1-1-2005

Reliability generalization of the California Psychological Inventory

Helen Zaikina

University of Nevada, Las Vegas

Follow this and additional works at: <https://digitalscholarship.unlv.edu/rtds>

Repository Citation

Zaikina, Helen, "Reliability generalization of the California Psychological Inventory" (2005). *UNLV Retrospective Theses & Dissertations*. 2000.

<http://dx.doi.org/10.25669/nwyp-55n0>

This Thesis is protected by copyright and/or related rights. It has been brought to you by Digital Scholarship@UNLV with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you need to obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself.

This Thesis has been accepted for inclusion in UNLV Retrospective Theses & Dissertations by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.

RELIABILITY GENERALIZATON OF THE CALIFORNIA PSYCHOLOGICAL
INVENTORY

by

Helen Zaikina

Bachelor of Arts
California State University, San Marcos
2001

A thesis submitted in partial fulfillment
of the requirements for the

Master of Arts Degree in Psychology
Department of Psychology
College of Liberal Arts

Graduate College
University of Nevada, Las Vegas
May 2006

UMI Number: 1436813

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[®]

UMI Microform 1436813

Copyright 2006 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Thesis Approval

The Graduate College
University of Nevada, Las Vegas

December 15, 2005

The Thesis prepared by

Helen Zaikina-Montgomery

Entitled

Reliability Generalization of the California Psychological Inventory

is approved in partial fulfillment of the requirements for the degree of

Master of Arts in Psychology

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

ABSTRACT
**Reliability Generalization of the California Psychological
Inventory**

By

Helen Zaikina

Dr. N. Clayton Silver, Examination Committee Chair
Associate Professor of Psychology
University of Nevada, Las Vegas

The present study examined the reliability of the scores of the California Psychological Inventory (CPI). The reported reliability coefficients were located via PsycInfo and recorded into a data set. A multiple regression analysis was performed on the coefficients for the Socialization scale as well as on the three factors of the CPI. No predictors were found to be significant. The test retest reliabilities were significantly higher than the internal consistency reliabilities across all scale factors. A lack of significant results may be explained by the fact that the CPI is a multi-subscale instrument and not all of the scales are widely used to provide enough reliability coefficients in the literature.

TABLE OF CONTENTS

ABSTRACT.....	iii
CHAPTER 1 INTRODUCTION AND LITERATURE REVIEW	1
What is Reliability?	4
Three Key Concepts in Reliability	5
Factors Affecting Reliability	10
Types of Reliability Methods	11
What is Reliability Generalization and Why is it Important	15
Meta-Analysis: Advantages and Disadvantages	17
Reliability Generalization: Advantages and Limitations ...	21
Current Reliability Generalization Work	23
The California Psychological Inventory	26
CHAPTER 2 METHOD	29
Data	29
Procedure	31
CHAPTER 3 RESULTS	33
Descriptive Statistics	33
Socialization Subscale	35
Analyses of the Individual Factors	36
CHAPTER 4 DISCUSSION	39
REFERENCES	45
VITA	62

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

In light of the recent proposal of a new meta-analytic method by Vacha-Haase (1998), termed reliability generalization (RG), numerous studies have been published reporting the cumulative reliability of well-established personality tests and inventories (e.g., Beretvas, Meyers, & Leite, 2002; Caruso, 2000; Viswersvaran & Ones, 2000; Yin & Fan, 2000). This paper will provide a comprehensive definition of reliability generalization, from both theoretical and practical viewpoints, outline some problems with the method of conducting reliability generalization studies as well as meta-analysis in general, along with providing the results of a reliability generalization analysis of the California Psychological Inventory (Gough, 1957).

The current body of literature on reliability generalization, however, makes no mention of the origin of the term, which was conceptualized by Kennedy and Turnage (1991). They drew a parallel between validity and reliability generalization by demonstrating that reliability coefficients from 10 studies showed a constancy of reliability. Hence, their work provided evidence that reliability coefficients across studies are highly correlated and reliable. Kennedy and Turnage (1991) took their

concept of reliability generalization directly from the idea of validity generalization, which involves calculating the best estimate of validity. Similarly, reliability generalization attempts to compute the best estimate of reliability. This idea of reliability generalization is somewhat different from Vacha-Haase's (1998) definition. Vacha-Haase (1998) viewed reliability generalization, not only as a method of estimating reliability, but also as a method that characterizes the *sources* of the variability of the measurement variances across measures.

It has become apparent from the recent influx of literature on the subject of reliability generalization that the problem of referring to the reliability of an instrument has gained considerable attention from psychological and social behavioral researchers and may begin to be corrected in the literature as well as in the everyday language of the professional academic setting. Some researchers have pointed out that referring to the reliability of test scores as that of an instrument, may be a semantic mistake (Thompson, 1994). Unfortunately, this mistake is too often repeated in the literature when authors refer to the reliability of an instrument and report reliability coefficients. The issue of confusing the reliability of a particular set of test scores with the reliability of a measurement instrument with which those test scores were obtained may simply be one of sloppy speaking, according to Vacha-Haase (1998). This sentiment is echoed by Thompson (2002), when he pointed out that the issue of sloppy speaking may unconsciously lead to sloppy thinking, which

in turn leads to sloppy practice. Pedhazur and Schmelkin (1991) mentioned that the issue of reliability confusion and the practice of incorrectly generalizing a reliability coefficient to the whole instrument is the responsibility, at least partly, of doctoral programs across the country. However, many doctoral students do not undergo a thorough study of psychometric theory. The recent report by the American Psychological Association's (APA) Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999) pointed out the need for reporting reliability of test scores in each study for which a test is used. Wilkinson and colleagues (1999) stated,

It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test for a particular population of examinees. Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric.

Interpreting the size of observed effects requires an assessment of the reliability of the scores (p. 596).

Thus, it is evident that reliability generalization studies are necessary and important in psychological research. Before one can study reliability generalization, one must first define and examine the concept of reliability.

What is Reliability?

Many definitions of reliability exist in the current literature. According to Gronlund and Linn (1990), reliability is primarily statistical. They stated, that reliability does not refer to the instrument itself, but instead to the *results* obtained with an evaluation instrument. Thus, it would be more appropriate to refer to the reliability of the 'test scores' or the 'measurement' than of the 'test' or the 'instrument'. They also pointed out that reliability is one of the most important characteristics of test results because it: a) provides the consistency that makes validity possible; and b) indicates how much confidence we can place in our results.

