomena may be predicted and generalizations can occur from an experiment involving falling apples to situations involving other falling objects. Happiness, intelligence, anxiety, etc. are other abstract attributes that researchers are interested in measuring. These things do not exist in the physical sense (e.g., vials of happiness cannot be collected); regardless, they must be measured in order to further theories. Constructs are not limited to unseen forces or learning processes. Any real thing or event that is related, either directly or indirectly, to behavior or experience may be defined as a construct (Murphy & Davidshofer, 2001, pp. 154-155). To determine if an assessment tool is a good measure of a specific construct, the abstract construct must first be translated into concrete, behavioral terms. One way of systematically describing constructs is to identify behaviors that may relate to the construct under consideration. Then identify other constructs that may or may not be related to the construct being measured. And finally, identify any additional behaviors related to each of these additional constructs and determine whether each behavior is related to the construct being measured. Although this procedure does not define exactly what a construct is, it does show how that construct relates to a number of behaviors. The more that is known about a construct, the prospects are greater for determining whether an instrument is an adequate measure of that construct. There are a number of methods used to test the predictions based on a description of the construct such as calculating differences between control and experimental groups, patterns of relationships, individual variations in the respondent's process, and changes over time. The most basic method is to correlate scores on the measure in question with scores on other measures (Joint

Committee, 1999; Murphy & Davidshofer, 2001). The scores of any assessment can be expected to correlate substantially with the scores of other measures of the same construct. For example, a high correlation of scores would be expected between two scholastic aptitude tests, but a much lower correlation between a scholastic aptitude and a musical aptitude test (Linn & Gronlund, 2000). Campbell and Fiske (1959, as cited in Murphy & Davidshofer, 2001) proposed a useful technique when a number of methods are used to measure more than one trait or construct; a *multitrait-multimethod (MTMM)* matrix or table is comprised of correlating traits or constructs within and between methods. This type of study provides a vast amount of information for establishing construct validity. First, each construct is measured using a number of different methods to produce comparable sources. If the data are in close agreement (i.e., the different constructs converge to yield similar results) this suggests the constructs under consideration are, in fact, the constructs that the researcher meant to measure. Thus, a high correlation between measures of the constructs demonstrates convergent validity and provides stronger evidence of construct validity. Similarly, correlations between different measures of different constructs should be smaller than the correlations between different measures of the same construct. If the constructs chosen are clearly different, the measures of those constructs will not correlate highly, thus, providing evidence of discriminant validity. Unfortunately, data from a MTMM matrix can be difficult to interpret and requires a large number of separate correlation coefficients. Therefore, more shorthand statistical methods have been proposed including analysis of variance and factor analysis (i.e., exploratory and confirmatory factor analysis) (Cohen & Swerdlik, 2002; Murphy & Davidshofer, 2001). However, Cronbach (1988, 1989, as

cited in Murphy & Davidshofer, 2001) criticized most MTMM methods “as mindless and mechanical; the selection of traits and methods is more often based on convenience rather than on compelling hypotheses” (p. 164).

Both considerations, content domain and construct, can yield evidence that the instrument measures what it is designed to measure. Combinations of content and construct validity represent very strong evidence for the validity of a measurement. Some researchers believe that to some degree most content categories are constructs (Murphy & Davidshofer, 2001). However, Murphy and Davidshofer (2001) offered this fundamental distinction, “Content validity is established if a test *looks* like a valid measure; construct validity is established if a test *acts* like a valid measure” (p. 166).

Assessment-criterion relationship validity evidence. Criterion validity, also referred to as “evidence based on relations to other variables”, involves analyzing the relationship of scores to some valued measure other than the assessment itself, called a *criterion* (Joint Committee, 1999, p. 13). These external variables or criteria may include some criteria that the instrument is expected to predict, as well as relationships to other instruments assumed to measure the same or different constructs. Essentially, the focus of criterion-related validity is the degree to which scores obtained at a later time can be used to predict criteria or the correlation between data and criteria information obtained at the same time (Joint Committee, 1999; Murphy & Davidshofer, 2001). For example, prediction occurs when a supervisor estimates a potential hire’s score on a measure of job performance based on their score on some other measure, such as a computer literacy test. Thus, an applicant with high test scores is predicted to perform well on the job. It is unknown how the new hire will actually perform; therefore, the better the test, the more

accurate the predictions, and the more correct decisions. The predictions are validated if the worker hired actually does perform at a higher level than those who were not hired. Although a predictive validity strategy is considered an ideal strategy for estimating validity, it is usually not realistic or timely. For instance, the performance measures for those persons hired are obtained at some later date in order to correlate these measures with initial scores, which were obtained before making the hiring decision. Another objection to a predictive validity study is that the population in the study must be similar to the general population of applicants. Thus, managers previously making decisions on a random basis would be forced to accept a system of selection based on those applicants with low test scores who are more likely to fail (Murphy & Davidshofer, 2001). A practical alternative with sufficiently similar outcomes to predictive validity study is the *concurrent validation strategy*. In this approach, both test scores and criterion scores are obtained at the same time and the correlation is calculated between the two. Correlations between test scores and criterion measures in a highly selective population selected according to their test scores are used to estimate the validity of a test (e.g., pre-selected sample of present employees already performing at acceptable levels). However, restrictions of range errors, inability to discriminate performance (i.e., superior performances from acceptable performances of applicants), and differences in the populations under consideration are just some of the problems with most concurrent studies (Murphy & Davidshofer, 2001).

Consequence validity evidence. Messick (1989) argued persuasively that the social (i.e., individual, institutional, or systemic) consequential basis of test use and interpretation should also become an integral part of validity. Determining if a measure

does what it is designed to do requires an evaluation of the intended and unintended social effects of its interpretation and use. Especially when results are used in high-stake decisions, the appropriateness of the intended purpose and the potential negative outcomes are major issues in the overall judgment of validity (Linn & Gronlund, 2000; Messick, 1989). Linn and Gronlund (2000) contended that the expansion of validity to include evidence based on consequences of use and interpretation of assessment results has been important in the recent movement toward more authentic alternative forms of assessment. For instance, proponents argue that a heavy reliance on standardized tests year after year, especially in state or school districts where students and teachers are being held accountable for results, has produced a number of unintended negative effects. The high-stakes associated with the test results lead teachers to focus narrowly on what is on the test and ignoring other important, untested parts of a curriculum. In addition to “teaching to the test”, scores are likely to rise over time changing the meaning of the results and the construct being measured (e.g., from mathematical concepts to memorization ability). Measures are designed with the hope that some benefit will be realized from their intended use. Therefore, a fundamental purpose of validation is to determine if these specific benefits are likely or not. Evidence about these consequences is collected, and in the case of standardized tests, the use of educational tests will improve student motivation or encourage changes in classroom instructional practices. Such claims are key aspects to the direct examination of consequences regarding validity. The validation would be confirmed by evidence in support of that claim (Joint Committee, 1999; Linn & Gronlund, 2000). Messick (1989) warned, “Even if adverse testing consequences derive from valid test interpretation and use, the appraisal of the

functional worth of the testing in pursuit of the intended ends should take into account all of the ends, both intended and unintended, that are advanced by the testing application” (p. 85).

The most powerful case can be made for validity if evidence is obtained regarding all four of these considerations. Although, one consideration may be of primary importance, an understanding of the other three are useful for greater validity and contribute to the validation of the meaning of assessment results and their interpretation (Linn & Gronlund, 2000). Many survey researchers consider reliability to be a part of validity, therefore, one cannot be discussed without the other. In addition, many survey researchers claim an assessment that produces totally inconsistent results cannot possibly provide valid information. Thus, full validation requires examination of the differences in scores to reveal important information about the components of disagreements and inconsistencies. The purpose is to better understand why scores differ toward the ultimate goal of improving their consistency (Linn & Gronlund, 2000; Uebersax, 2003).

Methods of Estimating Reliability

According to the *Standards for Educational and Psychological Testing* (1999), no developer is exempt from the responsibility of fully investigating reliability using the most ideal approach, which entails independent replication of the entire measurement process. Reliability refers to the changes or differences in the scores of such measurements when repeated on a population of individuals or groups. However, as a practical matter in many testing situations, replication is not possible. It is often logistically difficult, time-consuming, and expensive to attempt to administer the same instrument to the same individuals twice only to establish the reliability of scores often

departing from the original intent of the study (Joint Committee, 1999; Murphy & Davidshofer, 2001). Additionally, reactivity and carryover effects may occur in this type of test-retest method. Reactivity is when the experience of taking the test can change the individual's true score. For example, students who take a spelling test may look up the correct spelling of the words they were unsure of after taking the test resulting in a substantial change in true scores at the second administration. Also affecting true scores on the second test are carryover effects, when individuals recall their original answers and record the same pattern of right and wrong answers (Murphy & Davidshofer, 2001).

Therefore, some other broad categories of reliability have been recognized: 1) estimates derived from the administration of parallel forms in independent testing sessions (i.e., alternate-forms), 2) estimates obtained by administration of the same instrument on separate occasions (i.e., test-retest), 3) estimates based on the relationships among scores derived from individual items or subsets of the items within an instrument, all data accruing from a single administration (i.e., internal-consistency), and 4) estimates derived from scores that involve observations of behaviors or performances, evaluations of products, or requires a high level of judgment (Joint Committee, 1999). As one can see, there is more than one approach to estimating reliability and no single method of investigation can cover all situations and all relevant facts. Yet, the *Standards for Educational and Psychological Testing* (1999) ruled, "The reporting of reliability coefficients alone, with little detail regarding the methods used to estimate the coefficient, the nature of the group from which the data were derived, and the conditions under which the data were obtained constitutes inadequate documentation. General statements to the effect that a test is "reliable" or that it is "sufficiently reliable to permit

interpretations of individual scores” are rarely, if ever, acceptable” (p. 31). Basically, researchers are obligated to provide potential test users as well as consumers with sufficient data to make informed decisions about the confidence that can be placed in any measurement and the measurement process (Joint Committee, 1999).

However, evaluations such as these are the exception, rather than the norm (Fowler, 1993). Qualls and Moss (1996) examined the extent that researchers complied with these guidelines in published research. All too often, reliability and validity evidence supporting the use of a particular instrument was not reported. Score reliability information was reported for 41% of the instruments analyzed and validity information was offered for only 31.7%. Additionally, in more cases than not, score reliability was reported in the absence of validity evidence. In a later study, Whittington (1998) identified some of the most common reporting failures related to reliability and validity in educational research literature. He found that 61% failed to report any reliability evidence, 79% failed to report information about the development and/or piloting of new measures, and 45% failed to report evidence of inter-rater reliability of scores for open-ended measures or observations. And finally, Hogan, Benjamin, and Brezinski (2000) encountered a number of problems in their attempt to examine the frequency of use of various types of reliability in the APA-published *Directory of Unpublished Experimental Mental Measures*. In some instances the reliability reported in the directory was not based on the study cited but on some other source. Situational specificity of a measure also applies to score reliability just as it does for establishing validity. The authors warn, “The user of the directory might assume that the reliability is based on the article cited in the directory, but this is not a safe assumption” (pg. 66). Reporting adequate information

about a measure's score reliability allows the user to judge the adequacy of its results. Since most research is built on the work of others, it introduces a foundation of questionable evidence into studies that are used to inform decisions of great magnitude such as how pilots are trained to fly airplanes or how to identify life-threatening illnesses (Anastasi, 1988; Whittington, 1998).

Classical Test Theory vs Modern Test Theory: Approaches to Reliability Estimation

While traditional methods of estimating reliability of scores classify scores into two components: true scores and random errors of measurement, modern theories of measurement have given researchers powerful new ways of thinking about score integrity and the use of those score. Essentially, *generalizability theory*, or “*G*” *theory*, as delineated by Cronbach, Gleser, Nanda, and Rajaratnum (1972, as cited in Thompson, 2003b), has merged the distinctions between score reliability and score validity into a single concept which scrutinizes whether the models of measurement actually match the models of reality (Brennan, 2001; Thompson, 2003b). For instance, consider the classic test-retest method and the classic internal-consistency method, which estimates reliability using one measurement at two points in time and one measurement at a single point in time, respectively. It begs the question of whether or not a researcher should limit the interpretation of scores to only one set of items at only two points in time or even one point in time (Thompson, 2003b). Cronbach et al. (1972, as cited in Thomson 2003b) explained, “The score...is only one of many scores that might serve the same purpose. The [researcher] is almost never interested in the response given to a particular stimulus object or questions, to the particular tester, at the particular moment of the testing” (p. 57). Brennan (2001) further argued, “...if data are collected on a single occasion, an

estimate of reliability based on such data will almost certainly overestimate reliability when interest is in generalizing over occasions” (p. 20). And, this is exactly what researchers want to do, view scores as either adequate or inadequate across any and all forms. In other words, many researchers treat reliability estimates as if they are stable across target samples, administrations, and local situations even though this is not the case. In effect, the classical measurement model does not actually honor the models of reality at all. For this reason, “G” theory is becoming increasingly popular and is likely to replace the more traditional theory of reliability, which has dominated the field for over 70 years (Thompson, 2003a; 2003b).

“G” theory. “G” theory can be thought of as an extension of the ideas in the classic theory, which seeks to estimate the portion of a score that is attributable to error. In “G” theory, the true score is analogous to the *universe score*, which represents the score that is desired depending on the universe being considered. *Facets* in the universe score include such things as the number of items, the amount of training the raters have had, and the purpose of the measure. The influence of these particular facets on a measure’s score is represented by *coefficients of generalizability*, which is the ratio of the universe score to the expected observed variance score. Instead of attempting to conceive of all the variability in a score as error, “G” theory compartmentalizes score variability into a universe score component and various facet components. In other words, by systematically studying the many sources of consistency and inconsistency in scores one can begin to generalize from one set of measures to another set of credible measures. “G” theory makes it possible to separate observed score variance into a number of subcomponents, specifically, measurement error variance into its own variance

subcomponents. The repeated measures ANOVA form of the general linear model provides a method for measuring the systematic effects of several variables on the consistency of scores and suggests that scores can be more accurately generalized over time, over scorers, or to different tests. Measurement reliability is much more a function of the circumstances under which the measure is developed, administered, scored, and interpreted; if scores differ systematically according to when, where, or how the survey was taken, these differences will affect the generalizability of survey scores. They are not, however, random sources of error and the test theories that treat them as such leads to erroneous conclusions. Therefore, “G” theory is useful when the *conditions of measurements* are likely to affect scores or when the scores are to be used for several different purposes. For example, physical conditions of measurement might strongly affect measures such as a color discrimination test taken in a room with fluorescent light than in a room with natural light. Or, systematic differences in human judgement (e.g., some judges may give higher scores to women than men or attractive people may receive more favorable ratings than unattractive people) could be considered conditions of measurement. Also some measures may be more stable over time than others (e.g., measures of basic values as opposed to opinions of a future sporting event). The “G” theory approach recognizes that error is not always random and that it often is useful to identify specific, systematic sources of inconsistency in measurement. It is important to note that when there are no systematic differences in test scores associated with conditions of measurement, classical theory and “G” theory are mathematically identical (Cohen & Swerdlik, 2002; Murphy & Davidshofer, 2001).

Consider the example of a set of graders evaluating a set of essay questions. Any one of the essay items constitutes an admissible condition of measurement for the item facet, and any one of the raters constitutes an admissible condition of measurement for the rater facet. This design involves a sample of n_r raters to evaluate each of the responses by a sample of n_p persons to a sample of n_i essay items and is denoted by $p \times i \times r$ implying that they are all crossed. Thus, each person answers each essay item and each rater grades each essay. Therefore, there are seven sources of variance that can be independently estimated from the data and referred to as “*G*”-study variance components, which are described in Table 2 (Feldt & Brennan, 1989, p. 128).

Source of Variance	Description
Essay items (<i>i</i>)	Some essay items may be more difficult meaning that scores on one item are likely to be higher or lower than scores on other items.
Persons taking test (<i>p</i>)	Individual differences among the test-takers means that some test-takers will score higher or lower than others.
Raters (<i>r</i>)	Differences among the raters may produce systematic differences in scores meaning some raters may consistently assign higher or lower scores than others.
$i \times p$	Certain test-takers may have more knowledge about a given essay item than another item. Thus, a particularly high-performing test-taker may encounter one item which they have little experience or knowledge, producing a lower score on that particular item than on other items.
$i \times r$	A particular rater may be more severe or lenient for a subset of essay items. Thus, the rater is inconsistent across items.
$p \times r$	Stylistic cues of a given test-taker may interact with the expectations of a given rater. Thus, the particular test-taker receives higher or lower scores by a particular rater, not because of domain knowledge, but due to some unique “relationship” between the test-taker and the rater.
Error $i \times p \times r$	A three-way interaction may occur meaning that any one of the above two-way interactions only exists for one or more of the levels of the third facet, or the residual.

Table 2 Description of seven independent sources of variance

“G” theory replaces classical theory’s representation of a person’s observed score assigned by a raters on a particular test item, X_{ipr} , as the sum of his or her true score, T_p , plus an undifferentiated error, e_{ipr} , $X_{ipr} = T_p + e_{ipr}$. In “G” theory, there is variance associated with each component of an observed score, except for the constant μ , an expression used to compute variance components, or the grand mean. The magnitude of the variance components indicates how much each facet contributes to measurement error. If, for example, raters are considered as introducing error, which is denoted by $\mu_r - \mu$, or the rater effect, the score assigned by the raters for each person, which is denoted by $\mu_p - \mu$, or the person effect, is summed over items, which is denoted by $\mu_i - \mu$, or the item effect. So, the general linear model equation for the main effects of X_{ipr} is represented as $X_{ipr} = \mu + (\mu_i - \mu) + (\mu_p - \mu) + (\mu_r - \mu) + e_{ipr}$ (Crocker & Algina, 1996; Shavelson, Webb, & Rowley, 1989).

Furthermore, an interaction effect is when the effects of one independent variable on the outcome variable that is separate from one another and of the main effects influences the outcome variable over the levels of other independent variables. In a balanced design (i.e., an equal number in all of the design cells), all of the main effects and interaction effects will be perfectly uncorrelated with each other, separate, and cumulative. A good example of an interaction effect is a drug interaction. The average ability of a person driving a car may have little or no main effect with ingesting one drug (e.g., no antihistamine vs. one dosage of antihistamine). A second drug (e.g., no alcohol vs. one alcoholic drink) may have little or no main effect on an average person’s ability to drive. But, taking an antihistamine and drinking one alcoholic drink may have an effect over and beyond the sum of the two main effects separately (Thompson, 2003b, pp.

45-47). Yet, classical test theory can only create a single estimate of measurement error variance at a time and never estimates the interaction effects as if they never occur or are irrelevant. Conversely, ANOVA interaction effects are critically important in research and usually are of more interest than the main effects themselves. Simultaneously considering the combined effects of say, drugs and driving (i.e., their interaction effects) often results in significant variability that calls into question credibility of all scores if the variance cannot be partitioned into separate parts (e.g., Drug 1 on driving ability, Drug 2 on driving ability, interaction, and error) (Thompson, 2003b). Putting this in a real world context, Cronbach (1957, as cited in Thompson, 2003b) strongly recommended that researchers focus on identifying which treatments work best for different people as opposed to the main effects results that work best for everyone. Recognizing that traditional techniques are inadequate in many situations, the *Standards for Educational and Psychological Testing* (1999) even suggested the following regarding the advantageous use of this alternative approach to measurement:

Measurements derived from observations or behavior or evaluations of products are especially sensitive to a variety of error factors. These include evaluator biases and idiosyncrasies, scoring subjectivity, and intra-examinee factors that cause variation from one performance or product to another. The methods of generalizability theory are well suited to the investigation of the reliability of scores on such measures. Estimates of the error variance associated with each specific source and with the interactions between sources indicate the extent to which examinee scores may be generalized to a population of scorers and to a universe of products or performances. (Joint Committee, p.29)

Whereas, the researcher conducting a “G” study is primarily interested in the extent that a sample of measurements generalizes to a universe of measurements, a decision study, or “D” study is conducted to collect data specifically to make a decision (Crocker & Algina, 1996). Basically, the results from estimates of “G” study variance

components for one or more measurement procedures are used in a “D” study, which then uses the measurement procedures to collect data about the *objects of measurement*, or the universe to which decision makers want to generalize (Feldt & Brennan, 1989).

Cronbach (1970) explained why this is so important:

An erroneous favorable decision may be irreversible and may harm the person or the community. Even when reversible, an erroneous unfavorable decision is unjust, disrupts the person’s morale, and perhaps retards his development. Research, too, requires dependable measurement. An experiment is not very informative if an observed difference could be accounted for by chance variation. Large error variance is likely to mask a scientifically important outcome. Taking a better measure improves the sensitivity to of an experiment in the same way that increasing the number of subjects does. (p. 152)

Unfortunately, since its original publication over 30 years ago, “G” theory remains virtually unused. Directly related to its comprehensiveness is its complexity, which makes it more time-consuming and expensive to conduct as compared to simpler approaches to reliability estimation (Goodwin, 2001).

Scale-Dependent Metrics: Internal Consistency Reliability Coefficients

Within classical test theory, internal consistency reliability estimates are the most commonly used reliability estimate of scores from a single administration of a single instrument on one sample (Henson, 2001). Hogan, Benjamin, and Brezinski (2000) noted that internal consistency estimates were reported most often (75%) in a review of the APA published *Directory of Unpublished Experimental Mental Measures*. Fourteen years after the introduction of the *Kuder-Richardson formula 20* (1937, as cited in Anastasi, 1988), which is a split-half reliability estimate for determining the inter-item consistency of dichotomous items (i.e., items that can be scored right or wrong such as multiple-choice items), Cronbach (1951, as cited in Henson, 2001) proposed a variant known as *coefficient alpha* or *Cronbach’s alpha* (α) (Cohen & Swerdlik, 2002).

Subsequently, the more generalized formula is appropriate for use on tests containing any form, including both dichotomous and nondichotomous items (i.e., items that can individually be scored along a range of values such as a Likert-type scale) (Thompson, 2003a). Whereas split-half methods compare one half-test to another, internal consistency estimates compare each item to every other item. This method estimates the reliability of a test based on the number of items and the average intercorrelation among test items (Murphy & Davidshofer, 2001). This “item interrelationship” is called internal consistency, which suggests that the items should be highly interrelated because they assess the same construct of interest (Henson, 2001, p. 180). In essence, if the items are highly correlated (i.e., the items measure the same thing as all other items) then it is assumed that the scores on the test are highly reliable and the construct of interest has been measured to some degree of consistency (Henson, 2001; Murphy & Davidshofer, 2001). Furthermore, when the value for alpha is higher upon deleting an item it is inferred that the item is not tapping the same construct as all the others and should be removed from the test. More specifically, by holding the number of items constant, reliability will increase as the sum of item variances decreases and the total score variance increase. The formula for Cronbach’s alpha (α) is:

$$\alpha = \left(\frac{k}{k-1} \right) \left(1 - \frac{\sum s_k^2}{s_{TOTAL}^2} \right) \quad [4]$$

where k is the number of items on the test, $\sum s_k^2$ is the sum of all the k item variances, s_{TOTAL}^2 is variance of the total test scores (Henson, 2001). Consider the example in which 10 people ($p=10$) respond on three different occasions ($o=3$) to a 4-item attitude

measure using a nine-point Likert-type scale ($i=4$). Using only classical test theory, three different Cronbach's alphas would be computed, one for each of the three occasions. The resulting coefficients might indicate that only 10% of the observed score variance is measurement error. Or, at the other extreme, the resulting coefficients may be very different from one another (Thompson, 2003b). Regardless, classical reliability estimates (i.e., Cronbach's alpha) are separate and cumulative, which does not acknowledge the simultaneous and multiple influences (i.e., the impact of test occasion (o), the items (i), or their interaction ($o \times i$)) on any given measurement (Henson, 2001; Shavelson, Webb, & Rowley, 1989; Thompson, 2003b). If unable to estimate each potential source of error, how does one determine, in order to achieve a desired level of reliability, whether to alter the number of occasions, or the number of items, or a combination of both (Shavelson, Webb, & Rowley, 1989)? Henson (2001) added, "As an aside, generalizability theory allows for the simultaneous examination of these sources of error as well as the interactions between them by using analysis of variance methodology" (p. 182).

Rater-Dependent Metrics: Methods of Estimating Inter-rater Agreement and Reliability

Certain types of tests leave much of the judgement to a rater, in which case there is as much a need for a measure of rater reliability as there is for the more commonly used reliability coefficients (Anastasi, 1988). A common misconception in measurement is that IRA and IRR mean the same thing. Currently, researchers still tend to use the terms synonymously. Basically, IRA is the extent to which different judges assign exactly the same scores to the rated subject whereas IRR is the extent to which different judges order the ratings in the same way (Goodwin, 2001; Tinsley & Weiss, 1975). IRR can be thought of in terms of the same general pattern of the ratings between raters (Dwyer,

1986). The question of consistency or “reliability” is closely related to consensus or “agreement”, yet, if two judges both exactly agree, it does not automatically imply they are consistent. Conversely, if two judges have high consistency, it does not automatically imply that they have reached consensus (Kenny, 1991; Stemler, 2004). IRA is based on the assumption that two judges with exact agreement are using the rating scale in the same manner to score performances, items, or targets, and therefore, share a common interpretation of the construct. IRR is based upon the assumption that it may not be necessary for two judges to share a common meaning of the rating scale, but are consistent in their own definition of the scale. For example, if Judge A assigns a score of 3 to a set of essay questions and Judge B assigns a score of 1 to the same set of essays, then the two judges are not in consensus, however, the difference in how they use the rating scale is predictable (Stemler, 2004).

	Case 1: High IRA and high IRR			Case 2: Low IRA and high IRR			Case 3: High IRA and low IRR		
	Raters			Raters			Raters		
Counselor	1	2	3	1	2	3	1	2	3
A	1	1	1	1	3	5	5	4	4
B	2	2	2	1	3	5	5	4	3
C	3	3	3	2	4	6	5	4	5
D	3	3	3	2	4	6	4	4	5
E	4	4	4	3	5	7	5	4	3
F	5	5	5	3	5	7	5	5	4
G	6	6	6	4	6	8	4	4	5
H	7	7	7	4	6	8	5	5	4
I	8	8	8	5	7	9	4	5	3
J	9	9	9	5	7	9	5	5	5
\bar{X}	4.8	4.8	4.8	3.0	5.0	7.0	4.7	4.4	4.1
s_x	2.7	2.7	2.7	1.5	1.5	1.5	.5	.5	.5

Table 3 Hypothetical ratings of accurate empathy illustrating different levels of IRA and IRR for interval-scaled data (Tinsley & Weiss, 1975)

For example, Table 3 shows hypothetical data assuming interval-scale measurement in which three judges rated 10 counselors (i.e., targets) on the Truax and Carkhuff Accurate Empathy Scale (1967, as cited in Tinsley & Weiss, 1975) and \bar{X} is the mean rating of the judge and s_x is the standard deviation. Case 1 shows a set of ratings in which all three judges assign exactly the same ratings to each of the 10 counselors. Therefore, these ratings have both high IRA and high IRR. Case 2 represents a set of ratings with low IRA and high IRR. IRA is low because no two judges gave the same rating to a counselor. This lack of agreement is also reflected in their mean ratings, 3, 5, and 7, respectively. However, the ratings assigned to the counselors are proportional resulting in a high IRR even though the raters differed in their ratings of the counselors. As seen here, high reliability is no indication that the raters agree in an absolute sense and low reliability does not indicate that the raters are in disagreement. In Case 3, the high IRA is because the ratings of three raters are similar for 7 of the 10 counselors. Tinsley and Weiss (1975) suggested the IRR is low because of the restricted range of ratings given by the three raters (i.e., the variability of the ratings is small) and these ratings may have occurred because the counselors were highly similar in accurate empathy or the raters misinterpreted the rating scale. Furthermore, they indicated, in conjunction with a high IRA, the possibility exists the subjects may be homogeneous on the trait of interest and vice-versa (i.e., a low IRA means the raters do not necessarily regard the trait of interest in the same manner). This can be investigated in a secondary study by having the judges rate a sample of subjects known to be heterogeneous on the trait of interest. Tinsley and Weiss (1975) stated valid evidence of IRA and IRR is necessary before these ratings can even be accepted. The careful and responsible researcher should assess rater

effectiveness before applying their ratings to their study. As well as take steps to appropriately correct for discrepancies including the need for rater training and clearer specifications of the operational meanings of each point on the rating scale (Goodwin, 2001; Stemler, 2004). However, Tinsley and Weiss (1975) advised that ratings are of no value and should not be applied to research when IRA and IRR are low since the issue of generality is crucial in demonstrating that the obtained ratings do not reflect the raters' idiosyncratic biases (Tinsley & Weiss, 1975). On the contrary, lack of IRA and IRR can still be very informative along the lines of generalizability. A set of properly trained judges responding to a well-developed instrument with a well-defined construct may still disagree reflecting their legitimate differences of opinion on the trait of interest (S. M. Hall, personal communication, June 18, 2004). As these different judges possess different policies on the same topic they cannot be simply trained to come to consensus (Stemler, 2004). However, once group membership is accounted for (i.e., groups of common constituents), internal agreement should be higher. Although, even a low IRA within a constituent group may not be a function of poor survey construction, but a function of their truly collective ideological differences (S. M. Hall, personal communication, April 13, 2005).

Among researchers there is little consensus about what statistical methods are best to analyze rater reliability. Saal, Downey, and Lahey (1980) reported that there is even less agreement regarding their conceptual and operational definitions. These discrepancies lead to confusion whereby different researchers use different designs and data collection procedures making it difficult to suggest a straightforward statistical index of rating quality (Saal, Downey, & Lahey, 1980). However, there are usually one or two

methods best for a particular application and each of the various approaches can yield quite different results when applied to the same data. The choice of approach depends on a number of factors such as the scale of measurement (i.e., nominal, ordinal, interval, ratio), and the number and kind of sources of measurement error that the researcher wants to isolate (Goodwin, 2001; Stemler, 2004).

Inter-rater Agreement (IRA)

Indices for measuring IRR have been available for many years; however, indices for assessing agreement have only recently become available (Lindell, Brandt, & Whitney, 1999). More often today researchers in the fields of applied psychology, industrial/organizational psychology, and management are requiring knowledge of IRA as multiple raters evaluate characteristics of a single or multiple targets, such as task items for a job, team leadership, or an organization (Dunlap, Burke, & Smith-Crowe, 2003). As previously discussed, Lawshe's (1975) underlying rationale regarding the essentiality of an item suggests that *content validity ratio*, or *CVR* is actually a measure of IRA. However, today there are better alternative measures of IRA than CVR (Lindell, Brandt, & Whitney, 1999). One such estimation of IRA that is fairly straightforward is *percentage agreement*. Two or more raters must independently score a target then the percentage agreement is obtained by summing the number of times both raters assigned the same score, dividing that sum by the total number of observations, and multiplying the result by 100 (Linn & Gronlund, 2000). Although percentage agreement is easy to compute and easy to interpret it can be misleading. If one judge assigns "pass" to 100 percent of the targets, and another judge assigns "pass" to only 80 percent of the targets, then the percent agreement will be 80 percent. However, because there is no variance in

the first judge's ratings, there is no statistical relationship between the two sets of judges' ratings (Brannick, 2003). In other words, it does not account for chance agreement, which is the amount of agreement that is expected if the raters randomly assigned values on the rating scale. Additionally, percentage agreement treats agreements as all-or-none and views serious differences between ratings as minor. Basically, percentage agreement does not take into consideration that ratings of 4 and 5 may be more in agreement than ratings of 1 and 5. Also, since percentage agreement relates to the total item pool a disagreement on four items out of 100 is 96%, which is substantially different than disagreement on four items out of 10, or 60%. And finally, if there are restrictions of range rating errors the level of agreement will be inflated with this method (Dwyer, 1986).

A widely used statistic that provides an estimate of the amount of agreement due to chance is Cohen's kappa (k). Although this procedure is often reported in the literature for calculating IRA, it applies to nominally-scaled data only (Dwyer, 1986; Tinsley & Weiss, 1975). However, since many observational instruments are nominal-level and due to its popularity an explanation is presented here. The equation for k is:

$$k = \frac{P(A) - P(E)}{1 - P(E)} \quad [5]$$

where $P(A)$ is the proportion of times the judges actually agree and $P(E)$ is the proportion of times the judges are expected to agree by chance, calculated as:

$$P(E) = \sum_1^J p_j^2 \quad [6]$$

where p_j is the proportion of judgements in each category, which is determined by the total number of judgements, for example, in general there are K judges, N targets and m

categories, and if $K = 3$ judges and $N = 10$ targets, for a total of KN or 30 judgments or, C_j/KN (Brannick, 2003). Coefficient k can range from -1.0 to 1.0 . If raters agree at the same or a lesser level as is expected by chance alone, k will equal zero; and if the agreement exceeds the expected chance level, k will be greater than zero; and if perfect agreement is found k approaches 1.0 . A related statistic is *weighted kappa*, or k_w , which allows some disagreements to be “weighted” as if the raters partly agreed unlike k , which treats all disagreements as equally serious (Dwyer, 1986; Goodwin, 2001). Also, Fleiss (1971, as cited in Dwyer, 1986; Goodwin, 2001) revised Cohen’s kappa to estimate IRA when ratings are performed by more than two judges, or different sets of judges, but the number of judges must be the same for each subject. Basically, the computational formula is the same as k , but the calculation for $P(A)$ and $P(E)$ are different and are somewhat more complicated (Dwyer, 1986; Tinsley & Weiss, 1975).

There are other methods for calculating IRA often reported in the literature such as pairwise correlation, various chi-square tests, and Kendall’s coefficient of concordance, however, noted researchers (Cohen, 1960; Dwyer, 1986; Lu, 1971; Robinson, 1957, as cited in Tinsley & Weiss, 1975; Tinsley & Weiss, 1975) have argued that these methods do not have any merit. Otherwise, Tinsley and Weiss (1975) describe two measures of IRA permitted for ordinal/interval-scaled data, Lu’s (1971, as cited in Tinsley & Weiss, 1975) coefficient of agreement and Lawlis and Lu’s (1972) chi-square and T index (Tinsley & Weiss, 1975). Lu’s coefficient of agreement assumes that agreement varies along a continuum from absolute agreement to no agreement and that a disagreement of four points is more serious than a disagreement of one point. However, this method for calculating IRA is not often used since it is not as flexible as Lawlis and

Lu's chi-square and T index. Lawlis and Lu's approach allows the researcher to set an agreement criterion depending on the seriousness of the rating differences (i.e., ratings within two points of each other ($r = 2$), ratings within one point of each other ($r = 1$), or ratings that are an exact match ($r = 0$)). If this definition of agreement is changed the IRA results will be vastly different. First, a nonparametric chi-square is performed to test the significance of IRA. A chi-square that is not significant suggests that the scale is questionable and a significant chi-square suggests that the observed agreement exceeds the chance agreement. The chi-square test of significance is as follows:

$$x^2 = \frac{(N_1 - NP - .5)}{NP} + \frac{[N_2 - N(1 - P) - .5]}{N(1 - P)} \quad [7]$$

where N_1 is the number of agreements, N is the number of individuals rated, P is the probability of chance agreement on an individual, which is directly related to the range of acceptable ratings (r) and inversely related to the number of scale categories (A) and raters (k), $P = (1/A)^{k-1}$. To allow for a correction of continuity .5 is included in the formula, and N_2 is the number of disagreements. If the chi-square is significant then a T index is calculated to assess the magnitude of IRA. The T index is calculated as:

$$T = \frac{N_1 - NP}{N - NP} \quad [8]$$

where N_1 is the number of agreements, N is the number of individuals rated, and P is the probability of chance agreement on an individual. T is at a maximum when there is perfect agreement (i.e., $N_1 = 1.0$) and the probability of chance agreement is at a minimum. T is at a minimum when there is no agreement (i.e., $N_1 = 0$) and probability of chance agreement is at a maximum (Dwyer, 1986; Lindell & Brandt, 1999). The use of this procedure assumes that every rating has the same probability of being selected,

therefore, when restrictions of range rating errors are present the probability of chance agreement may be underestimated. This problem may be resolved by setting a more stringent level of significance or calculating the probability of chance agreement based on fewer scale categories such as 5 or 6 rather than 7 (Dwyer, 1986; Tinsley & Weiss, 1975). Furthermore, chi-square, as a distribution of variance, a central part of IRA, does not hold up to violations of normality. If the data were from a uniform distribution then the chi-square test of significance would be an accurate test, but a nonnormal population is often the case (Dunlap, Burke, & Smith-Crowe, 2003). Also, both Tinsley and Weiss (1975) and Lindell and Brandt (1999) indicated that Lawlis and Lu's (1972) formulas have typographical errors and are computationally complex when the range of acceptable ratings differ by no more than one or two points.