The meaning of the term "reliability" can be construed to have different applications within the psychological realm, but as applied to testing and measurement, reliability can be defined in the following four general points. First, reliability refers to the results of a test instrument. Second, an estimate of reliability always refers to a type of consistency, (i.e., over time, different samples, different raters, etc.). It is not always the case that test scores are consistent in all of these aspects, hence, the process of reliability generalization can only measure a limited amount of existing consistencies. Third, reliability is not a sufficient condition for the existence of validity. Although the results of a given test may be highly consistent, they may be consistently measuring the wrong construct, thus not ensuring a high degree of validity. Finally,

reliability is generally statistical. Most of the time, reliability is expressed as a correlational value known as a reliability coefficient. Another method for reporting reliability is by means of standard error of measurement (Gronlund & Linn, 1990).

Crocker and Algina (1986) defined reliability in practical terms, stating that reliability is the consistency or reproducibility of test scores when the same individuals are measured on the same test under similar circumstances more than once. Because every score is made up of a true score and error score component, in which

$$X = T + E, \quad (1)$$

where X represents a given score, T represents the true score component and E represents the error score component. Therefore, every reliability coefficient relies on the correlation of the relationship between true and error scores.

Three Key Concepts in Reliability

The Reliability Index

Before defining the reliability coefficient itself, it is important to mention the reliability index. When administering a test, the test administrator has access only to the observed scores yielded by the test administration. What the test administrator is really interested in, however, are the *true* scores yielded by the test. The question, then becomes, how closely related are the observed scores and the true scores? The

reliability index is an indication of this relationship, it is a correlation coefficient that measures the relationship between the true and observed scores on a test (Crocker & Algina, 1986). Another way to think of the reliability index is as a correlational score that provides the correlation of any one test score with the average of all other test scores (Nunnally & Bernstein, 1994). Mathematically, it can be expressed as the ratio of the standard deviation of true scores to the standard deviation of the observed scores:

$$\rho_{XT} = \frac{\sigma_T}{\sigma_X}, \quad (2)$$

in which ρ_{XT} is the correlation between true and observed scores, σ_T is the standard deviation of true scores and σ_X is the standard deviation of the observed scores (Crocker & Algina, 1986). The reliability index, therefore, is a theoretical concept, which represents the correlation between the true scores and all possible observed scores from all possible repeated testings. This concept, however, has little practical value, because true scores are not directly observable and it is impossible to obtain all observed scores from all of the possible testings (Crocker & Algina, 1986). Instead of using a complex theoretical concept such as the reliability index, a more practical concept of the reliability coefficient is used.

Reliability Coefficient

Because all forms of reliability are obtained by either testing the subjects multiple times with the same test version or by

testing the subjects with parallel test versions, the reliability coefficient can be defined as the correlation between scores on parallel test forms. Here, the term "parallel test forms" is used to represent any instance of test administration in which: a) each examinee has the same true score on both forms of the test; and b) the error variances for the two forms are equal (Crocker & Algina, 1986). This definition is not exclusive to equivalent forms reliability. It can also be applied to coefficient alpha (α), split-half reliability, test-retest reliability, and test-retest with equivalent forms. In all of these cases, each examinee has the same true score on both forms and the error variances for both forms of the test are equal. Therefore, the reliability coefficient, can be defined as the correlation between scores on parallel test forms and is expressed as: ρ_{X1X2} (Crocker & Algina, 1986). Mathematically, it is possible to establish a relationship between ρ_{XT} , the reliability index, and ρ_{X1X2} , the reliability coefficient. The reliability coefficient is defined as the ratio of true score variance to observed score variance:

$$\rho_{X1X2} = \frac{\sigma_T^2}{\sigma_X^2}, \quad (3)$$

in which, ρ_{X1X2} is the proportion of observed score variance that may be attributed to variation in the examinees' true scores, σ_T^2 is the true score variance, and σ_X^2 is the observed score variance (Crocker & Algina, 1986). Another method for measuring and reporting reliability is the standard error of measurement.

The Standard Error of Measurement

Gronlund and Linn (1990) defined the standard error of measurement as the estimate of the expected amount of variation in test scores, if the tests were administered to the same person many times. Standard error of measurement is usually taken into account when interpreting test scores and is usually available in the test manuals. Thus, the standard error of measurement provides a different way to describe reliability than does the reliability coefficient, which provides the ratio of observed score variance to true score variance. The standard error of measurement describes how measurement error affects the interpretation of persons' test scores. A more technical definition of the standard error of measurement, based on the classical test theory is provided by Crocker and Algina (1986), who stated that each respondent's personal distribution of possible observed scores around their true score has a standard deviation and it is that the average of these standard deviations for the group is called the *standard error of measurement*. The standard error of measurement is mathematically related to the reliability coefficient in the following manner: If the standard error of measurement is the average of the *error standard deviations*, it can be denoted as σ_E and expressed in terms of an observed score equation as

$$\sigma_T^2 + \sigma_E^2 = \sigma_X^2, \quad (4)$$

expressed here in terms of variance. Standard deviation is, of course, the square root of variance. If both sides of this equation are divided by σ_x^2 , the equation becomes,

$$\frac{\sigma_T^2}{\sigma_x^2} + \frac{\sigma_E^2}{\sigma_x^2} = 1. \quad (5)$$

Recall from Equation 3 that the first term of the left side of the equation is the mathematical expression of the reliability coefficient, ρ_{X1X2} , and to solve for σ_E ,

$$\frac{\sigma_E^2}{\sigma_x^2} = 1 - \rho_{X1X2} \quad (6)$$

and

$$\sigma_E = \sigma_x \sqrt{1 - \rho_{X1X2}} \quad (7)$$

From the above formulae, it is evident that the reliability index, reliability coefficient, and the standard error of measurement are mathematically related to one another. Given that the maximum reliability value is 1.0, from the above formula, it can be seen that as the random error increases, the reliability coefficient will decrease. Thus, it is important for scores to be free from random error to be reliable. Random error, however, is not the only factor that has an effect on the reliability of scores. There are other factors that likewise affect score reliability.

Factors Affecting Reliability

Thus, it is evident that a few quantifiable factors affect test score reliability. There are also other factors that play a role in the scores' reliability outcome. One of these factors is group heterogeneity/homogeneity. If the group to which a test or a scale is administered is homogeneous, then the reliability of the yielded scores will be higher than if the test were administered to a heterogeneous group (Fan & Yin, 2003).

Another factor that affects score reliability is the performance level of a group at the time of the test administration. Because score reliability is directly linked to error score variance, it is reasonable to hypothesize that error score variance is affected by the groups' performance. In turn, the group's performance will affect the reliability of that group's scores (Fan & Yin, 2003).

A third factor that can affect the reliability of test scores is the translation of the scale into a language other than the original language in which the scale was constructed. Translated versions of the scale may yield consistently lower or higher reliability scores, which are problematic in a situation when score reliability is being generalized over various population parameters (Arce-Ferrer & Ketterer, 2003). Presently, there are quite a few methods of computing and reporting reliability and before the concept of reliability generalization can be discussed and understood it would be useful to review the various types of reliability.