Next, the r_{wg} statistic proposed by James, Demaree, and Wolf (1984, as cited in James, Demaree, & Wolf, 1993; Lindell & Brandt, 1999) to assess agreement among a single group of judges on a single variable regarding a single target was originally cast as a form of IRR because it was based on Finn's (1970) *reliability coefficient*, r , and the *interchangeability (agreement) index* of inter-rater reliability developed by Shrout and Fleiss (1979) (James, Demaree, & Wolf, 1993; Kozlowski & Hattrup, 1992; Schmidt & Hunter, 1989). Although the authors of r_{wg} believe there are a number of unresolved and debatable issues that need to be addressed in future research, the theory underlying r_{wg} as an IRA index is "...logical, legitimate, and meaningful" (James, Demaree, & Wolf, 1993, p. 309). It is based on the ratio of actual variance in ratings to a theoretical benchmark for responses attributable totally to random measurement errors. Some researchers have criticized this hypothetical pattern of maximum possible variance designed to represent

random responding as too lenient and unrealistic (Dunlap, Burke, & Smith-Crowe, 2003; James, Demaree, & Wolf, 1993; Lindell, Brandt, & Whitney, 1999). Nonetheless, there are several advantages of utilizing this measure of agreement, it does not rely on true variance from classic measurement theory, and the obtained coefficient is sensitive to nonnormality, as well as the absolute values of ratings and their rank order reflecting absolute agreement in terms of both the pattern and the level of ratings (Dunlap, Burke, & Smith-Crowe, 2003; Law & Sherman, 1995). r_{wg} is represented as:

$$r_{wg} = \frac{\sigma E^2 - s_x^2}{\sigma E^2} = 1 - \frac{s_x^2}{\sigma E^2} \quad [9]$$

where s_x^2 is the observed variance and σE^2 is the theoretical variance of a discrete, uniform distribution and is calculated as $\sigma E^2 = (A^2 - 1)/12$, where A is the number of response categories (James, Demaree, & Wolf, 1993). They (James, Demaree, & Wolf, 1993) provided an example in which 10 judges rated whether a manuscript should be published on a five-point Likert-type response scale so σE^2 is equal to 2.0. If all 10 judges rate the manuscript a 5, then $s_x^2 = 0$, and $r_{wg} = (2.0 - 0)/2.0 = 1.0$, which indicates perfect agreement. If half the judges give the manuscript a 5 and the other half a 4, then $s_x^2 = .28$, and $r_{wg} = (2.0 - .28)/2.0 = .86$, which indicates a high but not a perfect level of agreement. If three judges each rate the manuscript a 5, a 4, and a 3 and one judge rates the manuscript a 2, then $s_x^2 = 1.067$, and $r_{wg} = (2.0 - 1.067)/2.0 = .47$, which indicates a low level of IRA. If two judges are represented at each value on the scale appearing to have randomly responded to the scale, then $s_x^2 = 2.22$, and $r_{wg} = (2.0 - 2.22)/2.0 = -.11$. Obviously, a critique of the James, Demaree, and Wolf (1984, as cited in Lindell & Brandt, 1999) approach is that negative values are obtained when the ratings have a

variance greater than that associated with a uniform distribution. James, Demaree, and Wolf (1993) recommended to avoid the consequences of negative values that r_{wg} be set to 0 to maintain the interval $0 \leq r_{wg} \leq 1.0$. Furthermore, they (1993) suggested that r_{wg} is best used when the judges are believed to be interpreting the rating scales in a similar manner. Lindell, Brandt, and Whitney (1999) examined the negative values of r_{wg} and rather than automatically replacing them with zeros as advocated by James, Demaree, and Wolf (1984, as cited in Lindell & Brandt, 1999), they determined that just as IRA can be greater than expected by chance it can also be less than expected by chance. These concerns led Lindell and Brandt (1999) to produce values of r_{wg}^* that lie in the interval $-1.0 \leq r_{wg}^* \leq 1.0$ on the common Likert-type response scale ($A = 5$). For example, $A = 3$ yields a minimum value of $r_{wg}^* = -.5$, and $A = 4$ yields a minimum value of $r_{wg}^* = -.8$, whereas $A = 7$ yields a minimum value of $r_{wg}^* = -1.25$. Any value of $A \geq 4$ can yield $r_{wg}^* \leq -1.0$ because the minimum values of r_{wg}^* corresponding to other values of A increase similarly (i.e., $A = 5$ yields a minimum value of $r_{wg}^* = -2.0$) (Lindell & Brandt, 1999). The r_{wg}^* index is adversely affected by extremely small samples, particularly $N \leq 10$ (Lindell, Brandt, & Whitney, 1999). Lindell and Brandt (1999) compared several indices of IRA to assess job relevance of a task item based on the distribution of endorsements by a panel of *subject matter experts*, or *SMEs*. Overall, r_{wg} and r_{wg}^* indices were widely applicable across a variety of situations and displayed statistical significance for items that most analysts believe to have a reasonable degree of agreement (Lindell & Brandt, 1999). James, Demaree, and Wolf (1984, as cited in Lindell & Brandt, 1999) and Lindell, Brandt, and Whitney (1999) extended these indices to assess agreement for multiple items represented as $r_{wg(J)}$ and $r_{wg(J)}^*$, respectively, where

J is the number of items in the scale. $r_{wg(J)}$ uses the average item variance in the numerator and applies the Spearman-Brown correction, however, Lindell, Brandt, and Whitney (1999) determined the Spearman-Brown correction was inappropriate since r_{wg} is a measure of agreement, not reliability. Thus, they suggested the most suitable index of multi-item agreement is:

$$r_{wg(J)}^* = 1 - \frac{s_x^{-2}}{\sigma E^2} \quad [10]$$

where s_x^{-2} is the obtained average variance of the items in the scale (Lindell, Brandt, & Whitney, 1999).

And finally, because of the interpretability problems associated with r_{wg} , Burke, Finkelstein, and Dusig (2003, as cited in Dunlap, Burke, & Smith-Crowe, 2003) proposed the *average deviation (AD)* index, which is the most recent addition to the catalogue of agreement indices. *AD* is computed by finding the absolute deviation of each rating from the mean or median of the group rating and then averaging the deviations. It differs from r_{wg} in allowing the interpretation of the actual categories of the Likert-type scaled used. It is represented as:

$$AD = \sum \frac{\bar{X} - X}{N} \quad [11]$$

The r_{wg} and *AD* statistics are considered complimentary, not competing and studies comparing their results indicate they are highly correlated and yield similar findings on agreement, however, it is difficult to make comparisons with other existing research since *AD* has not been widely applied (Dunlap, Burke, & Smith-Crowe, 2003).

It is clear there is a broad range of indices available to examine IRA, not to mention the tremendous debate among researchers about the appropriateness of one

method over another. Nevertheless, when deciding what IRA index to use and how it can best be utilized two basic issues should be considered according to Dunlap, Burke, and Smith-Crowe (2003). First, it should be determined that a reasonable consensus exists (i.e., sufficiently strong or sufficiently weak) for a group to aggregate individual level data to the group level. Second, it should be concluded that the apparent agreement for the group is significantly different from chance responding, consequently, there is as current trend by researchers to use chi-square tests and bootstrapping methods to assess the significance of IRA indices (Dunlap, Burke, & Smith-Crowe, 2003). Unfortunately, there is no obvious choice of an index of agreement, instead, a better strategy of measuring IRA is a combination of statistical techniques, an awareness of the scale and the sources of error associated with the ratings, and what kinds of decisions will be made with the scores (Goodwin, 2001).

Inter-rater Reliability (IRR)

The “proportionality of ratings” is a central concept to IRR and usually reported in terms of correlational or analysis of variance indices (Tinsley & Weiss, 1975, p. 361). Basically, IRR represents the degree of relationship from one judge’s score to other judges’ scores although the absolute number used to express this relationship may differ from judge to judge. At the nominal-level, rating categories do not differ quantitatively and disagreements do not differ in severity. Ratings are either in agreement or disagreement; therefore, proportionality is useless (Tinsley & Weiss, 1975).

Assuming the rater data is interval-level or ordinal scales that assume interval properties, one of the most popular statistics for calculating the degree of consistency between two judges is the Pearson product-moment correlation coefficient, which

indicates the linearity or how closely related the two rater's data are to one another.

However, as previously noted in several other cases, the upper limits of the correlation coefficient can be attenuated if the data on the rating scale is more skewed toward one end of the distribution or the other (Stemler, 2004). Furthermore, the Pearson correlation is “mean-free”, so the differences or similarities between the raters in their level of ratings are not taken into account causing the correlations between two raters' sets of scores to be very high, although their means may be very different from one another (Goodwin, 2001). Instead and specifically for ordinal scales, Tinsley and Weiss (1975) recommended the use of Finn's (1970) r as a measure of IRR, which uses a one-way analysis of variance, or a simple ANOVA, when the ratings are on more than one target to obtain the average amount of within groups variability called *within-groups mean square*, WMS . WMS is an average sum of squares derived from the sum of differences between each individual score in a group and the mean of that group, which is squared, then divided by the appropriate df . For the within-groups estimate, df is represented as $k(n-1)$, where k equals the number of groups and n equals the number of participants in each group. Conversely, BMS is the *between-groups mean square*, and is the average sum of squares derived from the sum of the differences between the mean of all scores and the mean of each groups' score, which is squared. It shows how different each groups' mean is from the overall mean. The sum of squares term is divided by the df term, which for the between-groups estimate is represented as $k-1$. After computing all of the sum of squares and the corresponding df terms, in essence, the MS is synonymous with variance (Salkind, 2000). Finn's (1970) r index is as follows:

$$r = 1 - \frac{WMS}{s_e^2} \quad [12]$$

where s_e^2 is the expected variance as if ratings were assigned randomly and is calculated as $k^2 - 1/12$, where k is the number of scale categories. Subtracting the ratio from 1.0 results in the proportion of the total variance in the ratings that is due to non-random factors. Additionally, the degree to which the observed variance (WMS) is less than the expected variance is an indication of the amount of non-chance variance in the ratings. However, in actual practice when a small sample of judges are used the observed variance, by chance, may be less than the expected variance. The use of chi-square can test the hypothesis that the observed variance is equal to the chance variance, but it assumes that the ratings will be normally distributed and that both the raters and the targets will be randomly selected. The chi-square test is:

$$\chi^2 = \frac{N(K-1)s_o^2}{s_e^2} \quad [13]$$

where N is the number of subjects, K is the number of raters, s_o^2 is the observed within-groups variance, and s_e^2 is the expected variance. A stringent critical value such as $p < .01$ is recommended since the normality assumption is frequently violated. Although one advantage of Finn's (1970) r is that it is not reduced by low variance within-judges, unlike other indices of IRA and IRR, however, if the judges avoid the extreme categories of the rating scale chance variance will be higher than the "true" chance variance causing r to be exceptionally high. This is because the computation of expected variance requires the assumption that judges' ratings are purely random and that every rating has the same probability. As a measure of IRR, Finn's (1970) r is fairly recent lacking systematic evidence, however, in comparative studies between product-moment correlation, other indices of IRR, and the *intraclass correlation coefficient (ICC)*, the *ICC* was preferable

and recommended as the best measure of IRR for ordinal, which assume interval properties, and interval-level data (Tinsley & Weiss, 1975). This is because when *ICC* is used to estimate IRR, it permits the error variance to include variance due to rater differences, it allows for an estimation of the precision of the reliability coefficient, and it uses familiar statistics and computational procedures (Goodwin, 2001; Tinsley & Weiss, 1975). As previously discussed, IRA and IRR are distinctly different concepts requiring different measuring techniques, however, it is possible to measure both absolute agreement and consistency using *ICC*, which is an alternative statistic for measuring both the relationships among variables and their variance, not only for pairs of measurements but also for larger sets of targets and multiple judges. *ICC* is preferred, however, when sample size is small (i.e., <15) (McGraw & Wong, 1996).

Intraclass Correlation Coefficient (ICC)

Many available reliability indices are basically versions of the *ICC*, represented as R , which is a ratio of the variance of interest over the sum of the variance of interest plus error. For instance, instead of working from item scores, internal consistency methods (i.e., Cronbach's alpha) may also apply to judgements made by k raters. The resulting value of alpha estimates the consistency of k ratings with that from k other raters or setting $k = 1$ estimates the consistency of single raters with each other (Cronbach, 1990). Nevertheless, Shrout and Fleiss (1979), the sponsors of *ICC*, provided a more rigorous definition. *ICC* is a correlation between one measurement on a target and another measurement obtained on that target. In addition, the modern form of *ICC* uses mean squares (MS) from an ANOVA-type model (McGraw & Wong, 1996; Shrout & Fleiss, 1979). In actuality, "G" theory is an extension of the basic meaning of *ICC* allowing for

the isolation of multiple sources of error, rather than just one source (Goodwin, 2001).

There is more than one formula available for *ICC* depending on the specific situation, the experimental design, and the intent of the study and each one can give different results when applied to the same data (Shrout & Fleiss, 1979). Therefore, Shrout and Fleiss (1979) supplied a set of guidelines for choosing the appropriate form of the *ICC*, which calls for three different decisions and three different designs resulting in six forms of the *ICC*. McGraw and Wong (1996) formally considered additional forms that were omitted by Shrout and Fleiss (1979) because they were not correlations in the strict sense, but are still important for measuring degree of relationship in terms of absolute agreement. Each possible combination calls for slightly altered formulas although common to all *ICC* models is n randomly selected objects of measurement (i.e., targets where the structure of the data is in rows) rated by k variables (i.e., judges where the structure of the data is in columns). For example, the analysis may entail data with just one or two sources of variance (i.e., a one-way or two-way design). Also, whether the same judges rate each target and are the only judges of interest or the judges are a sample from some larger population of judges, called fixed or random respectively, which effects the generalizability of the results. Furthermore, in applied contexts, data is typically collected to make important decisions, for that reason, a difference in means across judges becomes critical when one judge passes a performance that another judge fails. Consequently, single ratings are rarely very reliable since they cannot provide fine discrimination among the raters. As observed in Case 2 in Table 3 in which the judges differed in their average ratings, 3.0, 5.0, and 7.0, respectively. These differences are the only differences in the ratings given by the three judges. And finally, the degree of

consistency or absolute agreement among comparative judgements made about the objects of measurement (Brannick, 2003; Dwyer, 1986; McGraw & Wong, 1996; Nichols, 1998; Tinsley & Weiss, 1975).

Each model decomposes ratings made by i th judge on the j th target with various possible effects, such as effects for the i th judge, the j th target, the interaction between the judge and the target, the constant level of ratings, and the random error component. The three different cases that apply to different rater reliability study designs are: 1) the judges are randomly selected from a larger population of judges and for each target, which are randomly selected from a larger pool of targets, k different judges rate this target, 2) a random sample of k judges from a larger population of judges rate n targets, which are randomly selected from a larger pool of targets, and 3) each randomly selected target is rated by k judges and these are the only judges of interest (Shrout & Fleiss, 1979). Basically, in Case 1, the judges who rate the target are not necessarily the same as those who rate another. A realistic example of this situation is a survey in which the responses are anonymous. Since there is no way to associate the k variable with a particular judge variability due to specific judges, interactions of judges with targets, and measurement error cannot be partitioned out. All of these potential sources of measurement error are combined in the within-target variance and treated as error yielding two MS s, one for object of measurement (WMS) and one for the residual sources of variance (BMS). It is a one-way random effects model (i.e., row effects random) (McGraw & Wong, 1996; Nichols, 1998). Basically, in Case 2, the same set of k judges which are a random sample from a population of potential judges rate each target. Therefore, judge is considered a random effect and the ICC estimates the reliability of the

larger population of judges. “In practical terms, one knows that the levels of a variable are random, when a change in the levels of the variable would have no effect on the question being asked”, explained McGraw and Wong (1996, p. 37) calling this the “replaceability test”. In other words, judges constitute a random factor since the particular judges selected for any study are always replaceable by others from the same population (McGraw & Wong, 1996). This corresponds to a fully-crossed (judge X target) two-way random effects model (i.e., column and row effects random). Unlike Case 2, Case 3 only applies to the k judges in the study and cannot be generalized to other judges. In essence, these are the only judges available. In contrast to the “replaceability test” where the particular judges selected for a study are always replaceable, an example of a fixed effect variable is a biological relationship in a study such as a mother and a child or a population of judges that only possess the appropriate training and standardization of task. It is also a fully-crossed (judge X target) two-way mixed effects model (i.e., column effects fixed and row effects random). Then, each case can be analyzed based on an individual rating or the mean of all ratings producing a total of six different versions of *ICC* (McGraw & Wong, 1996; Shrout & Fleiss, 1979; Uebersax, 2003; Yaffee, 1998). Whereas there is just one one-way model for *ICC* in which only absolute agreement is measurable, there are four additional formulas taken from Shrout and Fleiss (1979) and designated by McGraw and Wong (1996) as 2A and 3A, respectively. The A extension to the case numbers indicates the interaction component is absent from the form and measures correlation using an absolute agreement definition for single scores and average scores (McGraw & Wong, 1996). The difference between consistency and absolute agreement measures is defined by how the variability due to

judges is treated. If that variability is considered irrelevant, it is not included in the denominator of the *ICC*, thus a measure of consistency is produced. If differences among the levels of ratings are considered relevant, it contributes to the denominator of the *ICC*, thus a measure of absolute agreement is produced (Nichols, 1998). Shrout and Fleiss (1979) also designated the intraclass reliability coefficients as *ICC*(case, expected unit of reliability measurement). *ICC*(1,1) is Case 1, a one-way random effects model for a single rating and *ICC*(1,*k*) is a one-way random effects model for the mean of the *k* judges. If, for instance, there are four judges, then this is called *ICC*(1,4). The *ICC* for Case 2 and the measurement of a single rating is called *ICC*(2,1) and *ICC*(2,*k*) is a two-way random effects model average measure or *ICC*(2,4) for the four judges (Yaffee, 1998). Thus, *ICC*(2A,1) is the degree of absolute agreement among measurements for a two-way, single score model and *ICC*(2A,*k*) is for measurements that are averages based on *k* independent measurements on randomly selected objects (McGraw & Wong, 1996). *ICC*(3,1) is designated as the two-way mixed effects model single measure and *ICC*(3,*k*) is the two-way mixed effects model average measure or *ICC*(3,4) for the four judges (Yaffee, 1998). Additionally, *ICC*(3A,1) is the absolute agreement of measurements made under the fixed level and *ICC*(3A,*k*) is the degree of absolute agreement for measurements that are averages based on *k* independent measurements made under the fixed levels (McGraw & Wong, 1996).

The basic linear model for a fully-crossed Shrout and Fleiss (1979) analysis of variance design in which the same *k* judges rate all *n* targets is:

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij} \quad [14]$$

where the component μ is the overall population mean of the ratings. The a stands for the judge demarcating their disposition (e.g., leniency or severity) and a_i is the difference from μ of the i th judge's ratings, b stands for the target demarcating the differences in the trait of interest and b_j is the difference from μ of the j th target's true score determined by repeated ratings on j th target, (ab) stands for the interaction and $(ab)_{ij}$ is the degree that the i th judge departs from their usual ratings when confronted with the j th target, and e_{ij} is the random error in the i th judge's scoring of the j th target. However, there are different assumptions made about a_i and $(ab)_{ij}$ in each case. In Case 2, a_i is a random variable that is normally distributed with a mean of zero and a variance σ_a^2 . In Case 3, it is a fixed variable that is constrained to $\sum a_i = 0$ and the corresponding parameter to θ_j^2 is $\sum a_i^2 / k(k-1)$. Since repeated ratings by each judge on each target are not required the interaction $(ab)_{ij}$ and the random error (e_{ij}) are not separately estimable. In Case 2, all the components of $(ab)_{ij}$ ($i = 1, \dots, k; j = 1, \dots, n$) are mutually independent with a mean of zero and a variance σ_{ab}^2 , except for in Case 2A where there is no interaction effect $(ab)_{ij}$. However, in Case 3, independence can only be assumed for interaction components that involve different targets, except for in Case 3A where there is no interaction effect $(ab)_{ij}$. This consequence is also observed in Case 1 in which the effects due to the judges, the interaction between judges and targets, and the random error are not separately estimable. Therefore, the basic linear model for this ANOVA design is:

$$x_{ij} = \mu + b_j + w_{ij} \quad [15]$$

where w_{ij} is a residual component equal to the sum of the inseparable effects of the judge, the interaction, and the error (Shrout & Fleiss, 1979).

Table 4 displays a fully-crossed (judge X target) Shrout and Fleiss (1979) analysis of variance and *MS* expectations design in which each judge rates each target once. Note that in Case 2 or 3, the within-target sum of squares is partitioned into a between-judges sum of squares and a residual sum of squares. The corresponding *MS*s are denoted *JMS* and *EMS* (Shrout & Fleiss, 1979).

Source	Label	<i>df</i>	Case 1 MS (One-way Random Judges)	Case 2 MS (Random Judges with interaction and without)	Case 3 MS (Fixed Judges with interaction and without)
Between Targets	<i>BMS</i>	$n - 1$	$k\sigma_t^2 + \sigma_w^2$	$k\sigma_t^2 + \sigma_i^2 + \sigma_e^2$ $k\sigma_t^2 + \sigma_e^2$	$k\sigma_t^2 + \sigma_e^2$ $k\sigma_t^2 + \sigma_e^2$
Within Targets	<i>WMS</i>	$n(k-1)$	σ_w^2	$\sigma_j^2 + \sigma_i^2 + \sigma_e^2$ $\sigma_j^2 + \sigma_e^2$	$\theta_j^2 + k/(k-1)\sigma_i^2 + \sigma_e^2$ $\theta_j^2 + \sigma_e^2$
Between Judges	<i>JMS</i>	$k - 1$		$n\sigma_j^2 + \sigma_i^2 + \sigma_e^2$ $n\sigma_j^2 + \sigma_e^2$	$n\theta_j^2 + k/(k-1)\sigma_i^2 + \sigma_e^2$ $n\theta_j^2 + \sigma_e^2$
Error	<i>EMS</i>	$(k-1)(n-1)$		$\sigma_i^2 + \sigma_e^2$ σ_e^2	$k/(k-1)\sigma_i^2 + \sigma_e^2$ σ_e^2

Table 4 displays a fully-crossed (judge X target) Shrout and Fleiss (1979) analysis of variance and mean square expectations design in which each judge rates each target once

Using this information, Shrout and Fleiss (1979) devised *ICC* formulas to estimate, ρ , the population value. For example, in Case 1, *WMS* is the unbiased estimate of σ_w^2 and by subtracting *WMS* from *BMS* and dividing the difference by the number of judges per target it is possible to obtain an unbiased estimate of the target variance σ_t^2 . Also, the w_{ij} components are assumed to be independent, thus, σ_t^2 is equal to the covariance between two ratings on a target. Since the covariance of the ratings is a variance term, the index takes the form of a variance ratio:

$$\rho = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_w^2} \quad [16]$$

Then, the estimate can take the following form:

$$ICC(1,1) = \frac{BMS - WMS}{BMS + (k-1)WMS} \quad [17]$$

Under the assumptions of Case 2 that judges are randomly sampled, the covariance between two ratings on a target is again σ_t^2 , but is different from that obtained under Case 1. Because the effects of judges is the same for all targets under Cases 2 and 3, interjudge variability does not affect the expectation of BMS and its observed value will be smaller than that under Case 1. By subtracting EMS from BMS and dividing the difference by k an estimate of the target variance σ_t^2 can be obtained. The expression for the parameter ρ is again a variance ratio:

$$\rho = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_j^2 + \sigma_i^2 + \sigma_e^2} \quad [18]$$

It is estimated by:

$$ICC(2A,1) = \frac{BMS - EMS}{BMS + (k-1)EMS + k(JMS - EMS)/n} \quad [19]$$

Case 3 differs from Case 2 in that the judges are fixed, therefore, σ_t^2 is no longer equal to the covariance between ratings on a target since the interaction terms on the same target are correlated. The actual covariance is equal to $\sigma_t^2 - \sigma_j^2/(k-1)$ and the total variance is equal to $\sigma_t^2 + \sigma_j^2 + \sigma_e^2$, and thus the parameter ρ variance ratio:

$$\rho = \frac{\sigma_t^2 - \sigma_j^2/(k-1)}{\sigma_t^2 + \sigma_j^2 + \sigma_e^2} \quad [20]$$

This is estimated as:

$$ICC(3,1) = \frac{BMS - EMS}{BMS + (k-1)EMS} \quad [21]$$

Tables 5 and 6 shows the *ICC* formulas for cases of consistency and absolute agreement, and whether it is an individual rating or the mean of all the ratings (Shrout & Fleiss, 1979).

	Case 2 (Consistency, Random Judges)	Case 3 (Consistency, Fixed Judges)
1 Judge	$\frac{BMS - EMS}{BMS + (k - 1)EMS}$	$\frac{BMS - EMS}{BMS + (k - 1)EMS}$
All Judges	$\frac{BMS - EMS}{BMS}$	$\frac{BMS - EMS}{BMS}$

Table 5 shows the *ICC* formulas for cases of consistency and whether it is an individual rating or the mean of all the ratings (Brannick, 2003)

	Case 2A (Absolute Agreement, Random Judges)	Case 3A (Absolute Agreement, Fixed Judges)
1 Judge	$\frac{BMS - EMS}{BMS + (k - 1)EMS + k(JMS - EMS)/n}$	$\frac{BMS - EMS}{BMS + (k - 1)EMS + k(JMS - EMS)/n}$
All Judges	$\frac{BMS - EMS}{BMS + (JMS - EMS)/n}$	$\frac{BMS - EMS}{BMS + (JMS - EMS)/n}$

Table 6 shows the *ICC* formulas for cases of absolute agreement and whether it is an individual rating or the mean of all the ratings (McGraw & Wong, 1996)

Shrout and Fleiss (1979) explained it would be unlikely that the formulas under Cases 2 or 3 would ever be applied in a Case 1 study since the appropriate *MSs* are not available, however, it would be more likely that *ICC*(1,1) would be used erroneously on data from Cases 2 or 3. In most practical applications, there are multiple judges and each judge rates each target where generalizations to other raters are permitted or Case 2. Consider the paired scores (2,4), (4,6), and (6,8), they are perfectly correlated using a consistency definition *ICC*(2,1) $R = 1.0$, but using an absolute agreement definition *ICC*(2A,1) $R = .67$ (McGraw & Wong, 1996). Note that the formulas derived for both the random and the fixed models are identical even though the population *ICCs* are defined differently in the two cases. In either case, the value is the same, however, the distinction is in the interpretation of the *ICC* not the calculation. When judges are

randomly sampled one can generalize beyond the data, but not when they are fixed (McGraw & Wong, 1996; Nichols, 1998). Also, $ICC(2,k)$ and $ICC(3,k)$ are equal to Cronbach's alpha, which is based on the number of judges (or items) and the ratio of the average inter-judge covariance to the average judge variance (Nichols, 1998).

The theoretical limits of the ICC are considered to be 0 and 1.0. As a general rule, R is "excellent" if it is larger than .75, "good" if R is between .4 and .75, and "poor" if R is less than .4. The higher the estimate of R indicates there is less of an impact of raters and the more equivalent they are; meaning one could substitute for another in their ratings. However, as Lahey, Downey, and Saal (1983) pointed out that large negative values and large positive values are possible. A negative ICC can be obtained if BMS is less than either EMS or WMS indicating the items were reversely coded (e.g., a judge assigns 0 as the highest score and 10 as the lowest score), the scale was unsuitable for its intended purpose, or the samples were badly correlated. For this reason, they (1983) suggested that researchers always test for the significance of the target (BMS) before calculating the ICC . Shrout and Fleiss (1979) presented the traditional F test to test for the significance of the ICC values. In the analysis of variance, the F value tests whether the raters significantly differ in their assessments. Here, if a result is significant it confirms there is sufficient evidence to believe the ICC value is significantly different than zero or some other arbitrary value such as .75, which may more appropriately suggest that there is some correlation among the raters' scores. To compute a value for F it is the ratio of BMS to either EMS or WMS and also tests for the significance of BMS . To evaluate the F ratio, the table, *Critical Values for Analysis of Variance or F Test*, is used to establish a critical region for this test presenting a given value of F that would

occur by chance no more than 5% of the time or no more than 1% of the time. The F distribution depends on the df for the numerator and the denominator. If the obtained value is more extreme than the critical value then the ICC value is significant (Lahey, Downey, & Saal, 1983; Salkind, 2000). Consider the three cases of data in Table 3, $ICC(1,1) R = 1.0, .18,$ and $.10$, respectively. Tinsley and Weiss (1975) explained that negative values of R are mathematically possible. Basically, a large negative value that exceeds its theoretical range occurs when the differences between targets approach zero and measurement error is greater than zero (Lahey, Downey, & Saal, 1983). For example, the ratings of Case 3 may have had low IRR primarily due to the restricted range of ratings given by all three raters. However, when negative values are observed in actual practice it implies Rater X Ratee interaction (Tinsley & Weiss, 1975).

Fortunately, there is an easier way to calculate the Shrout and Fleiss (1979) and McGraw and Wong (1996) coefficients by using one of the popular software statistical packages available today. Still, the researcher must select the appropriate ICC to apply to their data set by first determining whether the data are to be treated via a one-way or a two-way ANOVA model based on the source(s) of systematic variability. One systematic source of variance is associated with differences among objects measured (i.e., targets), which is always treated as random in an ANOVA and since it is unknown which scores were given by which judge, a one-way random effects model should be used or Case 1. Variability among the judges is generally treated as a second source of variance, suppose they are a random sample from a larger pool of judges. Then, a two-way random effects model should be used or Case 2. Suppose the set of judges are unique in some way and not considered part of a larger pool of judges, then a two-way mixed effects

model should be used or Case 3. The option of defining agreement in terms of consistency, although it is not estimable under Case 1, or in terms of absolute agreement is the next step in selecting the appropriate *ICC*. And finally, the researcher must decide whether to obtain the reliability estimate for a single judge's rating or for the mean of k ratings. Another advantage of calculating the *ICC* with a software statistical package is that it also produces significance tests (i.e., F value tests) as well as confidence intervals for both one-way and two-way models. Confidence intervals are an alternative test to the use of significance and are generally expressed as a range called the *lower bound* and *upper bound*. In survey research, for example, it is important to understand how close the reliability estimate is to the true value if the entire sample of the population had been surveyed. Therefore, a confidence interval is a calculated range for the true value, which is often set at 95%, meaning at a 95% confidence level, 95 times out of 100 the true value will fall within this confidence interval (Nichols, 1998; Yaffee, 1998).

Increasing Reliability

Error is introduced into the overall evaluation of score reliability by way of such factors as the instrument quality, the assumptions involved in scaling, the conditions of administration, the subjective judgment, and the method used to assess it. Beyond improving measurement quality by time spent on thoughtful design, when there is a lack of agreement and consistency in the data the question becomes what has happened and how to improve it. Systematic rater errors (i.e., restriction of range), interaction effects (i.e., judge X target), homogeneity, item ambiguity, construct confusion, or other survey idiosyncrasies can be sorted out via *ICC* then researchers may exercise the following strategies to achieve a higher degree of IRA and IRR (Cohen & Swedlik, 2002; Lahey,

Downey, & Saal, 1983; Light, Singer, & Willett, 1990). Rater training may improve reliability among judges, however, generic lectures to raters are not as effective at enhancing reliability as group discussion along with practice exercises on rater accuracy (Cohen & Swerdlik, 2002). Some of the techniques used in such training programs are role playing, computer simulation, and watching themselves on videotape as well as discussion of a particular rating format, scoring rubrics, job requirements, and observational elements (Cohen & Swerdlik, 2002). Brannick (2003) suggested frame of reference training and training that is specific to the construct of interest. Additionally, developing assessment forms in cooperation with the judges using them, as well as making their tasks simpler may yield better rater reliability. Concrete, observable behaviors are easier to evaluate than abstract concepts. However, simplifying the measure of an intended construct, for example, hostility, and defining it in terms of a specific quantitative pattern of behavior, such as shouting, may not fully capture the concept of hostility. The judges are only able to consider shouting in order to infer the degree of hostility, which can be expressed in many other different ways (Brannick, 2003). If the problem is scoring criteria, then rewriting the scoring rules might add clarity for the judges or together determining agreed-upon scoring rubrics and training raters to use those rubrics (Cohen & Swerdlik, 2002; Linn & Gronlund, 2000).

Another effective strategy to improve measurement quality is to increase the number of items on the instrument as long as each item equivalently measures the same construct under investigation. A single-item test is particularly vulnerable to measurement error since the score of one item, even a well-written item, can be very different from their true score. Estimating over many items, in which some scores will be

high and some low, but on average, they should be nearer to the individual's true score (Light, Singer, & Willett, 1990). This same strategy can be applied to the number of judges; as judges are added, the variation attributed to measurement error is exchanged for the increasing variability in the judges' true scores. Unfortunately, for many researchers it may be too expensive to increase the number of judges (Brannick, 2003).

According to Landy and Farr (1980) rater and ratee characteristics are fixed and, unlike the test items, rating format scale, or scoring rubrics, they cannot be easily changed. Therefore, training programs are probably ineffective in eliminating specific biases. Still, little is known about the dynamics of demographic characteristics, such as race and gender, and their affect on ratings. They suggested, "We must learn much more about the way in which potential raters observe, encode, store, retrieve, and record performance information, if we hope to increase the validity of ratings" (p. 100).

This comprehensive view of survey methodology, measurement error, and score validity and reliability points to an empirical evaluation of the meaning of scores and their usefulness. These issues are fundamental whenever evaluative judgements and subsequent decisions are made. The scope of this study investigates the degree of correspondence between expert ratings among judges to better understand the factors that cause raters to disagree, with the ultimate goal of improving their ratings. The extent to which judges agree (or disagree) with one another in their impressions of a common target is the primary focus. Basically, determining there is unreliability of the construct or rating format will certainly lower validity, however, discovering the variation is due to raters' legitimate differences in perspective will not. Yet, these legitimate differences in opinion among groups of raters often appear as lower IRA and IRR correlations. As a

result, if group membership is considered in the analysis the IRA value should be higher. In this study, limited consideration is given to one particular form of group membership (i.e., profession) in a rating process. Data was collected from expert raters responding to an authentic survey instrument, maintaining important characteristics of the judges and the task, designed to measure the value and impact of Engineering Criteria 2000 (EC2000). Although measures such as percentage agreement, Cohen's kappa, and Cronbach's alpha have been commonly used to measure IRA and IRR, these coefficients have serious disadvantages. A score's usefulness largely depends on how accurately the observed scores allow researchers to generalize about a person's behavior in some wider set of situations. "G" theory actually extends the basic meaning of *ICC*, which is calculated using an ANOVA, providing the most information bearing on the degree of agreement and consistency in scores. More appropriate assessment of these rater differences and identification of the specific sources of inconsistency in measurement can help guide revisions and improvements and facilitate more informed interpretations and uses of scores.

Statement of Hypothesis

When group membership is considered in the analysis of data via *ICCs* collected from two distinct groups of expert judges, who collectively hold legitimate differences in opinions, as a function of their group membership, it is expected the IRA correlation will be higher than the analysis which does not account for group membership.

Method

Aim of Study

The main purpose of this study is to examine the reliability evidence of expert ratings among judges responding to a survey instrument. Estimation of IRA and IRR are pre-requisites for evaluation of scores derived from high levels of judgement to assure that the results provide valid information to make important decisions. By analyzing IRA and IRR via *ICC* details are revealed about the properties of scores and the dispositions of judges that are central to validation of the instrument.

Participants

Data were gathered from a diverse group of constituents in the engineering profession, ranging from industry hiring recent graduates from a number of different schools, graduate engineering degree program faculty teaching recently entering graduate students, to the institutions themselves. Two of these constituent groups from the engineering community, industry and graduate degree program faculty, had a perception of the overall preparation of their new hires and grad students and the need for specific criteria (i.e., EC2000 student learning outcomes criteria (a-k)). Administrators/faculty from undergraduate-level engineering programs, which make up the third constituent group, were better able to evaluate the importance of the institutional process-oriented EC2000 criteria toward program effectiveness as well as their current degree of implementation. A total of 91 expert raters from a sample of ABET representatives as well as non-ABET members responded to a set of EC2000-related questions specific to their profession. Age, gender, and ethnicity of respondents were not evaluated. However, the majority of respondents (78 or 85.7%) had 10 or more years of experience

in the engineering profession and 57 (62.6%) reported a Ph.D. as their highest degree earned. The industry constituent group, hereafter referred to as industry, consisted of 41 raters and the graduate-level program faculty constituent group, hereafter referred to as education, consisted of 26 raters. On their own accord, an additional six respondents attempted to answer the set of EC2000-related questions for both industry and education as indicated by their professional experience in both occupational areas, however, only four fully completed each section, hereafter referred to as both. A total of 43 raters responded to the process-oriented criteria, hereafter referred to as administrators. Of these, 18 answered only this set of EC2000-related questions, two of the raters also responded to the industry questions and 23 also responded to the education questions. However, there were no raters that responded to all three sets of EC2000-related questions specific to their professions (i.e., industry, education, and administrators). Most of those from industry (20 or 48.8%) reported the amount of time they spend with recent graduates from ABET accredited, undergraduate engineering programs as “some”, 12 (29.3%) reported “little”, 7 (17.1%) reported “almost none”, and 2 (4.9%) reported “almost all”.

Measures

The study included the development of a survey instrument titled ABET Engineering Criteria Survey (Appendix B) designed in collaboration with the Office of Institutional Research at Embry-Riddle Aeronautical University. The survey gathered general data from the respondents (i.e., current profession/position, years in the profession, highest degree), then presented a separate set of EC2000-related questions specific to their profession with clear instructions of how to proceed. Those from

industry were asked about the amount of time they spend with recent graduates from ABET-accredited, undergraduate engineering programs and to then consider these recent graduates and respond to EC2000 student learning outcomes criteria (a-k) questions. Industry representatives then evaluated the importance of each skill for successful performance in an entry-level engineering position at their company, as well as the actual job performance of recent graduates (Part A). Graduate-level engineering program representatives provided similar evaluations, but with respect to success at and actual performance in a graduate-level engineering program (Part B). Respondents who were administrators/faculty in an undergraduate-level engineering program were also posed with the institutional process EC2000 criteria, and asked to evaluate each item's importance toward their program's effectiveness as well as the current degree of implementation at their university (Part C). All items possessed a five-point Likert-type rating scale, the value of five being the highest, with anchors at the two endpoints and a label at the intermediate point. Attached to the questionnaire was a one-page comment form with ten open-ended questions soliciting feedback from the respondents to evaluate the survey instrument itself (Appendix C and C1). The survey was designed and printed using a standard word processor and laser printer, and consisted of a cover letter followed by five pages of survey questions. A cover letter (Appendix B1-B4) preceded the survey and was designed to appear as if the survey came directly from ABET. All data gathered was maintained in Microsoft Excel and then downloaded into SPSS for data generation. All free form comments were maintained in Microsoft Word.

Procedure

Distribution of the survey instrument was carried out in three different ways. In its first distribution, the American Society of Mechanical Engineers (ASME) agreed to participate in the pilot study of the survey instrument. A contact within ASME was established. The contact person represented the Council on Education division of ASME, comprised of three boards: Engineering Education, Professional Development, and Pre-College Education. 48 surveys were included in the registration materials that were mailed by the contact person to board members who had pre-registered for the ASME Summer Meeting, held June 4-8, 2000, in Providence, Rhode Island. Participants were instructed to complete the survey and bring it to the meeting. Next, in order to obtain a larger sample of respondents, a second distribution consisted of the entire membership of the three boards of the Council on Education division of ASME, who received the survey dated February 2, 2001 by mail. In addition, the Academic Affairs Committee of the American Institute of Aeronautics and Astronautics (AIAA) also received the survey dated February 2, 2001 by mail. Members were assigned code numbers to track their status. Those who did not complete the first mailing were sent a second copy of the survey dated March 8, 2001. The mailing package consisted of a cover letter, the five page survey, the comment form, and a stamped reply envelope. Generated from these membership lists a total of 76 surveys were mailed. Finally, in the third distribution, a remaining 37 surveys were administered through personal contacts to various entities involved in the engineering field to include an engineering firm, a municipality, two government-sponsored agencies, and a university. In total, 161 surveys were distributed and 91 surveys (56.5%) were completed and returned.