Types of Reliability Methods

Several different methods exist to calculate test score reliability. Among these are methods of internal consistency, stability over time, equivalence, and inter-rater reliability.

The measures of internal consistency include coefficient alpha (α) and split-half reliability. The measures of stability over time include test-retest, test-retest with equivalent forms, and equivalence which is measured by equivalent forms reliability. Inter-rater reliability is usually measured by means of computing a Pearson product-moment correlation between the raters or by Cohen's Kappa (κ) (Cohen, 1960), in the case of categorical data.

Internal Consistency

Coefficient alpha or Cronbach's alpha (α) (Cronbach, 1951) is one of the most widely used measures of internal consistency. It is useful because it sets an upper limit for the reliability of test scores. If α is very low, the test may be too short or the items may have little in common and a revision of the test items would be necessary before obtaining other measures of reliability (Nunnally & Bernstein, 1994).

Another measure of internal consistency estimation is split-half reliability. In this form of reliability estimation, the test is given only once and the results of the equivalent halves of the test are correlated with each other. This method produces a simple Pearson product-moment correlation coefficient, r . The utility of this method consists of its ease to be computed by hand, however, with the wide availability of computers this

method has become almost obsolete. Although it is easily calculated by hand, split-half reliability allows us to compute a correlation between specific splits of only one version of a test; that is, per every calculation of this reliability, the test can only be split one way (i.e. the number correct of odd-items correlated with the number correct of even-items). Because Cronbach's α allows us to see the average of all possible correlations of the test split in many different ways and the advent of statistical computer software allows for a speedy computation of Cronbach's α , the split-half reliability is no longer a common way to estimate reliability (Nunnally & Bernstein, 1994).

Stability Over Time

Test-retest reliability is perhaps the most commonly used measure not only of stability over time, but also of reliability in general (Gronlund & Linn, 1990). It is a measure where the test is administered twice to the same individuals and the two sets of scores obtained from the different administrations of the test are correlated to compute the reliability coefficient (r). Although a commonly used method of reliability estimation because of the ease of administration, this method may be flawed because the second test administration is usually influenced by the memory of the first administration of the test and there may be carryover and practice effects. These usually occur because the memory of the first test influences the responses on the subsequent retest. Subjects might repeat their remembered

responses and make guesses similar to those that they made during the previous test administration (Nunnally & Bernstein, 1994). There are various recommendations in the literature as to the length of time that should elapse between the two administrations of a test in a test-retest situation and no one universal agreed upon time interval exists. The general recommendation for an average amount of time that should pass between the administrations of the two testing sessions is two weeks (Nunnally & Bernstein, 1994).

Test-retest reliability with equivalent forms is a combination of two methods of reliability estimation, and is a measure of both stability and equivalence. Two forms of a test are administered to the same group of participants with a specific time interval elapsing between the administrations of the two test forms. Of course, this is an improvement over the test-retest method, because it enables the researchers to control the possible carryover effects from the first administration, but this method is economically infeasible and time consuming, and therefore it is rarely used.

Equivalence

Equivalent forms reliability is a measure of correlation between scores obtained from the administration of two different forms of the same test. Although this method alleviates the problem of time passage between the administration of the test in methods such as test-retest, there is the economic problem of devising a different version of the test. In addition to this, the two test

versions must have equal means and variances as well as having similar correlations among other test variables. Because this method provides a measure of both stability and equivalence, this is the most rigorous method of reliability estimation, although it is rarely used due to the constraints stated above (Gronlund & Linn, 1990).

Inter-Rater Reliability

Another type of score reliability that is commonly reported in the literature is inter-rater reliability. This type of reliability refers to the component of the scores that is stable across raters. In cases of non-categorical data, a simple correlation is computed between the scores of the different raters. In cases when the data is categorical, a common measure of inter-rater reliability is Cohen's Kappa (Cohen, 1960). Cohen (1960) developed this measure to determine the extent to which obtained nominal scale categorizations are reliable. Usually, these categorizations are performed by people (e.g., clinical psychologists, social workers) and because their judgment may be faulty, this may lead to lower reliability. This measure is analogous to the concept of a coefficient of equivalence used with alternate forms testing, hence the judges can be compared to the alternate forms and the nominal data to scores on these forms. Thus, Kappa is a chance-corrected measure of the percentage of rater agreement. All of the above methods of estimating score reliability are commonly used in reliability generalization meta-analysis studies.

What is Reliability Generalization and Why is it Important?

Reliability generalization was first mentioned by Kennedy and Turnage (1991). These researchers applied the strategy of validity generalization (Schmidt, Pearlman, Hunter, & Hirsch, 1985) to assess test-retest reliability from nine studies. They found that the coefficients from the nine studies are consistent across test administrations and are generalizable. Vacha-Haase (1998) proposed a method of reliability generalization, that can be used to characterize the mean measurement error variance across studies. Vacha-Haase, Henson, and Caruso (2002) described this procedure as a meta-analytic procedure that explores the variability in reliability estimates while characterizing the source or sources of this variance. Both Kennedy and Turnage (1991) and Vacha-Haase (1998) outlined essentially the same methodology for conducting a reliability generalization study. In both articles, the authors stated that reliability generalization is an extension of a process called "validity generalization" (Hunter & Schmidt, 1990) and follows the same procedure as a validity generalization study, in which validity coefficients across studies are used as dependent variables and the studies are used as a unit of analysis with means, standard deviations and other descriptive statistics computed for these coefficients. When conducting their study, Kennedy and Turnage (1991), proposed that they were looking for a 'best' estimate of the variability of the coefficient among a set of reliabilities that are constrained by its own situational specifics. They

wanted to look at the extent to which reliability coefficients from nine studies of different populations will be constant in their relative rank positions. Vacha-Haase (1998) tried to characterize the mean measurement error variance across studies and demonstrated the extent to which reliability across studies is generalizable. The goals of the two studies were essentially the same, although, in the case of Vacha-Haase (1998), more emphasis was devoted to features of the studies under analysis that best predict the sources of measurement error across the reliability coefficients.

Essentially, a reliability generalization study is conducted by gathering all of the articles that report any type of a reliability coefficient for a particular psychological test or inventory, creating a data set consisting of all of the known reported coefficients, and then analyzing this data set, using the reliability estimates as the dependent variables.

The importance of reliability generalization studies lies in several different factors. The results of reliability generalization studies allow researchers to gain a deeper understanding of various factors that affect reliability and the construct of reliability in general. These aspects include the effect of reliability on statistical power, effect sizes, and populations in which the test under examination is to be used (Vacha-Haase, et. al., 2002). Reliability generalization studies also contribute to the administrators' knowledge about the test and allow the test administrators and test users to make better

decisions concerning the use of tests. Another important aspect of reliability generalization studies is that their recent appearance in the literature may draw attention from the psychological research community to the importance of reliability and may serve as a reminder to researchers currently conducting studies to report the reliability of the scores in their studies.