Reliability Analysis

To assess the rater reliability, both absolute agreement (IRA) and consistency (IRR), intraclass correlation coefficients (*ICC*) were calculated for the expert raters (i.e., industry and education) that subjectively judged the same items (i.e., student learning outcomes in Part A or Part B). As described earlier, *ICC* seeks to compare the variability of different ratings to the total variation across all scores and all targets. Recall that rater agreement indices are based on the extent to which judges assign identical ratings to targets, whereas, reliability indices are based on the extent to which judges produce similar rank orderings for targets (Lahey, Downey, & Saal, 1983). Of the forms of *ICC* identified by Shrout and Fleiss (1979) and McGraw and Wong (1996), in this case, the two-way random effects model best fit the data of a fully-crossed design where the same set of k judges rated each target. Although the selection of participants was not random, each judge was selected from a much larger population of professional and teaching engineers. Therefore, it is expected that the ratings from each judge would generalize to this population. Since combining multiple ratings generally produces more reliable results, the mean of k ratings will be applied to the appropriate *ICC*. However, according to Shrout and Fleiss (1979), both units of reliability should be reported, given that the average measure reliability is always greater in magnitude than the reliability of individual ratings.

The focal point in this case was the various judges that rated the same EC2000-related items (i.e., student learning outcomes criteria in Part A or Part B) specific to their profession. The first analysis of the study aimed to find out, regardless of their group membership, did the judges' ratings covary (i.e., reliability) and was there a reasonable level of agreement among the judges? The initial reliability analysis included the ratings

of eight randomly chosen judges, four that completed Part A, or industry, and four that completed Part B, or education. Hereafter this group was referred to as diverse. Recall that high levels of agreement indicate the various judges share a similar meaning of the construct being measured, regardless of whether or not they apply the construct similarly across targets. High levels of reliability indicate a consistent application of criteria across targets, even if each judge applies the criteria differently (i.e., some judges may be more lenient than others, but their resulting ratings covary). However, an analysis performed without regard to group membership may result in an overall low *ICC* (i.e., low *IRA* and/or *IRR*) since it is possible the different groups collectively hold legitimate differences of opinion about the construct being measured. Subsequently, as part of the second analysis, the diverse group was separated into their specified constituent groups, industry or education. Then, the ratings of four randomly chosen judges from industry and four randomly chosen judges from education were newly added to the two constituent groups, for a total of eight judges in each group. Once group membership was considered agreement and consistency within a constituent group became the key issue and as a result the *IRA* value should increase. Combining the judges from each of their respective constituent groups, industry and education, into a single group of judges, diverse, should increase the amount of variance in their ratings, and thus, increase the standard error of the *IRR* and *IRA* estimates resulting in the decreased *IRR* and *IRA* values. Still, common judges responding to a well-developed survey with a well-defined construct may truly disagree with one another (i.e., low *IRA*), therefore, in conjunction with a high *IRR* value indicates a consistent within-group application of criteria.

To illustrate the computations of *ICC* consider the Truax and Carkhuff Accurate Empathy Scale hypothetical data from Table 3. For an ANOVA analysis the dependent variable and the independent variables must first be identified. The dependent variable in this research is the individual ratings (i.e., scores) for each counselor on the Truax and Carkhuff Accurate Empathy Scale (1967, as cited in Tinsley & Weiss, 1975) and the independent variables are the three raters (i.e., judges) and the 10 counselors that were evaluated (i.e., targets). The initial results calculated by Tinsley and Weiss (1975), $R = 1.0$, $.18$, and $.10$, respectively, used the formula $ICC(1,1)$ to estimate *ICC*. Recall there is only one *ICC* of each type for one-way data, absolute agreement of a single measure and an average measure. In other words, the reliability of a single rating was estimated for absolute agreement only indicating each counselor was rated by a different set of judges, meaning there is no way to disentangle variability due to specific raters, interactions of raters with targets, and measurement error. Instead and as specified in this current study, a two-way random effects model was applied to the same data. *ICC* was measured using both a consistency and an absolute agreement definition for both the individual ratings and the mean of all judges or $ICC(2,1)$ and $ICC(2,3)$ and $ICC(2A,1)$ and $ICC(2A,3)$.

Source	Label	Sum of Squares	<i>df</i>	Mean Square
Judge	JMS	0.00	2	0.00
Target	BMS	190.80	9	21.20
Judge * Target	EMS	0.00	18	0.00

Table 7 ANOVA summary table for Case 1

Source	Label	Sum of Squares	<i>df</i>	Mean Square
Judge	JMS	80.00	2	40.00
Target	BMS	60.00	9	6.67
Judge * Target	EMS	0.00	18	0.00

Table 8 ANOVA summary table for Case 2

Source	Label	Sum of Squares	<i>df</i>	Mean Square
Judge	JMS	1.80	2	.90
Target	BMS	3.20	9	.36
Judge * Target	EMS	8.20	18	.46

Table 9 ANOVA summary table for Case 3

Formulas	Calculations	<i>R</i>
1 Judge, Consistency		
$\frac{BMS - EMS}{BMS + (k - 1)EMS}$	$\frac{21.20 - 0.00}{21.20 + (2)0.00}$	1.00
All Judges, Consistency		
$\frac{BMS - EMS}{BMS}$	$\frac{21.20 - 0.00}{21.20}$	1.00
1 Judge, Absolute Agreement		
$\frac{BMS - EMS}{BMS + (k - 1)EMS + k(JMS - EMS)/n}$	$\frac{21.20 - 0.00}{21.20 + (2)0.00 + 3(0.00 - 0.00)/10}$	1.00
All Judges, Absolute Agreement		
$\frac{BMS - EMS}{BMS + (JMS - EMS)/n}$	$\frac{21.20 - 0.00}{21.20 + (0.00 - 0.00)/10}$	1.00

Table 10 Computation of *ICC* for Case 1

Formulas	Calculations	<i>R</i>
1 Judge, Consistency		
$\frac{BMS - EMS}{BMS + (k - 1)EMS}$	$\frac{6.67 - 0.00}{6.67 + (2)0.00}$	1.00
All Judges, Consistency		
$\frac{BMS - EMS}{BMS}$	$\frac{6.67 - 0.00}{6.67}$	1.00
1 Judge, Absolute Agreement		
$\frac{BMS - EMS}{BMS + (k - 1)EMS + k(JMS - EMS)/n}$	$\frac{6.67 - 0.00}{6.67 + (2)0.00 + 3(40.00 - 0.00)/10}$.36
All Judges, Absolute Agreement		
$\frac{BMS - EMS}{BMS + (JMS - EMS)/n}$	$\frac{6.67 - 0.00}{6.67 + (40.00 - 0.00)/10}$.63

Table 11 Computation of *ICC* for Case 2

Formulas	Calculations	<i>R</i>
1 Judge, Consistency		
$\frac{BMS - EMS}{BMS + (k - 1)EMS}$	$\frac{.36 - .46}{.36 + (2).46}$	-.08
All Judges, Consistency		
$\frac{BMS - EMS}{BMS}$	$\frac{.36 - .46}{.36}$.28
1 Judge, Absolute Agreement		

$\frac{BMS - EMS}{BMS + (k - 1)EMS + k(JMS - EMS)/n}$	$\frac{.36 - .46}{.36 + (2).46 + 3(.90 - .46)/10}$	-.07
All Judges, Absolute Agreement		
$\frac{BMS - EMS}{BMS + (JMS - EMS)/n}$	$\frac{.36 - .46}{.36 + (.90 - .46)/10}$.25

Table 12 Computation of *ICC* for Case 3

	<i>ICC</i>	95% Confidence Interval		F Test with True Value .75			
		Lower Bound	Upper Bound	Value	<i>df</i> 1	<i>df</i> 2	<i>p</i>
1 Judge, IRR	1.00	1.00	1.00
All Judges, IRR	1.00	1.00	1.00
1 Judge, IRA	.36	.014	.760	.185	9	18	.971
All Judges, IRA	.63	.041	.905	.556	9	18	.780

Table 13 Computed confidence intervals and *F* values (*BMS/EMS*), significant at .05 level for Case 2

	<i>ICC</i>	95% Confidence Interval		F Test with True Value .75			
		Lower Bound	Upper Bound	Value	<i>df</i> 1	<i>df</i> 2	<i>p</i>
1 Judge, IRR	-.08	.324	.386	.078	9	18	1.00
All Judges, IRR	-.28	-2.753	.654	.195	9	18	.992
1 Judge, IRA	-.07	-.292	.369	.072	9	18	1.00
All Judges, IRA	-.25	-3.509	.682	.182	9	18	.994

Table 14 Computed confidence intervals and *F* values (*BMS/EMS*), significant at .05 level for Case 3

Notice the IRR values, $ICC(2,1) R = 1.0, 1.0$, and $-.08$, respectively, were altered when a two-way random effects model was used instead of a one-way random effects model, $ICC(1,1) R = 1.0, .18$, and $.10$, respectively. The new *ICC* values for Case 1 and Case 2 were larger than .75, which was considered “excellent” (Lahey, Downey, & Saal, 1983). However, since none of the other values were nominally greater than .75, and since *R* actually equals 1.0 in Case 1 and Case 2, it is unnecessary to present the obvious results of the confidence intervals and the *F* value test, which tests for significance.

Shrout and Fleiss (1979) explained, for the same set of data, $ICC(1,1)$ will, on average,

produce smaller values than $ICC(2,1)$ and $ICC(3,1)$. Since the effect of judges is the same for all targets under $ICC(2,1)$ and $ICC(3,1)$ the inter-judge variability does not affect the expectation of BMS . Therefore, under $ICC(1,1)$, the MS for judges is included in the MS for error and the BMS value tends to be larger than that in $ICC(2,1)$ and $ICC(3,1)$. Thus, the difference between the two estimates of reliability were largest for Case 2 where the largest mean differences among the raters occurred (Shrout & Fleiss, 1979; Tinsley & Weiss, 1975).

Notice in Case 2, the ratings were in perfect agreement (i.e., R approaches 1.0) when a consistency definition was used, $ICC(2,1) R = 1.0$ and $ICC(2,3) R = 1.0$, but not an absolute agreement definition, $ICC(2A,1) R = .36$ and $ICC(2A,3) R = .63$. For consistency measures, the inter-judge variance (i.e., column effects) is excluded from the denominator because it is considered an irrelevant source of variance. For example, it is does not matter that Rater 1 assigns relatively low scores and Rater 3 high scores. In contrast, the absolute agreement measure includes total score variance in the denominator, therefore, the systematic differences among levels of ratings are considered relevant. In Case 2, low agreement among the raters was reflected in their mean ratings of 3, 5, and 7, respectively. However, since the counselors (i.e., targets) were ordered similarly the three raters were proportional in their ratings. Fundamentally, IRR fails to show that the raters differed in their ratings of the counselors. Clearly, both types of information are important in evaluating subjective ratings (McGraw & Wong, 1996; Tinsley & Weiss, 1975).

Finally, Tinsley and Weiss (1975) explained that negative values of R are mathematically possible. For example, the ratings of Case 3, $ICC(1,1) R = -.10$ and

$ICC(2,1) R = -.08$, may have had low IRR primarily due to the restricted range of ratings given by all three raters. However, when negative values are observed in actual practice (i.e., BMS is less than either EMS or WMS) it implies, not only a complete lack of reliability, but also a Rater by Ratee interaction, badly correlated samples, reversely coded items (e.g., a judge assigns 0 as the highest score and 10 as the lowest score), or an unsuitable scale (Lahey, Downey, & Saal, 1983).

Results

Table 15 shows the recorded judgements for eight random judges, represented as diverse, on the 22 EC2000 student learning outcomes criteria labeled targets. Judges 1 through 4 belonged to the constituent group, industry, and Judges 5 through 8 belonged to the constituent group, education. All of the judges earned a Ph.D., except for Judge 4 who earned a Master's degree. All of the judges reported 10 or more years of experience in their field, except for Judge 4 and Judge 8, who reported 4 to 6 years and 1 to 3 years, respectively. Appendix D provides additional general data gathered from each judge (i.e., current profession/position, survey distribution method). Table 15 also includes \bar{X} , which is the mean rating for a judge, and s_x , which is the standard deviation of the judge's ratings.

	Random Judges – Diverse Group							
Targets	1	2	3	4	5	6	7	8
1	5	5	3	5	5	5	5	5
2	4	4	3	2	5	4	4	5
3	5	5	4	5	5	5	5	5
4	5	5	2	5	5	5	4	5
5	5	3	3	3	4	5	2	5
6	4	5	5	4	4	3	4	5
7	4	3	2	3	5	3	3	5
8	5	4	4	4	5	4	4	5
9	5	4	2	4	5	4	5	5
10	4	5	3	4	4	3	3	5
11	5	5	2	4	4	5	4	5
12	2	3	2	2	3	4	3	4
13	1	3	3	3	3	3	4	4
14	4	4	3	2	4	3	4	4
15	3	3	2	2	4	3	4	4
16	2	3	4	3	3	3	3	4
17	2	4	2	3	3	3	4	4
18	2	4	2	2	3	4	3	3
19	2	5	3	3	3	4	2	4
20	2	3	3	1	3	4	2	4
21	2	5	3	3	3	3	3	3

22	3	4	3	3	3	3	4	4
\bar{X}	3.45	4.05	2.86	3.18	3.91	3.77	3.59	4.41
s_x	1.37	.84	.83	1.10	.87	.81	.91	.67

Table 15 Diverse groups' rating of EC2000 student learning outcomes criteria

In addition to means and the standard deviation of the judge's ratings, for an ANOVA the dependent variable and the independent variables must also be identified. The dependent variable was the individual ratings (i.e., scores) for each judge on the ABET Engineering Criteria Survey and the independent variables were the eight judges ($k = 8$) and the 22 EC2000 student learning outcomes criteria ($n = 22$) that were evaluated (i.e., targets). Table 16 shows the results from an ANOVA summary table for the diverse group necessary to compute the ICCs.

Source	Label	Sum of Squares	df	Mean Square
Judge	JMS	37.27	7	5.32
Target	BMS	71.73	21	3.42
Judge * Target	EMS	78.86	147	.54

Table 16 ANOVA summary table for diverse group

Formulas	Calculations	R
1 Judge, Consistency		
$\frac{BMS - EMS}{BMS + (k - 1)EMS}$	$\frac{3.42 - .54}{3.42 + (7).54}$.40
All Judges, Consistency		
$\frac{BMS - EMS}{BMS}$	$\frac{3.42 - .54}{3.42}$.84
1 Judge, Absolute Agreement		
$\frac{BMS - EMS}{BMS + (k - 1)EMS + k(JMS - EMS)/n}$	$\frac{3.42 - .54}{3.42 + (7).54 + 8(5.32 - .54)/22}$.32
All Judges, Absolute Agreement		
$\frac{BMS - EMS}{BMS + (JMS - EMS)/n}$	$\frac{3.42 - .54}{3.42 + (5.32 - .54)/22}$.79

Table 17 Computation of ICC for diverse group

The ICC values calculated for the diverse group and were tested against the desired value of .75. As noted in Table 18, the IRA and IRR values were not statistically different from the test value of .75 (alpha set at .05). The lower bound estimate for IRR

was .72 and for IRA was .62. Both of these values were respectable, but not above the desired value of .75.

	ICC	95% Confidence Interval		F Test with True Value .75			
		Lower Bound	Upper Bound	Value	df1	df2	p
1 Judge, IRR	.40	.243	.607	.255	21	147	1.00
All Judges, IRR	.84	.720	.925	1.592	21	147	.058
1 Judge, IRA	.32	.174	.532	.183	21	147	1.00
All Judges, IRA	.79	.620	.902	1.221	21	147	.260

Table 18 Computed confidence intervals and *F* values (*BMS/EMS*), significant at .05 level for diverse group

Table 19 shows the recorded judgements for eight random judges, represented as industry, on the 22 EC2000 student learning outcomes criteria labeled targets. The diverse group was separated into their specified constituent groups; the industry group consisted of the original Judges 1 through 4 and the ratings of four randomly chosen judges from industry, Judges 9 through 12, were newly added for a total of eight judges. All of the judges earned a Ph.D., except for Judge 4, who earned a Master's degree. All of the judges reported 10 or more years of experience in their field, except for Judge 4, who reported 4 to 6 years. Appendix D provides additional general data gathered from each judge (i.e., current profession/position, amount of time spent with recent graduates—industry only, survey distribution method). Table 19 also includes \bar{X} , which is the mean rating for a judge, and s_x , which is the standard deviation of the judge's ratings.

	Random Judges – Industry Group							
Targets	1	2	3	4	9	10	11	12
1	5	5	3	5	4	5	5	3
2	4	4	3	2	4	5	5	3
3	5	5	4	5	4	5	5	5
4	5	5	2	5	5	4	5	3

5	5	3	3	3	4	4	5	5
6	4	5	5	4	4	4	5	3
7	4	3	2	3	4	4	5	2
8	5	4	4	4	5	5	5	5
9	5	4	2	4	4	4	5	3
10	4	5	3	4	3	4	5	2
11	5	5	2	4	5	4	5	4
12	2	3	2	2	2	2	2	3
13	1	3	3	3	2	3	2	3
14	4	4	3	2	4	2	4	3
15	3	3	2	2	2	2	3	3
16	2	3	4	3	3	2	2	2
17	2	4	2	3	4	2	3	3
18	2	4	2	2	3	2	2	3
19	2	5	3	3	3	3	3	2
20	2	3	3	1	3	3	3	3
21	2	5	3	3	2	2	3	3
22	3	4	3	3	3	4	3	2
\bar{X}	3.45	4.05	2.86	3.18	3.50	3.41	3.86	3.09
s_x	1.37	.84	.83	1.10	.96	1.14	1.25	.92

Table 19 Industry groups' rating of EC2000 student learning outcomes criteria

Table 20 shows the results from an ANOVA summary table for the industry group necessary to compute the ICCs.

Source	Label	Sum of Squares	df	Mean Square
Judge	JMS	23.54	7	3.36
Target	BMS	104.17	21	4.96
Judge * Target	EMS	87.34	147	.59

Table 20 ANOVA summary table for industry group

Formulas	Calculations	R
1 Judge, Consistency		
$\frac{BMS - EMS}{BMS + (k - 1)EMS}$	$\frac{4.96 - .59}{4.96 + (7).59}$.48
All Judges, Consistency		
$\frac{BMS - EMS}{BMS}$	$\frac{4.96 - .59}{4.96}$.88
1 Judge, Absolute Agreement		
$\frac{BMS - EMS}{BMS + (k - 1)EMS + k(JMS - EMS)/n}$	$\frac{4.96 - .59}{4.96 + (7).59 + 8(3.36 - .59)/22}$.43
All Judges, Absolute Agreement		

$\frac{BMS - EMS}{BMS + (JMS - EMS)/n}$	$\frac{4.96 - .59}{4.96 + (3.36 - .59)/22}$.86
---	---	-----

Table 21 Computation of *ICC* for industry group

Table 22 shows the results of the confidence intervals and appropriate *F* value test indicating the computed values for the average ratings were statistically larger than the test value of .75.

	<i>ICC</i>	95% Confidence Interval		F Test with True Value .75			
		Lower Bound	Upper Bound	Value	<i>df</i> 1	<i>df</i> 2	<i>p</i>
1 Judge, IRR	.48	.315	.674	.334	21	147	.998
All Judges, IRR	.88	.786	.943	2.087	21	147	.006
1 Judge, IRA	.43	.270	.633	.278	21	147	.999
All Judges, IRA	.86	.745	.933	1.801	21	147	.026

Table 22 Computed confidence intervals and *F* values (*BMS/EMS*), significant at .05 level for industry group

Table 23 shows the recorded judgements for eight random judges, represented as education, on the 22 EC2000 student learning outcomes criteria labeled targets. The diverse group was separated into their specified constituent groups; the education group consisted of the original Judges 5 through 8 and the ratings of four randomly chosen judges from education, Judges 13 through 16, were newly added for a total of eight judges. All of the judges reported a Ph.D. as their highest degree earned. All of the judges reported 10 or more years of experience in their field, except for Judge 8, who reported 1 to 3 years. Appendix D provides additional general data gathered from each judge (i.e., current profession/position, survey distribution method). Table 23 also includes \bar{X} , which is the mean rating for a judge, and s_x , which is the standard deviation of the judge's ratings.