The effect of reliability on statistical power is an important aspect of testing. As reliability increases, so does the statistical power of the test. Perhaps more importantly, reliability has an impact on effect sizes, an issue that has been largely ignored in social behavioral research (Thompson, 2002). Effect sizes are tied to reliability because reliability depends on measurement error and measurement error attenuates the effect size. If the reliability is high, the measurement error is low and the effect size will be larger as compared to low reliability. Because the results of a reliability generalization study include studies with different samples, they can be helpful in determining populations for which a given test may be appropriate. Although reliability generalization is a useful and necessary procedure to help researchers better utilize tests, the meta-analytic procedure of a reliability generalization study contains both advantages and disadvantages.

Meta-Analysis: Advantages and Disadvantages

As the body of work of psychological and social behavioral studies grew, a new method to analyze the results of studies

dealing with the same issue appeared. The task of integrating the research findings from studies on the same topic is very complex if done in a subjective or qualitative fashion and the need for a quantitative method of accomplishing this task was great. Because hundreds of data sets exist on certain topics, the only "unbiased" way to make sense of this data and to integrate the findings is a quantitative methodology. In order to synthesize information on the same topic quantitatively, meta-analysis was developed. Since its inception in the late 1970's, meta-analysis has become a widely accepted statistical method. In fact, the number of meta-analytic studies that were found in the *PsycInfo* database from 1984 to 2004 has increased from 63 to 330.

Meta-analysis is a statistical procedure which involves analyzing the correlations or other statistical quantities (i.e. effect sizes, means) between various predictors and criterions, which in turn, permits the estimation of the mean validity and reliability while controlling for systematic sources of variance. Meta-analysis gives us an important and useful way of studying the extent to which predictor-criterion relationships are valid and generalizable (Schmidt, Pearlman Hunter, & Hirsch, 1985). This technique also allows us to determine the true variance across studies while getting rid of sampling and measurement error.

Because reliability coefficients are correlation coefficients, there are a few advantages and disadvantages inherent in this

meta-analytic procedure. Each reliability coefficient that is reported in a given study, carries with it an "artifact", a study imperfection due to human error; not something that occurs naturally (Hunter & Schmidt, 1990). Some of these artifacts include sampling error, error of measurement in the independent and dependent variables, deviation from perfect construct validity in the dependent and independent variables, reporting error, and variance due to extraneous factors (Hunter & Schmidt, 1990). When simple correlations are being considered in a meta-analytic procedure, one can eliminate most of the artifacts prior to using the correlation coefficient in the analysis. In the case of reliability generalization, artifactual information is either provided sporadically or not at all, thus eliminating the effect of the artifacts prior to analysis is virtually impossible. Hence, the data are analyzed without prior control for the effect of the artifacts. This is one of the disadvantages of using the meta-analytic method in a reliability generalization study.

Another disadvantage of meta-analysis is availability bias. According to Hunter and Schmidt (1990), availability bias is the argument that the studies obtained for meta-analysis are the only studies available in the literature, because only the studies that contain statistically significant results are published. In the case of reliability generalization, this may not be as severe of an issue as one might suspect. Most studies that are used for reliability generalization analysis are not ones which are

performed specifically for the purpose of establishing the reliability of the scores in question. The reliability coefficients that are published in these studies usually possess quite a large range and are reported because of the general requirement to report the reliability of the scores that are obtained in the particular assessment procedure done in the study. Moreover, this range of reliability coefficients allows us to better assess under which populations and circumstances the instrument under study is most appropriate (Hunter & Schmidt, 1990).

Another general criticism of meta-analysis is that it combines studies that are incompatible with one another for comparison. This criticism has been made against meta-analysis that combines studies with different independent and dependent variables; that is, the dependent and independent variables across studies vary in their constructs (Hunter & Schmidt, 1990). Sometimes this criticism extends to meta-analytic studies that combine studies in which the dependent and independent variables have the same construct. In a reliability generalization procedure, however, we are not interested in a meta-analysis of dependent or independent variables, instead we are interested in analyzing the reported reliability coefficients that can be seen almost as the byproducts of the studies under analysis.

Reliability Generalization: Advantages and Limitations

The procedure of reliability generalization contains within itself inherent advantages and disadvantages, as is the case with any existing statistical procedure. The practice of reliability generalization contains some obvious advantages. One of the advantages of the reliability generalization methodology is that it synthesizes the findings of published studies in one article, making those findings more accessible to researchers and practitioners who may use the test in their work. Instead of dealing with the task of an extensive literature review, a researcher can now reference only one article if they are interested in the reliability of the scores yielded by a particular instrument. Another advantage of the reliability generalization procedure is that different types of reliability coefficients may be used in this type of meta-analysis. Because there are many different types of reliability available to the researcher to be calculated, different types of reliability are reported in the literature. Reliability generalization not only allows us to compare different types of reliability coefficients in the same analysis (i.e. ANOVA, regression analysis) but also to analyze across different coefficients.

Along with its advantages, the process of reliability generalization contains some disadvantages. One of the disadvantages of the procedure is that it only encompasses a limited amount of consistency aspects. This disadvantage stems from the meta-analytic nature of the process of reliability

generalization in which only reliability coefficients that are available for use in the analysis are the ones reported in the literature. In turn, the types of reliability coefficients that are measured and reported in the literature are determined by the researchers in every individual study, and are therefore out of the scope of the reliability generalization researcher's control. Another disadvantage of reliability generalization studies relates to the previously mentioned one. In the process of reliability generalization we are sometimes unable to include low reliability coefficients in the analysis because they are often not reported in the literature. Once again, this issue of reporting bias is out of the scope of control of the reliability generalization researcher. Although this particular issue has not been addressed with regard to reliability generalization, Schmidt, et al. (1985) have addressed it with regard to validity generalization. They pointed out that evidence in the literature shows that reporting bias does not exist because a search of published and unpublished studies will produce a representative set of data for a validity generalization analysis. This argument, however, is not applicable to the process of reliability generalization. Most of the time when a reliability generalization study is conducted, the authors of such a study do not have access to unpublished coefficients, which are usually produced with samples for whom the test is not appropriate (i.e., clinical populations, children, foreign populations). Hence, these types of samples cannot be included in the analysis, making

the results of a reliability generalization study potentially erroneous and non-representative.