	Random Judges – Education Group							
Targets	5	6	7	8	13	14	15	16
1	5	5	5	5	5	5	3	5
2	5	4	4	5	3	5	3	4
3	5	5	5	5	5	5	4	5
4	5	5	4	5	5	5	4	5
5	4	5	2	5	3	5	3	4
6	4	3	4	5	4	5	3	4
7	5	3	3	5	3	4	2	3
8	5	4	4	5	4	5	3	4
9	5	4	5	5	5	5	4	5
10	4	3	3	5	4	4	2	5
11	4	5	4	5	5	5	5	5
12	3	4	3	4	2	4	3	2
13	3	3	4	4	3	4	3	2
14	4	3	4	4	2	4	2	2
15	4	3	4	4	2	3	3	2
16	3	3	3	4	3	3	3	3
17	3	3	4	4	2	3	3	3
18	3	4	3	3	3	3	4	2
19	3	4	2	4	2	3	3	3
20	3	4	2	4	2	3	2	3
21	3	3	3	3	3	3	3	2
22	3	3	4	4	3	3	3	2
\bar{X}	3.91	3.77	3.59	4.41	3.32	4.05	3.09	3.41
s_x	.87	.81	.91	.67	1.29	.90	.75	1.22

Table 23 Education groups' rating of EC2000 student learning outcomes criteria

Table 24 shows the results from an ANOVA summary table for the education group necessary to compute the *ICCs*.

Source	Label	Sum of Squares	<i>df</i>	Mean Square
Judge	JMS	28.25	7	4.04
Target	BMS	86.43	21	4.12
Judge * Target	EMS	56.75	147	.39

Table 24 ANOVA summary table for education group

Formulas	Calculations	<i>R</i>
1 Judge, Consistency		
$\frac{BMS - EMS}{BMS + (k - 1)EMS}$	$\frac{4.12 - .39}{4.12 + (7).39}$.55
All Judges, Consistency		

$\frac{BMS - EMS}{BMS}$	$\frac{4.12 - .39}{4.12}$.91
1 Judge, Absolute Agreement		
$\frac{BMS - EMS}{BMS + (k - 1)EMS + k(JMS - EMS)/n}$	$\frac{4.12 - .39}{4.12 + (7).39 + 8(4.04 - .39)/22}$.46
All Judges, Absolute Agreement		
$\frac{BMS - EMS}{BMS + (JMS - EMS)/n}$	$\frac{4.12 - .39}{4.12 + (4.04 - .39)/22}$.87

Table 25 Computation of *ICC* for education group

Table 26 shows the results of the confidence intervals and appropriate *F* value test indicating the computed values for the average ratings were statistically larger than the test value of .75.

	<i>ICC</i>	95% Confidence Interval		F Test with True Value .75			
		Lower Bound	Upper Bound	Value	<i>df</i> 1	<i>df</i> 2	<i>p</i>
1 Judge, IRR	.55	.383	.728	.426	21	147	.987
All Judges, IRR	.91	.833	.955	2.665	21	147	.000
1 Judge, IRA	.46	.285	.660	.302	21	147	.998
All Judges, IRA	.87	.757	.940	2.016	21	147	.015

Table 26 Computed confidence intervals and *F* values (*BMS/EMS*), significant at .05 level for education group

	IRR Consistency Definition		IRA Absolute Agreement Definition	
	1 Judge	All Judges	1 Judge	All Judges
Diverse Group	.40	.84	.32	.79
Industry Group	.48	.88	.43	.86
Education Group	.55	.91	.46	.87

Table 27 Summary of *ICC* computations for diverse, industry and education groups

In all three data sets, notice the IRR and IRA estimates for the single judge were lower than the average measure and since these values often contain too much uncertainty, the mean *ICC* ratings, or “All Judges”, were used for analysis purposes. The IRR and IRA values for the industry and education groups were at or above the desired value of .75, but the IRR and IRA values for the diverse group failed to statistically differ from the test value of .75. Combining the judges from each of their respective constituent groups, industry and education, into a single group of judges, diverse, increased the amount of variance in their ratings, and thus, increased the standard error of the IRR and IRA estimates resulting in the decreased IRR and IRA values. Figures 2 and 3 illustrate graphically the confidence interval data providing the lower bounds and the upper bounds estimate of the computed IRR and IRA values for each distinct group of judges, respectively.

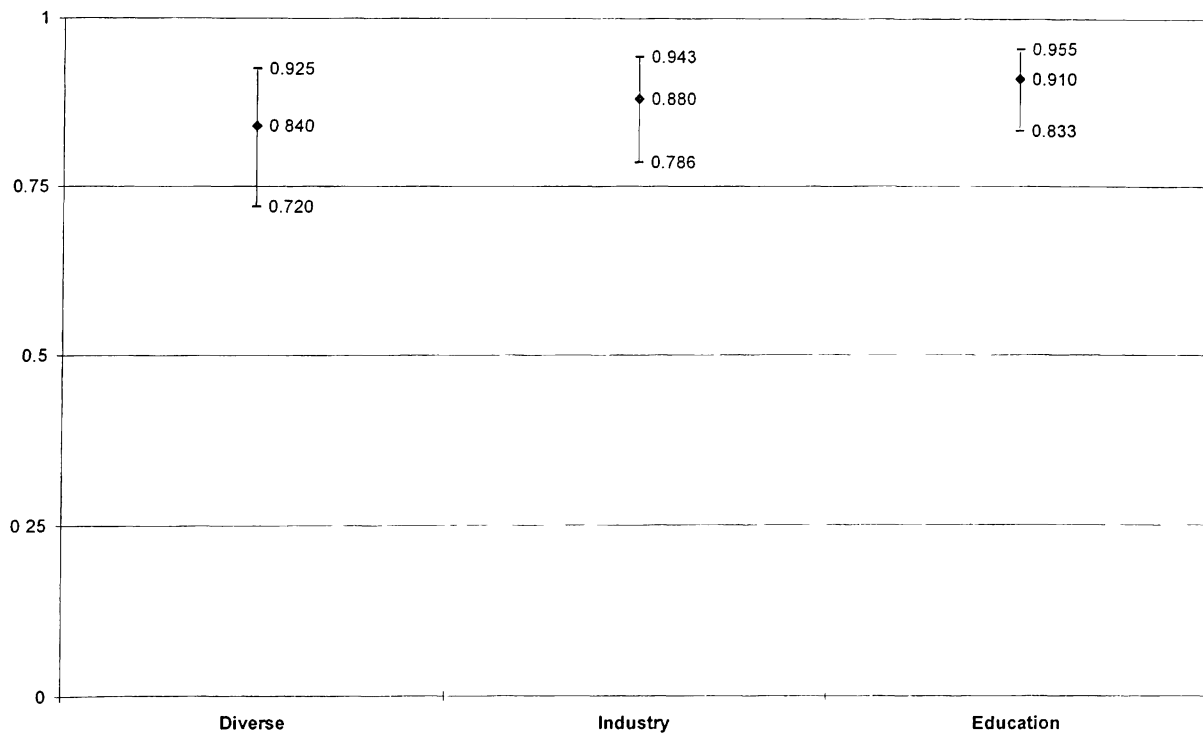


Figure 2 Computed IRR values and confidence intervals

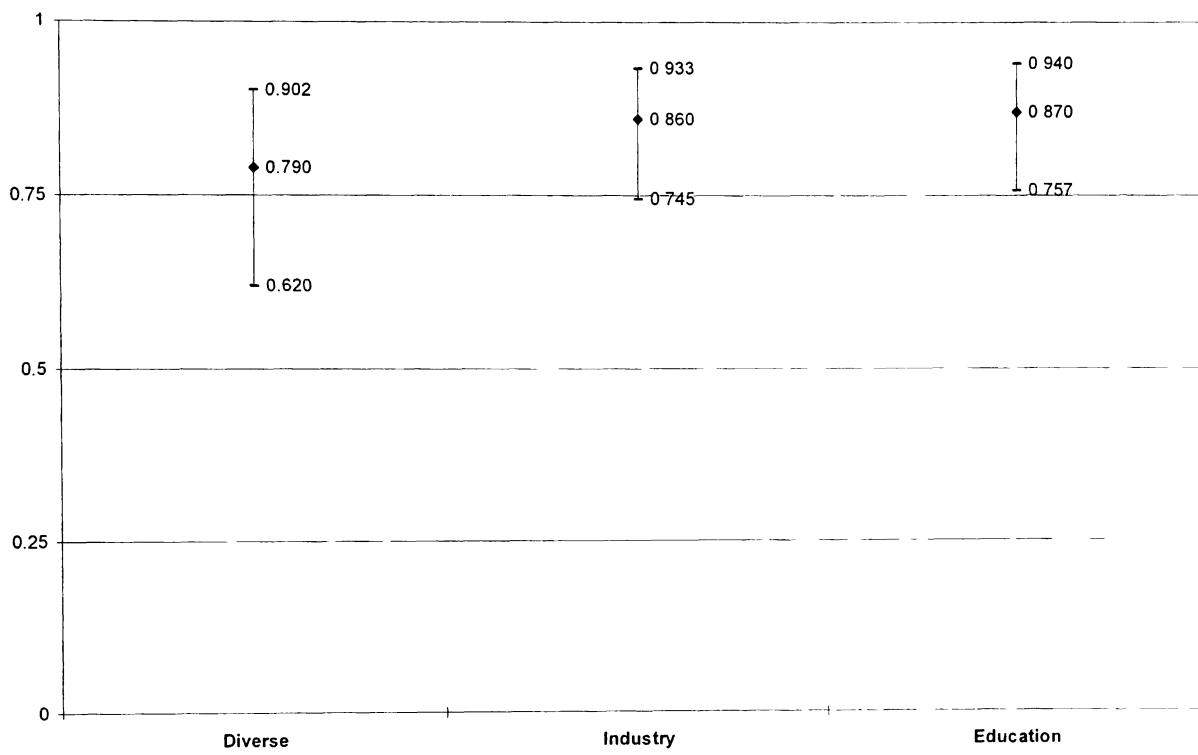


Figure 3 Computed IRA values and confidence intervals

Discussion

By simply looking at the standard descriptive statistics for the individual judges in each constituent group revealed interesting information about the judges and the rating process. In the diverse group, the low mean scores of Judge 3, Judge 4, and Judge 1 indicate they tended to score more harshly than the others did. The high mean scores of Judge 8, Judge 2, and Judge 5 showed they were more generous. The mean scores of Judge 7 and Judge 6 showed them to be middle-of-the-road. Additionally, the larger the standard deviation, the more spread the scores are, and the more different from one another. In that, the low standard deviation of Judge 8's ratings indicate that he/she marked deviations in performance quality on a finer scale than the others, while the high standard deviations for Judge 1 and Judge 4 showed they made free use of the rating scale in their range of scores. In all three data sets, the individual judges seemed to be quite different in their methods of scoring, however, there is much more to be learned in survey interpretation and their individual summary statistics do not provide enough detail.

Using *ICC* to measure the correlation among multiple observations (i.e., judges) on randomly selected objects of measurement (i.e., targets) offers real advantages over traditional analysis. In a single measure one can determine absolute rater differences on the one hand (i.e., IRA) and rank order similarities on the other (i.e., IRR). This level of detailed information is not available with other approaches. *ICC* is a derivative of "G" theory that uses a familiar statistical method, the repeated measures analysis of variance (ANOVA), to make use of details about the relative contributions of various factors, such as variability due to the raters and what is being rated (i.e., targets) and their interactions to the total score variance. These kinds of interaction effects are expected in most

measurement situations and are critically important in research. Especially when data are in the form of ratings, this measurement model provides the most accurate reliability method for measuring the systematic effects of several variables on the consistency of scores that simultaneously intrude in any given measurement application (Goodwin, 2001; Thompson, 2003b).

Overall, the ratings from all three distinct groups of judges, diverse, industry, and education, resulted in *ICC* values larger than .75, which were considered “excellent” (Lahey, Downey, & Saal, 1983). The high levels of agreement, .79, .86, and .87 for diverse, industry, and education, respectively, indicate the various judges shared a common interpretation of the construct being measured. For instance, a majority of judges rated the dominant skill as engagement in life-long learning (Skill 8). Since they classified the statement in exactly the same way it implies that the judges were essentially providing the same information. If some judges classified the major skill as ability to communicate (Skill 9) and some classified the major skill as the ability to apply knowledge of mathematics, science, and engineering (Skill 3), then a breakdown in their shared understanding would have occurred introducing standard error, thus, resulting in lower *IRA* values.

Overall, the high levels of reliability, .84, .88, and .91 for diverse, industry, and education, respectively, indicate a consistent application of the scoring criteria across targets, even if each judge applied the criteria differently (i.e., some judge may be more lenient than others). Basically, *IRR* estimates the extent to which the judges applied their ratings in a manner that is predictable and replicable. Regardless of whether the judges

exhibited exact agreement, here, what is most important was that each judge applied the rating scale consistently within his or her own definition of the scale.

In an analysis of data collected from distinct groups of judges who collectively hold legitimate differences in opinions, as a function of their group membership, disagreement among the rater groups would be expected. Clearly, if the responses among judges within a given constituent group showed little convergence (i.e., agreement), then the reliability and subsequently, the validity of the scores should be seriously questioned. Therefore, it seems that within-group IRA is a more appropriate requirement for assessing score reliability and rater validity.

These preliminary results suggested that the various raters showed a high degree of agreement regarding overall preparation of their new hires and grad students and the need for specific criteria (i.e., EC2000 student learning outcomes criteria (a-k)) though the IRA values declined from .87 to .86 to .79 for the education, industry, and diverse groups, respectively. When group membership was considered in the analysis, the IRA value was higher demonstrating there was, in fact, increased internal agreement within the two distinct constituent groups, industry and education. Furthermore, the issue of reporting IRA across constituent groups (i.e., diverse) appeared to recognize the role-related nature of ratings from various perspectives. This supports the hypothesis that basically, more constituency relevant information generally increased agreement, suggesting that raters who represented education or industry, and considered common judges, agreed more than the diverse group, who varied in their perspectives to some degree.

Another possible method of comparing IRA and IRR values was to look at the confidence interval data for each distinct group of judges. The purpose of a confidence interval is to obtain likely lower and upper bound values for a given outcome score. More importantly, the lower bound value actually indicates the lowest possible value that both the IRA and IRR correlations might be with a certain level of confidence. One can then use the lower bound condition in subsequent computations and decision-making processes as a worst case scenario. In the context of IRR and IRA, smaller confidence intervals, all other things being equal, indicated a higher degree of homogeneity with regard to expected ratings at the population level. The largest confidence intervals were reported for the both the diverse group's IRR and IRA reliability coefficients. They ranged from .720 to .925, or $\pm .205$ points and .620 to .902, or $\pm .282$ points, respectively. Oppositely, the smallest confidence intervals were found in the IRR and IRA reliability coefficients for the education group, which ranged from .833 to .955, or $\pm .122$ points and .757 to .940, or $\pm .183$ points, respectively. This suggests that the group of common judges, in this case, the education group, was much more consistent in their application of the rating criteria and had nearly identical scores regarding the trait of interest. Furthermore, in the absence of a statistical test for comparing *ICCs* and determining the observed difference between the two groups are not likely to occur by sampling error, the overlapping confidence intervals could be put to use instead. At the point estimates where each of the three distinct groups overlap may represent the true group differences if the study were repeated using different judges from the same population.

In contrast to IRA, when the means of the different judges are very different, IRR values may still be high such as the high levels of reliability, .84, .88, and .91 for diverse, industry, and education, respectively. A disadvantage of consistency estimates (i.e., IRR) is that they are highly sensitive to the distribution of data. For example, if all ratings by all judges for all targets were identical, then a reduced level of variation within the ratings causes the scores to appear unreliable (i.e., low IRR) even though they were accurately reflecting raters' true scores. Restricted variability within the judges' ratings may occur particularly if group membership is considered in the interpretation of the data due to the homogenous nature of a group of common judges. As the group of judges or the group of targets become more homogeneous, the power of the *ICC* to detect patterns in the data becomes limited. Other potential sources leading to lack of variability may be a presence of an interaction between judges and targets, a set of consensual judges (i.e., too lenient or too severe), or a rating system that is prone to range restriction. Once these sources of unreliability are ruled out and there is no correlation that exists among the raters indicates that either they are not in agreement about the construct being measured (i.e., low IRA) or the rating instrument is incapable of reflecting their "true" scores (i.e., low IRR). In essence, it is not enough to just state the ratings are unreliable, in order to understand the dynamics of the judgements it is necessary to further investigate the source or sources leading to the unreliability (Lahey, Downey & Saal, 1983).

Still, common judges responding to a well-developed survey with a well-defined construct may truly disagree with one another (i.e., low IRA and low IRR). The raters themselves may still vary in their perspectives and provide different answers to the same well-worded questions. Basically, a low IRA correlation in conjunction with a high IRR

correlation indicates that while the various judges may not have reached a consensus about the construct being measured, they do share a consistent within-group application of rating scale.

Regardless, each of these conditions affects the overall utility of the instrument. If the raters' scores varied because the construct was poorly defined or the rating scale was defective (i.e., survey idiosyncrasies) then survey validity surely will be diminished; however, legitimate differences of opinion (i.e., rater idiosyncrasies) will not necessarily diminish its validity. Traditional concepts of reliability and validity recommend whenever IRA or IRR are low, the scores are of no value and the instrument must be revised or discarded due to poor survey construction and not because of the "true" ideological differences among the judges' perspectives. To use a common analogy, comparing evaluations across different rater groups is like comparing apples and oranges; they both are fruits, just different kinds of fruits (Bozeman, 1997, p. 314). Therefore, and in response to any previous recommendations, it is proposed that the ratings obtained from different groups of common judges who may collectively hold legitimate differences of opinion still be considered reliable and valid, even if they do not exhibit high levels of agreement.

Recommendations

In this study, what was of interest was whether the groups of judges systematically differed in their ratings as a function of their group membership. These preliminary results suggested that IRA values increased when group membership was considered in the analysis indicating an improved within-group convergence in ratings between raters. However, explaining the convergence in ratings between raters is an

interesting hypothesis for future research. Little is known about the moderating influences on the evaluative judgement process, in which case exploring what raters do implicitly will enhance understanding of the psychology that underlies ratings. Potentially important moderating influences on the quality of ratings may be the judges' level of education and experience, as well as their roles, duties, and responsibilities. For example, experienced raters are more likely to have extensive knowledge of the worker characteristics needed to perform a certain job. They may be exposed to similar experiences over time, which will provide a common frame of reference when making judgements. In the same way, supervisors are more likely to evaluate an individual's performance differently than would a subordinate or a peer. Supervisors are evaluating the individual in his or her role as a subordinate and subordinates are rating the individual in his or her role as a supervisor. Essentially, the two distinct groups of judges are not evaluating the same thing, constituting different domains of job performance. This issue has implications for the creation of different assessment tools for use by different groups of judges. It is unrealistic to expect all of these judges with varying capacities to respond similarly to identical, generic rating formats (Bozeman, 1997).