Another criticism of the reliability generalization methodology has been proposed by Sawilowsky (2000). He asserted that although the procedure of reliability generalization has been invented with the purpose of investigating the reliability of scores across studies, the emphasis being put on the scores themselves, this does not occur in a reliability generalization study. What actually happens is that instead of examining an individual's *score* or a group's *set of scores* on the same test used in different studies, the reliability coefficients obtained for a test that was used by researchers in independently conducted studies are analyzed (Sawilowsky, 2000, p. 167).

Current Reliability Generalization Work

Since Vacha-Haase's (1998) resurrection of the concept of reliability generalization, quite a number of studies have been published that have performed a reliability generalization analysis on a widely used assessment instrument, questionnaire or scale (e.g., Beretvas, Meyers, & Leite, 2002; Caruso, 2000; Shields & Caruso, 2004; Viswesvaran & Ones, 2000). The reliability generalization studies that have been published to date have mainly utilized two types of analyses: ANOVA and multiple regression, although such analyses as correlations between the predictor variables and reliability coefficients have also been used.

ANOVA Analyses

Caruso (2000) published a reliability generalization study in which he analyzed the reliabilities of fifty-one samples employing the NEO personality scales. In his study, Caruso (2000) computed mean reliabilities and transformed them using Fisher's z transformations (Fisher, 1921). The z -scores were analyzed and later back transformed to correlation coefficients. The author used bivariate correlations between continuous sample characteristics and score reliability. ANOVA analyses and η^2 effect size measures were calculated for discrete (nominal) sample characteristics (Caruso, 2000). Scores from the Neuroticism scale appeared to be the most reliable of the scales with a mean reliability of .88. Scores from the Openness and Agreeableness scales were found to be least reliable with mean reliabilities of .77 and .73, respectively (Caruso, 2000).

Multiple Regression Analyses

Shields and Caruso (2003) published a study in which they examined the reliability of scores from the Alcohol Use Disorders Identification Test (AUDIT). Their study had two goals: to characterize the typical reliability of scores for the AUDIT, and to examine factors that may be related to the reliability of those scores. In this study, the authors employed multiple regression as a means of analyzing the data. Score variability, sample type, and proportion of male scale respondents were used as predictor variables and score reliability was used as the outcome variable (Shields & Caruso, 2003). The authors found

that score variability predicted a large, statistically significant amount of variance in score reliability and that none of the other predictors contributed significantly to this prediction (Shields & Caruso, 2003).

Correlational Analyses

A few reliability generalization studies have used various correlational analyses to analyze their data. For example, Yin and Fan (2000) have performed correlational analysis in their study, which involved assessment of the reliability of Beck Depression Inventory (BDI) scores. First, the authors calculated the standard errors of measurement (SEMs) for every source of variation (e.g., BDI form, BDI language, age range, etc.). The SEMs were then correlated with the corresponding reliability coefficients and the standard deviation of the scores in one of the analyses. In a second correlational analysis, the test-retest reliability coefficient was correlated with the length of interval between test and retest. The authors found that the measurement error associated with time (test-retest) is significantly larger than the error associated with internal consistency. However, the authors did not find any significant correlations between SEMs and their associated reliability coefficient of the BDI (Yin & Fan, 2000).

In another correlational study, Shields and Caruso (2004) examined the CAGE questionnaire, another widely used alcohol screening instrument. In this study, the authors calculated bivariate correlations between predictor variables, such as

gender, ethnicity, sample type and score reliability coefficients. None of the correlations were found to be statistically significant with the exception of the correlation between score reliability and participants' age (Shields & Caruso, 2004). The present study used a combination of some of the discussed analyses. Before the analyses in the present study are discussed, however, it is necessary to give a brief overview of the instrument in which reliability coefficients will be used in this study.

The California Psychological Inventory

The instrument under analysis in this study is the California Psychological Inventory (CPI); originally developed by Gough (1957). The test was designed to assess interpersonal personality traits within a normal population (Groth-Marnat, 1990). The California Psychological Inventory is a self-administered paper-and-pencil test comprised of 462 true-false statements. Although the test can be administered individually, it was originally designed for group administration. The test has been used to evaluate adults, although it was originally constructed for use with younger adults having a minimum of fourth-grade reading ability. The items of the CPI were designed to gather information about an individual's typical behavior patterns, usual feelings, opinions, and attitudes regarding social, ethical, and family matters (Groth-Marnat, 1990). The results of the test are plotted on 18 scales. The first set of

scales measure social factors (e.g., sociability), the middle set of scales measure internal qualities (e.g., responsibility), and the last set of scales measure "broadly stylistic variables related to different functional models" (e.g. flexibility) (Gough, 1957). The scales have also been divided by Gough (1987) into five statistically derived factors. Factor 1 (Dominance, Capacity for Status, Sociability, Social Presence, Self-acceptance, Independence, and Empathy) indicates one's level interpersonal effectiveness and social ability. Factor 2 (Responsibility, Socialization, Self-control, Tolerance, Good Impression, Communality, and Sense of Well Being) indicates a general level of mental well-being, ability for adjustment and social conformity. Factor 3 (Achievement via Independence, Flexibility, Tolerance, Intellectual Efficiency, and Psychological Mindedness) assesses the extent to which a person can think and behave independently. Factor 4 (Communality, Responsibility, Socialization, and Sense of Well-being) measures the extent of a person's adherence to social norms and expectations. Finally, Factor 5 which is composed of the Femininity/Masculinity subscale measures one's level of sensitivity and dependency. The test scores are reported in a T-scores format, with a mean of 50 and a standard deviation of 10. Gough (1957) constructed the items of the California Psychological Inventory based on what he referred to as "folk concepts", which are personality concepts that are relevant throughout different cultures, are easily understood, and have

"functional validity". Hence, Gough (1957) attempted to make the test more understandable, thereby augmenting the predictability of behavior.

CHAPTER 2

METHOD

Data

The data for the present study were the reliability coefficients from studies that administered the California Psychological Inventory or one of its subscales. These coefficients were the reliability coefficients for each of the eighteen subscales of the CPI.

A preliminary sampling of the data showed that there may not be enough reliability coefficients obtained for the analysis. At that time it was decided that some of the more recent publications' authors would be contacted and requested to provide a computed reliability estimate for their particular samples' subscales that were used. In the case that a computed reliability estimate was not available, the authors would be requested to provide their original data set for the purposes of computing any possible reliability estimates for their sample. The authors were assured that any information that they may provide would be used for the purposes of the meta-analysis in the current study as required by APA guidelines. It was determined that authors of studies published between the years of 1998 and 2004 would be contacted due to the fact that data is

usually stored for a period of up to seven years after the completion of the study. These criteria yielded 22 authors who were contacted. The response rate was approximately 85%, however, these responses yielded only six studies that were determined to be usable in the present analysis. The studies that were unusable were either from authors who no longer had their data sets, had incomplete data sets, or did not reply.

After the completion of the data collection, it was noted that some of the subscales were used more often than others (i.e., their reliabilities were computed and reported in the literature more often than the reliabilities of the other subscales). At this time it was noted that the most widely used scale of the California Psychological Inventory was the Socialization (So) scale. This scale was originally called the "delinquency scale" and is intended to measure antisocial behavior, the degree to which social norms are accepted and adhered to, along with intrapersonal controls (Groth-Marnat, 1990). Although descriptive statistics are provided for all of the reliability coefficients in the data set, inferential analyses were performed only on the internal consistency coefficients of the Socialization scale and the reliability coefficients for the five factors of the scale. The reliability coefficients from the Socialization scale and the coefficients for the five factors served as dependent variables in the study. The independent variables were various sample characteristics (e.g., gender of participants and age of participants in the samples).

Procedure

A list of articles relating to the California Psychological Inventory was generated using the American Psychological Association's PsycINFO database. The key words "California Psychological Inventory" and "CPI" were entered into the search and the generated list of articles was saved. The initial search yielded a list of 1200 such citations. The first step of the procedure consisted of identifying and eliminating articles that were published in Dissertation Abstracts International, as these articles are not commonly used in reliability generalization procedures (Caruso, 2000; Viswesvaran & Ones, 2000; Yin & Fan, 2000). Of the initial list, 271 articles were found to be published in Dissertation Abstracts International, and were thus eliminated, leaving a list of 929 articles for further examination. Each of the articles in this remaining list was located and examined individually. In this process the articles were separated into five categories:

1. articles that use the CPI, but do not mention or report the reliability of the scores or the scores of any of the CPI's subscales ($n = 710$; $710/929 = 76.4\%$),
2. articles that use the CPI or one of its subscales, mention the reliability of the scores, but do not report or compute the reliability of their sample scores ($n = 115$; $115/929 = 12.4\%$),

3. "false hits", articles that accidentally appear in the search, but after examination have no relevance to the CPI or any of its subscales ($n = 20$; $20/929 = 2.2\%$),

4. articles written in a foreign language ($n = 39$; $39/929 = 4.2\%$), and

5. articles that use the CPI or one of its subscales and report the computed reliability of one, some, or all of the subscales for their sample ($n = 45$; $45/929 = 4.8\%$).

Only the 45 articles in the last category were found to be usable in the present study, these articles are marked with asterisks in the References section.

CHAPTER 3

RESULTS

Descriptive Statistics

Descriptive statistics were used to characterize the internal consistency and test-retest reliabilities for each of the subscales of the CPI. A regression procedure was used to evaluate the predictors of reliability coefficients for the Socialization scale of the CPI.

As shown in Table 1, descriptive statistics showed that the internal consistency reliabilities for all scales of the CPI ranged from the highest for the Tolerance subscale ($M = .77$, $SD = .08$) to lowest for the Femininity subscale ($M = .51$, $SD = .17$). Descriptive statistics for the test-retest reliabilities ranged from the highest for the Socialization subscale ($M = .84$, $SD = .04$) to the lowest for the Achievement via Conformance subscale ($M = .64$, $SD = .06$). The means in this table were obtained by first transforming the reliability coefficients to Fisher's z' coefficients, then averaging the Fisher's z' coefficients, and then backtransforming those z' coefficients into the reported reliability coefficients. The transformation technique was used in accordance with Silver and Dunlap (1987) who found that transforming correlation coefficients to Fisher's z' before

performing calculations and then back transforming the results back into correlation coefficients provides a more accurate and less biased estimate for the population correlation, especially for correlations above .50, which is characteristic of most reliability coefficients.

Figure 1 shows a box-and-whisker plot of internal consistency coefficients of the individual subscales of the CPI. It is clear from this plot that most subscales lack the necessary number of coefficients to create an adequate picture of the distribution of the coefficients' variability. An acceptable cut-off for moderate reliability is .60 (Gronlund & Linn, 1990). Some of the scales in this figure do not reach that cut-off point (e.g, Self-acceptance, Achievement via Conformance). Figure 2 shows a box-and-whisker plot of the test-retest reliability coefficients of the individual subscales of the CPI. The number of test retest reliabilities for most scales was too low to provide a meaningful depiction of the variability of the coefficients and in a lot of cases, the number of coefficients was not sufficient to provide a plot (e.g., Dominance, Sociability, Tolerance subscales, etc.). However, in the case of test-retest coefficients, most scales that have enough reliability coefficients do exceed the acceptable moderate reliability cut-off of .60. Both figures show that the distributions of the subscales' coefficients are mostly asymmetric, some with extreme outliers and very low *ns*.

Socializations Subscale

As compared to other subscales, the distribution of the coefficients of this particular subscale is somewhat asymmetric, with most values and the median falling in the upper half of the distribution. There are two outliers and one extreme value, which comprise a somewhat large portion of the scores, considering $n = 16$ for this subscale (see Figure 1).

A standard multiple regression analysis was used to examine predictors for the Socialization subscale of the CPI. The α coefficients served as the dependent variable in the analysis with the sample gender, age, clinical vs. non-clinical sample status and scale language serving as independent variables. Hence, a total of four variables were entered into the regression equation. The analysis did not yield any significant results, $F(4, 15) = .518, p > .05, R^2 = .39$.

The internal consistency and test-retest reliabilities for the Socialization scale were further examined via a one-way ANOVA. They were found to be significantly different, with the test-retest reliability coefficients showing a significantly higher mean than that of the internal consistency coefficients, $F(1, 17) = 5.19, p < .03$.

A repeated measures ANOVA was used to examine the differences between the internal consistency coefficients and the test-retest coefficients of the Socialization subscale by sample gender, sample age, clinical status of sample, and scale language. No significant differences were found.

Analyses of the Individual Factors

Each factor was computed by averaging the reliability coefficients of the scales that are part of that particular factor according to Gough (1987). Table 2 provides the means and standard deviations of the internal consistency and test-retest reliability coefficients of the factors that were used in the analysis. Note that in the analyses of the test-retest coefficients, Factor 5 was omitted because no coefficients were available for that factor.

Figures 3 and 4 provide box-and-whisker plots of the internal consistency and test retest reliability coefficients by factor, respectively. It is evident from the box-and-whisker plots that the distributions of the internal consistency and test retest coefficients of the factors are more symmetrical than those of the individual subscales, due partly to the fact that the fact that the distributions of the factors' coefficients have much larger *ns*. The larger *ns* also contribute to each of the factors' exceeding the moderate reliability cut-off of .60, in the case of both internal consistency and test-retest coefficients.