Another recommendation would be to estimate the rater agreement and consistency in several different ways. Each of these approaches may produce different results and information about the consistency of scores. In particular, the results of a "G" study, one in which all the available information from raters, ratees, items, and occasions (i.e., the major sources of measurement error) are included as facets, provides the most information on the reliability of scores (Goodwin, 2001). Only "G" theory simultaneously examines all the measurement influences and the interaction effects, as

well as, estimates reliability coefficients in the context of different decisions (i.e., a decision study or “D” study). The ANOVA method is a powerful tool to estimate variance components and then use these variances to estimate score reliability. The magnitude of the variance components indicates how much each facet contributes to measurement error. First, the ANOVA provides mean squares (*MS*), from the *MS*, variance components can be estimated. From these estimated variance components the *coefficients of generalizability*, analogous to a reliability coefficient, can be calculated. Next, the “G” study then estimates the magnitude of as many potential sources of measurement error. And finally, the “D” study uses this information to design a measurement that minimizes error for a particular purpose (Shavelson, Webb, & Rowley, 1989).

However, it seems there is little indication that these recent advances in reliability methods are even being applied more generally in either published research literature or in popular psychological testing and assessment textbooks. Perhaps the lack of widespread use of these out-of-the-ordinary reliability coefficients in the technical literature is because they are too technical or many researchers’ misconceptions and unawareness may be due to decreased emphasis on measurement course work in doctoral programs. Furthermore, researchers may be operating in a measurement vacuum, in that they are unfamiliar with the analysis and reporting of reliability coefficients outside of their area of expertise (i.e., medical research vs. educational research). It is hoped that this study may motivate researchers to report *ICC* reliability coefficients for their own judgmental data, which could improve the quality of literature reviews and facilitate more informed interpretations of scores (Goodwin, 2001, Henson, 2001).

If important decisions are made with respect to specific scores, another potential problem area may be regarding the standards for accepting or rejecting inter-judge reliability coefficients. Should statistical analyses that involve judgmental data be held to a higher standard than internal consistency estimates since it is known that some reliability methods provide higher reliability coefficients than others? The fact that all three data sets easily fell into the “excellent” category suggests that the guidelines may be too liberal. If much higher inter-judge reliability coefficients were recommended, for example, values of .90 and above are considered “excellent”, then most reliability coefficients would tend to be lower than the established standards.

In the interpretation of results, confidence intervals were adopted in absence of formal standards for comparing *ICCs*. Applying this procedure lacks a strong empirical basis, which may not be completely satisfactory or perhaps, misused. Consequently, the question remains whether or not the conclusions would withstand for the three distinct groups of judges. A comparison of multiple studies that utilize *ICCs* to estimate IRA and IRR of subjective data would address this issue, if they were available.

Finally, in most studies, there is a limited budget for both money and time so naturally, researchers must determine the maximum number of judges and targets to be used in any given measurement situation. Unfortunately, the type and number of judges and targets influences the kinds of inferences that can be drawn from a study. For example, a study with smaller numbers of either may pose problems in generalizing the study to some populations of judges or targets. In order to provide clinicians with good instruments to work with, survey developers should strive not only for higher reliability, but also for more precise reliability coefficients by increasing the number of subjects

whenever possible (Brannick, 2003). Fortunately, when this is not possible, *ICC*, with perhaps more practical significance than statistical significance, can estimate reliability for either one rater or for more than one rater. In this study, more precise reliability determinations could be obtained by repeating the analysis with eight more different judges in each of the three distinct groups and examine the various reliability values.

In this application, the purpose of examining rater agreement and consistency was not to estimate the accuracy of ratings by a single judge. Instead, the aim was more to understand the factors that cause experts to disagree, with the ultimate goal of improving their ratings. The scope of this study was limited to the consideration of one particular form of group membership (i.e., profession) in a rating process, however, there is a wide variety of issues that should be further examined when measurement depends on judgmental indices. In addition to the group-related variables (i.e., education level, experience, and roles), which was previously discussed, other issues include: the types of rating formats, dynamic nature of dimensions to be rated, rating context (i.e., the purpose for the rating), characteristics of rater and ratees (i.e., age, gender, race, cognitive abilities, psychological state, knowledge of the ratee), results of the rating (i.e., actions based on this information) (Landy & Farr, 1980). Research in this area is long overdue to suggest a more unified and modernized approach in an effort to better understand individual rater differences and improve the validity of their judgements.

Summary

Many important decisions are made on the basis of subjective evaluations. As such, the quality of these ratings is an important concern in both research and practice. For instance, would the same scores be obtained if a different but equally knowledgeable judge rated the same items? This variation or unreliability in scores is generally unrelated to the purpose of the measurement and ultimately, reduces its usefulness. Thus, reliability (or lack thereof) carries with it implications for how ratings across judges impacts the validity of scores. In essence, unreliability of the construct or rating format will lower validity, however, variation due to raters' legitimate differences in perspective will not. Investigating the convergence in ratings between raters will help to better understand the disposition underlying ratings. That is, their inter-individual differences (i.e., rater idiosyncrasies), which are considered to be a part of their true variance, and unlikely to be trained away or irrelevant factors that can be better experimentally controlled (i.e., survey idiosyncrasies). The only way this question can be answered conclusively is through empirical analysis. Although there are numerous methods available to measure the degree of agreement and consistency between raters' scores, modern generalizability theory ("G" theory) offers the best estimates for true reliability. Intraclass correlation coefficient (*ICC*), which is an extension of "G" theory, estimates reliability using a repeated measures ANOVA to isolate multiple sources of error and the interactions among them, rather than just one. This approach provides the most information about rater differences allowing more accurate decisions to be made from the resultant scores. The focus of this study was to utilize *ICCs* to estimate both inter-rater agreement (IRA) and inter-rater reliability (IRR) of subjective data gathered from

engineering professionals (i.e., judges) responding to a questionnaire. The preliminary results suggested that the various judges showed a high degree of agreement in their ratings, however, IRA values increased when group membership was considered in the analysis indicating an improved within-group convergence in ratings between raters. Furthermore, the issue of reporting a lower IRA across constituent groups of various judges who possess true differences in opinions is still informative since their lack of agreement may not be due to poor survey construction, as demonstrated by a high IRR correlation. Ultimately, calculations using *ICCs* to determine if groups of judges systematically differ in their opinions demonstrates the need to further investigate the source or sources which lead to unreliability.

References

- ABET, Inc. (2005). *About ABET*. Retrieved March 15, 2005, from <http://www.abet.org/about.html>
- ABET, Inc. (Spring 2002). *Measuring success EC2000 longitudinal study gains board support and gathers momentum*. ABET Communications Link. Retrieved March 15, 2005, from <http://www.abet.org/news.html>
- ABET, Inc. (2004). *Sustaining the change: A follow-up report to the vision for change*, (pp. 1-10)
- Aldridge, M.D. & Benefield, L.D. (1998, May/June). A model assessment plan. *ASEE Prism*, 7(9), 22-26.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Banta, T.W. (Ed.). (1993). *Making a difference: Outcomes of a decade of assessment in higher education*. San Francisco: Jossey-Bass.
- Bordens, K.S. & Abbott, B.B. (1999). *Research design and methods: A process approach* (4th ed.). Mountain View, CA: Mayfield.
- Bozeman, D.P. (1997). Interrater agreement in multi-source performance appraisal: A commentary. *Journal of Organizational Behavior*, 18(4), 313-316.
- Brannick, M.T. (2003). Subjective Methods of Performance Measurement. Unpublished manuscript.
- Brennan, R.L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice*, 17(1), 8.
- Cohen, R.J. & Swerdlik, M.E. (2002). *Psychological testing and assessment: An introduction to tests and measurement*, (5th ed.). New York: McGraw-Hill.

- Coleman, G.D., VanAken, E.M., & Shen, J. (2002). Estimating inter-rater reliability of examiner scoring for a state quality award. *Quality Management Journal*, (9)4, 39-58.
- Crocker, L. & Algina, J. (1996). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group/Thomson Learning.
- Cronbach, L.J. (1977). *Educational psychology*, (3rd ed.). Harcourt Brace Jovanovich.
- Cronbach, L.J. (1970). *Essentials of psychological testing*, (3rd ed.). New York: Harper & Row
- Cronbach, L.J. (1990). *Essentials of psychological testing*, (5th ed.). New York: Harper & Row.
- Donald, J.G. & Denison, D.B. (2002, July/August). Quality assessment of university students: Student perceptions of quality criteria. *The Journal of Higher Education*, 72(4), 478-502.
- Dunlap, W.P., Burke, M.J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for r_{wg} and average deviation inter-rater agreement indexes. *Journal of Applied Psychology*, 88(2), 356-362.
- Dwyer, D. (1986). Inter-rater agreement and inter-rater reliability of rating scales. Unpublished manuscript.
- Feldt, L.S. & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational Measurement*, (3rd ed.), (pp. 105-146). New York: Macmillan.
- Fink, A. & Kosecoff, J. (1985). *How to conduct surveys: A step-by-step guide*. Newbury Park, CA: Sage.

- Fowler, F.J., Jr. (1993). *Survey research methods*, (2nd ed.). (Applied Social Research Methods Series Vol. 1). Newbury Park, CA: Sage.
- Goodwin, L.D. (2001). Inter-rater agreement and reliability. *Measurement in Physical Education and Exercise Science*, 5(1), 13-34.
- Gwet, K. (2001). *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters*. Gaithersburg, MD: Stataxis Publishing.
- Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, 34, 177-189.
- Hogan, T.P., Benjamin, A., & Brezinski, K.L. (2000). Reliability methods: A note on the frequency of use of various types. In B. Thompson (Ed), *Score reliability: Contemporary thinking on reliability issues* (chap. 4, p. 59). Newbury Park, CA: Sage.
- James, L.R., Demaree, R.G., & Wolf, G. (1993). r_{wg} : An assessment of within-group inter-rater agreement. *Journal of Applied Psychology*, 78(2), 306-309.
- Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Kenny, D.L (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review*, 98(2), 155-163.

- Kozlowski, S.W.J. & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling the issues of consistency versus consensus. *Journal of Applied Psychology*, 77(2), 161-167.
- Krueger, D.W. (1993). Total quality management. In T.W. Banta (Ed.), *Making a difference: Outcomes of a decade of assessment in higher education* (chap. 18, p. 269). San Francisco: Jossey-Bass.
- Lahey, M.A., Downey, R.G., & Saal, F.E (1983). Intraclass correlations: There's more than meets the eye. *Psychological Bulletin*, (93)3, 586-595.
- Landy, F.J. & Farr, J.L. (1980). Performance rating. *Psychological Bulletin*, 87(1), 72-107.
- Law, J.R. & Sherman, P.J. (1995). *Do raters agree? Assessing inter-rater agreement in the evaluation of air crew resource management skills*. Proceedings of the 8th International Symposium on Crew Resource Management. pp. 608-612.
- Lawlis, G.F. & Lu, E. (1972). Judgements of counseling process: Reliability, agreement, and error. *Psychological Bulletin*, 78(1), 17-20.
- Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563-575.
- Lenth, R.V. (2001). Some practical guidelines for effective sample-size determination. *The American Statistician*, 55, 187-193.
- Light, R.J., Singer, J.D., & Willett, J.B. (1990). *By design: Planning research on higher education*. Harvard University Press.

- Lindell, M.K., Brandt, C.J., & Whitney, D.J. (1999, June). A revised index of inter-rater agreement for multi-item ratings if a single target. *Applied Psychological Measurement*, 23(2), 127-115.
- Lindell, M.K. & Brandt, C.J. (1999). Assessing inter-rater agreement on the job relevance of a test: A comparison of the CVI , T , $r_{wg(J)}$, and $r^*_{wg(J)}$ indexes. *Journal of Applied Psychology*, 84(4), 640-647.
- Linn, R.L., Baker, E.L., & Dunbar, S.B. (1992). *Complex Performance-based Assessment: Expectations and Validation Criteria* (CSE Tech. Rep. 331). Los Angeles: University of California Center for Research on Evaluation, Standards, and Student Testing.
- Linn, R.L. & Gronlund, N.E. (2000). *Measurement and assessment in teaching*, (8th ed.). Upper Saddle River, NJ: Prentice Hall.
- Linn, R.L. (Ed.). (1989). *Educational measurement*, (3rd ed.). New York: Macmillan.
- McGraw, K.O. & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30-46.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement*, (3rd ed.), (pp. 13-103). New York: Macmillan.
- Murphy, K.R. & Davidshofer, C.O. (2001). *Psychological testing: Principles and applications*, (5th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Nichols, D.P. (1998). *Choosing an intraclass correlation coefficient*. (SPSS Keywords, No. 67). Retrieved July 26, 2005, from <http://www.utexas.edu/its/rc/answers/spss/spss4.html>

Peterson, G.D. (1997, September). Engineering Criteria 2000: A bold new change agent.

ASEE Prism, 7, 30-34.

Psychometrics. (2005). In *Britannica Concise Encyclopedia*.

Qualls, A.L. & Moss, A.D. (1996). The degree of congruence between test standards and test documentation within journal publications. In B. Thompson (Ed), *Score reliability: Contemporary thinking on reliability issues* (chap. 12, p. 191).

Newbury Park, CA: Sage.

Questionnaires. (n.d.). Retrieved October 18, 2000, from

<http://at.jji.de/USINACTS/tutorial/quest.html>

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.

Salkind, N.J. (2000). *Statistics for People Who (Think They) Hate Statistics*. Newbury Park, CA: Sage.

Sanders, M.S. & McCormick, E.J. (1993). *Human Factors in Engineering and Design*, (7th ed.). New York: McGraw-Hill.

Sawilosky, S.S. (2000). Reliability: Rejoinder to Thompson and Vach-Haase. In B. Thompson (Ed), *Score reliability: Contemporary thinking on reliability issues* (chap. 9, p. 149). Newbury Park, CA: Sage.

Schmidt, F.L. & Hunter, J.E. (1989). Inter-rater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology*, 74(2), 368-370.

Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability theory. *American Psychologist*, 44(6), 922-932.

- Shrout, P E & Fleiss, J L (1979) Intraclass correlations Uses in assessing rater reliability *Psychological Bulletin*, 86(2), 420-428
- Stemler, S E (2004) A comparison of consensus, consistency, and measurement approaches to estimating inter-rater reliability *Practical Assessment Research & Evaluation* (9)4 Retrieved May 17, 2005, from [http //PAREonline net/getvn asp?v=9&n=4](http://PAREonline.net/getvn.asp?v=9&n=4)
- Suskie, L A (1992) *Questionnaire Survey Research What Works* (Association for Institutional Research Resources for Institutional Research, No 6) Tallahassee, FL
- Thompson, B (Ed) (2003) *Score reliability Contemporary thinking on reliability issues* Newbury Park, CA Sage
- Thompson, B (2003a) Understanding reliability and coefficient alpha, really In B Thompson (Ed), *Score reliability Contemporary thinking on reliability issues* Newbury Park, CA Sage
- Thompson, B (2003b) A brief introduction to generalizability theory In B Thompson (Ed), *Score reliability Contemporary thinking on reliability issues* (chap 3, p 43) Newbury Park, CA Sage
- Tinsley, H E A & Weiss, D J (1975) Inter-rater reliability and agreement of subjective judgements *Journal of Counseling Psychology* 27(4), 358-376
- Trochim, W M K (1999) Types of reliability In *Research Methods Knowledge Base*, (2nd ed) Retrieved on March 30, 2000, from [http //trochim human cornell edu/kb/reliabl.html](http://trochim.human.cornell.edu/kb/reliabl.html)

Uebersax, J. (2003, May 10). *Intraclass correlation and variance component methods*.

Retrieved on May 26, 2004, from

<http://ourworld.compuserve.com/homepages/jsuebersax/icc.html>

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. In B. Thompson (Ed), *Score reliability: Contemporary thinking on reliability issues* (chap. 11, p. 173).

Newbury Park, CA: Sage.

Yaffee, R.A. (1998, March 11). *Reliability analysis enhancement: An SPSS application of intraclass correlation*. Retrieved on December 10, 2002, from

<http://www.nyu.edu/its/socsci/Docs/intraccls.html>

Appendix A

Accreditation Board of Engineering and Technology (ABET)

ABET is the recognized accreditation agency for programs leading to degrees in applied science, computing, engineering, and technology at colleges and universities across the nation. For over 70 years, ABET has provided leadership and advancement of education in the engineering community which includes industry, academe, and government and embodies a federation of 32 professional and technical societies representing these fields. ABET currently accredits over 2,500 programs at 550 colleges and universities in the U.S. It has been recognized by CHEA since 1997 as a specialized professional accrediting organization that accredits programs, not institutions. As one of the most respected accreditation agency in the U.S., most state licensing boards and certification programs require graduation from an ABET-accredited program in order to practice professionally (ABET, 2005, About ABET).

Engineering Criteria 2000 (EC2000)

In 1994, ABET responded to industry and educational leaders concerned about engineering graduates not being adequately prepared to perform in a new, modern engineering environment. These forward-thinking leaders believed that graduates lacked skills such as the ability to work in a team and communicate effectively, and they had little knowledge of customer service, continuous quality improvement, and global and societal responsibilities. At the same time, ABET's accreditation criteria and process was criticized by its constituents as too rigid and prescriptive thus, discouraging program innovation to prepare graduates facing a new working environment. So in 1994, ABET held three consensus-building workshops, which included members of professional and

technical engineering societies, university presidents, deans, and faculty, engineers in industry and private practice, government researchers and regulators, state engineering licensure and registration board members, and ABET leaders and board members. With the help of workshop participants from all facets of the engineering community, in 1997, ABET instituted a strategic plan and revolutionary engineering criteria to continuously address the most current challenges engineering graduates face and the tools needed to meet them (ABET, Inc., 2004). The central focus of ABET's newly adopted outcomes-based Engineering Criteria 2000 (EC2000) is on what is learned rather than what is taught. This new assessment process is more flexible, informed by the mission and goals of its individual programs and institutions, and reinforces the call for continuous improvement (ABET, 2005, About ABET). Today, EC2000 is applied in the evaluation of all ABET accredited programs. To further its commitment to continuous improvement and assuring quality, ABET initiated a longitudinal study focusing on the preparedness of graduates educated under EC2000 and the efficacy of the new criteria. In a sense, EC2000 and ABET must itself be evaluated to determine if its new goals have been achieved and to what extent. The study will address some key questions: Is the preparation of engineering students entering the profession improving? Are the criteria helping institutions better prepare students to enter the profession? Are students better prepared today under the new outcomes-based criteria than under the conventional criteria? The length of the study will provide benchmarking measurements, a baseline and end points, for longitudinal analysis in the years ahead as well as ongoing feedback on the progress of the new criteria in accomplishing its goals (ABET Communications Link, 2002).

The criteria themselves have been designed by ABET to emphasize student learning outcomes and the related institutional processes that promote the value of assessment, ensure the fulfillment of educational objectives, create an atmosphere of continuous quality improvement, and encourage self-accountability (ABET Communications Link, 2002). Accordingly, the eight criteria may be viewed as three subcategories:

1. student learning outcomes criteria (items a-k under General Criteria 3)
2. institutional process criteria (the remaining items in General Criteria 1 through 7)
3. program-specific criteria (the final General Criteria 8)

The new engineering criteria (a-k) under Criteria 3-Program Outcomes and Assessment departs from such rigid requirements as the 16-credit hours of engineering design, yet still maintains the traditional core of math, science, and engineering requirements.