Multiple regression analyses were performed to examine the significance of the predictors for each of the factors' internal consistency coefficients. Sample gender, scale language, age, and clinical vs. non-clinical sample status were used as predictor variables with the factor coefficients serving as the dependent variables. None of the predictors were found to be significant for any of the factors. The results of the multiple

regression analyses are as follows. Factor 1, $F(4, 8) = .565$, $p > .05$, $R^2 = .22$; Factor 2, $F(4, 25) = .940$, $p > .05$, $R^2 = .131$; Factor 3, $F(4, 14) = 1.09$, $p > .05$, $R^2 = .238$; Factor 4, $F(4, 15) = .676$, $p > .05$, $R^2 = .153$; Factor 5's internal consistency coefficients could not be examined using a multiple regression analysis because the number of coefficients was too small. A between groups ANOVA was used to examine any differences among the internal consistency coefficients of the five factors. No significant differences among the factors' coefficients were found, $F(4, 81) = 1.50$, $p > .05$.

Multiple regression analyses were likewise used to examine the predictors for the test-retest reliability coefficients of the five factors. The same set of predictors was used in this analysis as with the internal consistency coefficients. None of the predictors were found to be significant for any of the examined factors' coefficients. The results of the analyses were as follows. Factor 1 $F(4, 2) = 1.58$, $p > .05$, $R^2 = .760$; Factor 2 $F(4, 3) = .629$, $p > .05$, $R^2 = .456$; Factor 3 $F(2, 3) = 1.73$, $p > .05$, $R^2 = .776$; Factor 4 $F(3, 5) = .047$, $p > .05$, $R^2 = .066$. There were no test retest coefficients available for Factor 5, hence this factor was not included in the multiple regression analysis.

The difference between the internal consistency and the test retest coefficients of the five factors was examined an ANOVA. They were found to be significantly different, with test retest

reliabilities showing a significantly higher mean than the internal consistency coefficients, $F (1, 24) = 8.62, p < .01$.

A series of repeated measures ANOVAS were performed to examine the difference between the internal consistency coefficients and the test-retest coefficients by sample gender, sample age, clinical status of sample, and scale language for all five factors. No significant differences were found for any of the factors.

CHAPTER 4

DISCUSSION

It is surprising that the test-retest reliability coefficients scores used in this analysis are significantly higher for the subscales and the factors than the internal consistency coefficients. One explanation for this would be that a much smaller number of test-retest coefficients were reported than internal consistency coefficients and this, in turn, minimized the variance of the test-retest coefficients as opposed to the variance of the internal consistency coefficients. It is, of course, much easier to find a significant difference between a group of coefficients with significantly different variances than two groups where variances are relatively equal. Furthermore, very little is known about the validity or quality of the test retest coefficients that were used in this study because each of the authors' procedure for computing the test retest reliability is different (e.g., the length of time that has elapsed between test administrations) and this may affect the outcome of the coefficient calculation. These results have not been replicated in the literature. Barnes, Harp, and Jung (2002) have compared internal consistency and test-retest coefficients, but found the internal consistency coefficients to be higher than test-retest,

as expected. Another explanation for this finding is that scores with lower internal consistency may have greater validity because different items (or in this case subscales) may be measuring the same concept Achenbach (1991 c.f. Silver, Ross, Alvarez, Bensaheb, Karwoski, LaSota, Silverman, & Zaikina, 2004). This is probably the most plausible explanation for the present results, due to the fact that there is an overlap in the subscales and some subscales belong to more than one factor (see Table 2).

Overall, the current study yielded few significant results. This may have occurred for a number of reasons. One of the more salient of these was a paucity of data points available for analysis. This lack of data can be explained by a few different factors. Although the California Psychological Inventory is a widely used scale (recall that the original search yielded over twelve hundred results!) and has been used since its inception in 1957, it is not possible to compute a reliability coefficient for the entire instrument. Hence, each article that uses the CPI and reports a computed reliability coefficient for the sample used in that particular study does not provide us with a reliability estimate for the same scale. Instead, the reliability estimates provided are for different scales, largely depending on the subject of the article itself and the needs of the author or authors. If a scale which measures one concept, for which only one reliability estimate could be computed was used, then every time a computed reliability estimate was encountered, there would be a 100% chance of that reliability estimate being one computed

for the whole instrument. Because the California Psychological Inventory is composed of eighteen subscales, the chance of finding a reliability estimate for any one particular scale is reduced to five percent.

Another explanation for the small number of data points under analysis may be the general trend among researchers to omit reliability estimates for their samples. As mentioned earlier, it is not a common practice for authors to compute or report reliability estimated for their respective samples. Despite the author's attempt to overcome this situation by contacting researchers who could potentially provide reliability estimates for their study's samples and given an exceptional response rate, there were still not enough current studies in the literature to make up a sufficiently large number of data points to provide adequate statistical power.

A well-established practice among social scientists is the non-reporting of reliability of their sample's scores. This may be due to a variety of factors. One of these may be the necessity to save space in publication journals and therefore the requirement to delete information from the original manuscript. Of course, in publications where the reliability of the California Psychological Inventory is not the primary focus, this information is more likely to be omitted when submitting to a journal. Another factor may be that most university programs that prepare social science professionals do not emphasize the importance of score reliability in general and specifically of

reporting the reliability of the scores yielded by their particular samples (Sawilowsky, 2000). Moreover, authors tend to shy away from reporting non-significant or low reliability results and this fact may have also minimized the availability of data.

Another very notable limitation of the current study is the structure of the five factors of the CPI. Gough (1987) outlined the five factors (they were not given descriptive names, just simply named Factor 1, Factor 2, etc.) in such a way that some of the subscales were placed into one or more factors. For example, the Tolerance subscale is part of both factors 2 and 3 (see Table 2). This overlap of the subscales is problematic due to the fact that it creates a shared variance among the factors. When the difference between the reliability coefficients scores of the factors is examined, it becomes unclear whether each factor has its own unique variance. In effect, the analyses of the internal consistency coefficients of the individual factors of the CPI violates a basic assumption of independence.

The box-and-whisker plots show that as the number of reliability coefficients increases, the reliability crosses the moderate cut-off point of .60. For example, the internal consistency and test-retest reliability of the factors is much higher than the reliability and internal consistency of the individual subscales due to a much higher number of coefficients in each factor as compared to each scale. Nevertheless, it is problematic that the scale does not have clearly defined factors that measure various

separate aspects of social development, as the scale was originally meant to do.

As mentioned previously elsewhere in this paper, usually psychological instruments are not used with populations that they are not intended for and when reliability coefficients are reported, it is possible to make a definitive statement as to the reliability of the scores yielded by an instrument because the coefficients that are reported are produced by the appropriate populations. This is not the case with the CPI. The scores used in the present analyses were obtained with the intended populations (i.e., normal non-clinical adolescents) as well as with unintended populations (e.g., adults seeking psychiatric treatment). This means that although a picture of reliability for the scores of the intended populations cannot be seen, the reliability that is measured may be a more genuine example of the "true" reliability of the instrument's scores, since most psychological instruments are not used solely with the populations that they are intended for.