However, an important new skill set has been added that includes teamwork and communication, global and societal awareness, ethical responsibility, and continuing education (ABET, Inc., 1993; Peterson, 1997). In designing EC2000, ABET emphasized the practice of continuous improvement through a broader approach focusing on input of constituencies, processes, and linking program objectives to resources such as students, faculty, facilities, and institutional and financial support. These resource requirements constitute the other half of EC2000. And finally, Criteria 8-Program Criteria are more discipline specific in nature. Certain engineering programs must satisfy a number of additional requirements related to curricular topics and faculty qualifications (ABET, Inc., 2003; Aldridge & Benefield, 1998).

ABET ENGINEERING CRITERIA SURVEY

GENERAL QUESTIONS – This section is to be completed by all respondents

1. Current Profession and Position (Check all that apply):

Practicing Engineer and/or Engineering Management

- ☐ Corporate Vice President, Director, etc.
☐ Chief Engineer, Senior Engineer, etc.
☐ Project Manager, Manager, Supervisor, etc.
☐ Other: _____

Engineering Education

- ☐ Graduate Faculty Member in an Engineering Program
☐ Undergraduate Faculty Member in an Engineering Program
☐ Dean/Chair of an Undergraduate Engineering Program
☐ Other: _____

Other (includes non-engineering positions)

- ☐ Describe your profession/position: _____

2. Total number of years in the engineering profession (includes teaching engineering curriculum):

- ☐ Less than one year ☐ 1-3 years ☐ 4-6 years ☐ 7-9 years ☐ 10+ years

3. Highest degree earned (Mark only one):

- ☐ High School diploma
☐ Associate's degree
☐ Bachelor's degree
☐ Master's degree
☐ Ph.D
☐ Other: _____

GENERAL INSTRUCTIONS FOR REMAINDER OF THE SURVEY

The survey is divided into three parts indicating a specific profession, as shown below. Please complete the part(s) that most closely applies to your profession. If more than one section applies to you, complete each one.

Part A (page 2) – Practicing Engineers and/or Engineering Management

Part B (page 3) – Faculty Members in a Graduate-Level Engineering Program

Part C (page 4) – Deans/Chairs and Faculty Members in an Undergraduate-Level Engineering Program

PART A
PRACTICING ENGINEERS AND/OR ENGINEERING MANAGEMENT
This section is to be completed by those currently employed as engineers

Directions: For the questions below, consider the duties required of any entry-level engineering position at your company. Also consider individuals whom you have worked with or supervised who had recently completed a four-year undergraduate engineering program from an ABET-accredited institution and had no prior experience in the field.

1. How much of your time do you spend with recent engineering graduates?

☐ Almost All ☐ Most ☐ Some ☐ Little ☐ Almost None

2. For each general skill below, circle the numerical value in each column that most accurately describes:

How Critical: How critical the skill is for recent engineering graduates to be successful at an entry-level engineering position at your company.

Performance: How well the recent engineering graduates you have worked with in the last year meet your company's expectations for an entry-level engineering position.

SKILL	HOW CRITICAL TO SUCCESSFUL PERFORMANCE IN THE POSITION					ACTUAL PERFORMANCE IN THE POSITION				
	<i>Not Required</i> 1	2	<i>Useful</i> 3	4	<i>Essential</i> 5	<i>Needs Improvement</i> 1	2	<i>Meets</i> 3	4	<i>Exceeds</i> 5
Ability to design and conduct experiments, as well as to analyze and interpret data	1	2	3	4	5	1	2	3	4	5
Ability to design a system, component, or process to meet desired needs	1	2	3	4	5	1	2	3	4	5
Ability to apply knowledge of mathematics, science, and engineering	1	2	3	4	5	1	2	3	4	5
Ability to identify, formulate, and solve engineering problems	1	2	3	4	5	1	2	3	4	5
Understanding of professional and ethical responsibility	1	2	3	4	5	1	2	3	4	5
Ability to function on multi-disciplinary teams	1	2	3	4	5	1	2	3	4	5
Broad education necessary to understand the impact of engineering solutions in global and societal context	1	2	3	4	5	1	2	3	4	5
Recognition of the need for, and an ability to engage in life-long learning	1	2	3	4	5	1	2	3	4	5
Ability to communicate effectively	1	2	3	4	5	1	2	3	4	5
Knowledge of contemporary issues	1	2	3	4	5	1	2	3	4	5
Ability to use the techniques, skills, and modern engineering tools necessary for engineering practice	1	2	3	4	5	1	2	3	4	5

3. What other skills and/or qualities do you expect an entry-level engineer at your company to possess?

**If you are not in education, please stop here and bring your
completed survey to the 2000 Summer Annual Meeting.
Thank you for your time and assistance!**

PART B GRADUATE FACULTY MEMBERS

This section is to be completed by those currently teaching engineering at the graduate level

Directions: For the questions below, consider students who recently entered a graduate-level engineering program at your institution, who had just completed a four-year undergraduate engineering program from an ABET-accredited institution and had no prior experience in the field.

1. For each general skill listed below, circle the numerical value in each column that most accurately describes:

How Critical: How critical the skill is for entering students to be successful at a graduate-level engineering program at your school.

Performance: How well the most recent entering students meet your expectations for graduate-level engineering coursework.

SKILL	HOW CRITICAL TO SUCCESSFUL PERFORMANCE IN THE PROGRAM					ACTUAL PERFORMANCE IN THE PROGRAM				
	<i>Not Required</i>		<i>Useful</i>		<i>Essential</i>	<i>Needs Improvement</i>		<i>Meets</i>		<i>Exceeds</i>
	1	2	3	4	5	1	2	3	4	5
Ability to design and conduct experiments, as well as to analyze and interpret data	1	2	3	4	5	1	2	3	4	5
Ability to design a system, component, or process to meet desired needs	1	2	3	4	5	1	2	3	4	5
Ability to apply knowledge of mathematics, science, and engineering	1	2	3	4	5	1	2	3	4	5
Ability to identify, formulate, and solve engineering problems	1	2	3	4	5	1	2	3	4	5
Understanding of professional and ethical responsibility	1	2	3	4	5	1	2	3	4	5
Ability to function on multi-disciplinary teams	1	2	3	4	5	1	2	3	4	5
Broad education necessary to understand the impact of engineering solutions in global and societal context	1	2	3	4	5	1	2	3	4	5
Recognition of the need for, and an ability to engage in life-long learning	1	2	3	4	5	1	2	3	4	5
Ability to communicate effectively	1	2	3	4	5	1	2	3	4	5
Knowledge of contemporary issues	1	2	3	4	5	1	2	3	4	5
Ability to use the techniques, skills, and modern engineering tools necessary for engineering practice	1	2	3	4	5	1	2	3	4	5

2. What other skills and/or qualities do you expect entering graduate students at your school to possess?

**If you are not a Dean/Chair or a faculty member of an undergraduate engineering program,
please stop here and bring your completed survey to the 2000 Summer Annual Meeting.
Thank you for your time and assistance!**

PART C

DEANS/CHAIRS & FACULTY MEMBERS OF AN UNDERGRADUATE ENGINEERING PROGRAM

This section is to be completed by those currently overseeing or teaching an undergraduate level engineering program at a postsecondary institution.

1. Consider the activities of your institution's undergraduate engineering program(s). For each item below, circle the numerical value in each column that most accurately describes:

How Critical: How critical the item is to your program's overall effectiveness in preparing graduates for the practice of engineering at a professional level.

Implementation: Indicate the degree to which this item is currently functioning in your program(s).

ITEM	HOW CRITICAL TO PROGRAM'S EFFECTIVENESS					DEGREE OF IMPLEMENTATION				
	<i>Not Required</i> 1	2	<i>Useful</i> 3	4	<i>Essential</i> 5	<i>Not Defined</i> 1	2	<i>Defined</i> 3	4	<i>Fully Functional</i> 5
Evaluation, advisement, and monitoring of students in order to determine the program's success in meeting its objectives.	1	2	3	4	5	1	2	3	4	5
Existence and enforcement of policies for acceptance of transfer students and for the validation of courses taken for credit elsewhere.	1	2	3	4	5	1	2	3	4	5
Existence and enforcement of procedures to assure students meet program requirements.	1	2	3	4	5	1	2	3	4	5
Detailed, published educational objectives that are consistent with the mission of the institution and ABET accreditation criteria.	1	2	3	4	5	1	2	3	4	5
A process based on the needs of the program's various constituencies in which the objectives are determined and periodically evaluated.	1	2	3	4	5	1	2	3	4	5
Curriculum and processes that ensure achievement of the program's objectives.	1	2	3	4	5	1	2	3	4	5
A system of ongoing evaluation that demonstrates achievement of the program's objectives and uses the results to improve effectiveness of the program.	1	2	3	4	5	1	2	3	4	5
An assessment process, with documented results, which demonstrates that the outcomes important to the mission of the institution and the objectives of the program are being measured.	1	2	3	4	5	1	2	3	4	5
Application of assessment results to the further development and improvement of the program.	1	2	3	4	5	1	2	3	4	5
A curriculum that culminates in a major design experience based on the knowledge and skills acquired in earlier course work and incorporating engineering standards and realistic constraints.	1	2	3	4	5	1	2	3	4	5
A curriculum that includes one year of college level mathematics and basic sciences (some with experimental experience) appropriate to the discipline.	1	2	3	4	5	1	2	3	4	5
A curriculum that includes one and a half years of engineering topics, consisting of engineering sciences and engineering design appropriate to the student's field of study.	1	2	3	4	5	1	2	3	4	5
A curriculum that includes a general education component that complements the technical content and is consistent with the program and institution objectives.	1	2	3	4	5	1	2	3	4	5

CONTINUED ON THE FOLLOWING PAGE...

DEANS/CHAIRS & FACULTY MEMBERS OF AN UNDERGRADUATE ENGINEERING PROGRAM

This section is to be completed by those currently overseeing or teaching an undergraduate level engineering program at a postsecondary institution.

(Continued from previous page)

ITEM	HOW CRITICAL TO PROGRAM'S EFFECTIVENESS					DEGREE OF IMPLEMENTATION				
	<i>Not Required</i> 1	2	<i>Useful</i> 3	4	<i>Essential</i> 5	<i>Not Defined</i> 1	2	<i>Defined</i> 3	4	<i>Fully Functional</i> 5
Faculty which are of sufficient number and have the competencies to cover all the curricular areas of the program.	1	2	3	4	5	1	2	3	4	5
Sufficient faculty to accommodate adequate levels of student-faculty interaction, student advising and counseling, university services activities, professional development, and interaction with industrial and professional practitioners and employers, as well as employers of students.	1	2	3	4	5	1	2	3	4	5
Classrooms, laboratories, and associated equipment that are adequate to accomplish program objectives and provide an atmosphere conducive to learning.	1	2	3	4	5	1	2	3	4	5
Computing and information infrastructures which support the scholarly activities of the students and faculty and the educational objectives of the institution.	1	2	3	4	5	1	2	3	4	5
Institutional support, financial resources, and constructive leadership that are adequate to assure the quality and continuity of the program.	1	2	3	4	5	1	2	3	4	5
Sufficient resources to attract, retain, and provide for the continued professional development of a well-qualified faculty.	1	2	3	4	5	1	2	3	4	5
Sufficient resources to acquire, maintain, and operate facilities and equipment appropriate for the engineering program.	1	2	3	4	5	1	2	3	4	5

2. What other activities do you feel are essential to overall program effectiveness?

This image shows a single sheet of white paper with horizontal blue or grey ruling lines. The lines are evenly spaced and run across the width of the page. There are approximately 20 lines visible. The paper has a slightly textured appearance and some minor discoloration or shadows, suggesting it's a physical scan of a real object. There is no handwriting or other markings on the paper.

Thank you for your time and assistance!

May 22, 2000

Dear ASME member,

The Accreditation Board for Engineering and Technology (ABET), the organization that accredits engineering and technology programs at over 500 colleges and universities in the United States, is requesting your assistance. Recognizing the challenges facing engineers to meet the needs of technology and the industry in the 21st century, ABET rewrote a new set of criteria for accreditation. You may already be familiar with these new criteria, known as Engineering Criteria 2000 (EC2000), which were adopted in 1997 for full implementation in the year 2000.

In an effort to embrace continuous quality improvement, ABET has undertaken a longitudinal study of the new criteria and their impact on an engineering graduate's preparation for practice. As a member of ASME, you can contribute valuable information to this ongoing investigation. The attached survey is part of a pilot study to assess EC2000. Please aid us in this endeavor by completing the survey, which should take no longer than 20 minutes of your time. Your candid responses are important. You need not give your name, however, we ask that you volunteer your name below in case we wish to contact you to help us refine the survey instrument itself (question wording, clarity, etc.).

The survey will be collected at the ASME Summer Annual Meeting, June 4-8, 2000, so you may wish to complete it now then bring it with you to Providence. Your assistance is greatly appreciated!

Sincerely,

The Accreditation Board for Engineering and Technology

Contact Information:

Name: _____

Phone: _____

E-mail: _____

February 2, 2001

Dear «TITLE» «LAST_NAME»,

The Accreditation Board for Engineering and Technology (ABET), the organization that accredits engineering and technology programs at over 500 colleges and universities in the United States, is requesting your assistance. Recognizing the challenges facing engineers to meet the needs of technology and the industry in the 21st century, ABET rewrote a new set of criteria for accreditation. You may already be familiar with these new criteria, known as Engineering Criteria 2000 (EC2000), which were adopted in 1997 for full implementation in the year 2000.

In an effort to embrace continuous quality improvement, ABET has undertaken a longitudinal study of the new criteria and their impact on an engineering graduate's preparation for practice. As a member of «ORG», you can contribute valuable information to this ongoing investigation. The attached survey is part of a pilot study to assess EC2000. Please aid us in this endeavor by completing the survey, which should take no longer than 20 minutes of your time. Also attached is a feedback form to help us refine the survey instrument itself. Your candid responses are important.

Be assured that all of your responses are completely confidential. A postage-paid envelope is included for your convenience. Please respond by **February 21, 2001**.

Sincerely,

Accreditation Board for Engineering and Technology

March 8, 2001

Dear «TITLE» «LAST_NAME»,

Recently, you received the Engineering Criteria Survey asking about the new set of criteria for accreditation, EC2000, which was rewritten and adopted by the Accreditation Board for Engineering and Technology (ABET) in 1997 and fully implemented last year. We know that you are busy, but we hope you can find time to fill out and return the enclosed survey. By doing so, you can contribute valuable information to ABET's ongoing investigation of its new criteria and their impact on an engineering graduate's preparation for practice.

The survey should take no longer than 20 minutes to complete. Also attached is a feedback form to help us refine the survey instrument itself. Be assured that all of your responses are confidential. A postage-paid envelope is included for your convenience. Please respond by **March 23, 2001**.

If you have already sent out your reply, kindly disregard this notice. Thank you!

Sincerely,

Accreditation Board for Engineering and Technology

The Accreditation Board for Engineering and Technology (ABET), the organization that accredits engineering and technology programs at over 500 colleges and universities in the United States, is requesting your assistance. Recognizing the challenges facing engineers to meet the needs of technology and the industry in the 21st century, ABET rewrote a new set of criteria for accreditation. You may already be familiar with these new criteria, known as Engineering Criteria 2000 (EC2000), which were adopted in 1997 for full implementation in the year 2000.

In an effort to embrace continuous quality improvement, ABET has undertaken a longitudinal study of the new criteria and their impact on an engineering graduate's preparation for practice. In your position, you can contribute valuable information to this ongoing investigation. The attached survey is part of a pilot study to assess EC2000. Please aid us in this endeavor by completing the survey, which should take no longer than 20 minutes of your time. Also attached is a feedback form to help us refine the survey instrument itself. Your candid responses are important.

Be assured that all of your responses are completely confidential.

Sincerely,

Accreditation Board for Engineering and Technology

Subject: Your feedback is requested on the ABET Engineering Criteria Survey

Last month, you completed the ABET Engineering Criteria Survey. The valuable information you supplied will assist ABET in a longitudinal study of its new criteria, EC2000. The survey was part of a pilot study, in which one goal was to produce a survey form that is usable, reliable and valid, and provides ABET with the information needed.

Will you please assist us once more by commenting on the survey instrument itself? Your input will help to ensure that in the future your colleagues will have little difficulty contributing to the on-going investigation through this survey. Please respond to the 10 questions listed below by answering yes or no, then elaborating as appropriate. Any recommendations you may have in the areas where the survey could be improved would be greatly appreciated. For reference, the same ABET Engineering Criteria Survey you completed is attached.

You may submit your feedback via return e-mail to: litz@aug.com. We look forward to hearing from you soon and thank you again for your assistance.

Sincerely,

Accreditation Board for Engineering and Technology

Were the questions clear?

Were you able to understand the directions provided?

Were the procedures simple?

Were you able to follow the general format easily?

Were the rating scales appropriate?

Did it take you 20 minutes or less to complete the survey?

Was the length of survey acceptable?

Was the survey easy to read?

Was there adequate space for making appropriate marks?

Was there adequate space for comments?

ABET ENGINEERING CRITERIA SURVEY INSTRUMENT FEEDBACK

The survey you just completed is intended to provide ABET with usable, reliable and valid information. Please help us to evaluate the survey instrument by giving your critique below. Your input will help to ensure that in the future your colleagues will have little difficulty contributing to the ongoing investigation through this survey.

Were the questions clear?

Were you able to understand the directions provided?

Were the procedures simple?

Were you able to follow the general format easily?

Were the rating scales appropriate?

Approximately how long did it take you to complete the survey?

Was the length of survey acceptable?

Was there adequate space for making appropriate marks?

Was there adequate space for comments?

Please offer any other suggestions you may have about the survey.

Thank you again for your time and assistance!

Appendix D

Judges	Current Profession/Position	Time spent w/ recent graduates (Industry only)	Survey Distribution Method
1	Industry: Other—Director of Engineering (Ret)	little	2 nd
2	Industry: Other—Professional Engineer	some	3 rd
3	Industry: Chief Engineer, Senior Engineer, etc.	some	3 rd
4	Industry: Project Manager, Manager, Supervisor, etc.	little	3 rd
5	Education: Graduate Faculty Member in an Engineering Program		1 st
6	Education: Dean/Chair of an Undergraduate Engineering Program		2 nd
7	Education: Graduate/Undergraduate Faculty Member in an Engineering Program		3 rd
8	Education: Graduate/Undergraduate Faculty Member in an Engineering Program		3 rd
9	Industry: Corporate Vice President, Director, etc.	most	2 nd
10	Industry: Other—Director of Engineering Projects	most	2 nd
11	Industry: Other—Military	most	2 nd
12	Industry: Other—Previously Practicing Engineering & Recently entered Graduate Faculty Member in an Engineering Program	some	3 rd
13	Education: Graduate Faculty Member in an Engineering Program		2 nd
14	Education: Graduate/Undergraduate Faculty Member in an Engineering Program		2 nd
15	Education: Graduate/Undergraduate Faculty Member in an Engineering Program		2 nd
16	Education: Graduate Faculty Member in an Engineering Program		2 nd

Note. The response scale consisted of “almost all”, “most”, “some”, “little”, and “almost none”. 1st = members of the Council on Education division of the American Society of Mechanical Engineers (ASME) who attended the ASME Summer Meeting; 2nd = entire membership of the Council on Education division of ASME and the Academic Affairs Committee of the American Institute of Aeronautics and Astronautics (AIAA) received the survey by mail; 3rd = personal contacts to various entities involved in the engineering field to include an engineering firm, a municipality, two government-sponsored agencies, and a university.