Reliability generalization studies, such as the present one offer an in-depth look at the reliability of the scores of psychological instruments, which may otherwise be assumed as being acceptable by their administrators and users. When one selects an instrument for the purposes of assessment or evaluation, it is difficult to estimate a true reliability of its scores other than the one provided in the instrument's manual, usually obtained with the intended population. A reliability

generalization study can be a useful tool for an estimate of the scores' reliability of a given instrument; scores obtained not only from the intended population, but ones that are obtained in the every-day application of the instrument. These studies can offer a look at not only the reliability of an instrument's scores, but also at the appropriate populations that the evaluation instrument can be used with.

REFERENCES

- *Adams, S.H. (1994). Role of hostility on women's health during midlife: A longitudinal study. *Health Psychology, 13*(6), 488-495
- *Adams, S.H., Cartwright, L.K., Stewart, A.J., Ostrove, J.M., Wink, P. (1998). Psychological predictors of good health in three longitudinal samples of educated midlife women. *Health Psychology, 17*(5), 412-420.
- *Ahmed, I. (1986). Initial psychometric evaluation of Urdu version of California Psychological Inventory (CPI). *Pakistan Journal of Psychological Research, 1*(1-2), 3-16.
- *Alterman, A.I., McDermott, P.A., Mulvaney, F.D., Cacciola, J.S., Rutherford, M.J., Searles, J.S., Lynch, K, Cnaan, A. (2003). Comparison of embedded and isolated administrations of the California Psychological Inventory's Socialization subscale. *Psychological Assessment, 15* (1), 64-70.
- *Antill, J.K. & Cunningham, J.D. (1982). Sex differences in performance on ability tests as a function of masculinity, femininity, and androgyny. *Journal of Personality and Social Psychology, 42*(4), 718-728.
- Arce-Ferrer, A.J. & Ketterer, J.J. (2003). The effect of scale tailoring for cross-cultural application on scale

- reliability and construct validity. *Educational and Psychological Measurement*, 63(3), 484-501.
- Barnes, L.L.B., Harp, D., Jung, W.S. (2002). Reliability generalization of scores on the Spielberg State-Trait Anxiety Inventory. *Educational and Psychological Measurement*, 62(4), 603-618.
- *Barnett, O.W. & Hamberger, K.L. (1992). The assessment of martially violent men on the California Psychological Inventory. *Violence and Victims*, 7(1), 15-28.
- Beretvas, S.N., Meyers, J.L, & Leite, W.L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement*, 62(4), 570-589.
- Caruso, J. (2000). Reliability generalization of the NEO Personality Scales. *Educational and Psychological Measurement*, 60 (2), 236-254.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement*, 20, 37-46.
- *Cooney, N.L., Kadden, R.M., Litt, M.D. (1990). A comparison of methods for assessing sociopathy in male and female alcoholics. *Journal of Studies on Alcohol*, 51(1), 42-48.
- Crocker, L. & Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York NY: Holt, Rineheart and Winston.
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

- *Day, D.V., Bedeian, A.G., Conte, J.M. (1998). Personality as a predictor of work-related outcomes: Test of a mediated latent structural model. *Journal of Applied Social Psychology*, 28(22), 2068-2088.
- *DeFrancesco, J.J. (1997). On the reliability of the revised Socialization scale. *Psychological Reports*, 81, 1002.
- *Donohue, M.V. & Lieberman, H. (1992). Social competence of male and female psychiatric patients: Sociability, social presence, socialization and age of onset of psychosis. *Occupational Therapy in Mental Health*, 12(1), 25-46.
- Fan, X. & Yin, P. (2003). Examinee characteristics and score reliability: An empirical investigation. *Educational and Psychological Measurement*, 63 (3), 357-368.
- *Feingold, A. (1981). Testing equity as an explanation for romantic couples "mismatched" on physical attractiveness. *Psychological Reports*, 49, 247-250.
- *Farley, F.H. & Cohen, A. (1980). Common items and reliability in personality measurement. *Journal of Research in Personality*, 14, 207-211.
- Fisher, R.A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1, 1-32.
- *Goldberg, L.R. & Kilkowski, J.M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology*, 48(1), 82-98.

- Gough, H. (1957). California Psychological Inventory Manual. Palo Alto, CA: Consulting Psychologists Press
- Gough, H. (1987). California Psychological Inventory: Administrators Guide. Palo Alto, CA: Consulting Psychologists Press
- *Gough, H. (1994). Theory, development, and interpretation of the CPI Socialization scale. *Psychological Reports*, 75, 651- 700.
- Gronlund, N.E. & Linn, R.L. (1990). Measurement and Evaluation in Teaching. 6th Edition. New York NY: Macmillan Publishing Company.
- *Groth-Marnat, G. (1990). Handbook of Psychological Assessment. 2nd Edition. New York NY: A Wiley- Interscience Publication.
- *Hare, R.D. (1984). Performance of psychopaths on cognitive tasks related to frontal lobe function. *Journal of Abnormal Psychology*, 93(2), 133-140.
- *Hare, R.D. (1985). Comparison of procedures for the assessment of psychopathy. *Journal of Consulting and Clinical Psychology*, 53(1), 7-16.
- *Heilbrun Jr., A.B. (1979). Psychopathy and violent crime. *Journal of Counseling and Clinical Psychology*, 47(3), 509-516.
- *Helson, R. & Roberts, B.W. (1994). Ego development and personality change in adulthood. *Journal of Personality and Social Psychology*, 66(5), 911-920.

- *Helson, R., Stewart, A.J., Ostrove, J. (1995). Identity in three cohorts of midlife women. *Journal of Personality and Social Psychology*, 69(3), 544-557.
- *Helson, R. & Wink, P. (1987). Two conceptions of maturity examined in the findings of a longitudinal study. *Journal of Personality and Social Psychology*, 53(3), 531-541.
- Henson, R.K. & Thompson, B. (2001, April). *Characterizing measurement error in test scores across studies: a tutorial on conducting "Reliability Generalization: analyses*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
- Hunter, J.E. & Schmidt, F.L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Newbury Park: Sage Publications.
- *Ittenbach, R.F. & Eller, B.F. (1988). A personality profile of Southeastern Conference football officials. *Journal of Sport Behavior* 11(3), 115-125.
- *Ittenbach, R.A. & Harrison, P.L. (1990). Predicting ego-strength from problem-solving ability of college students. *Measurement and Evaluation in Counseling and Development*, 23(3), 128-136.
- *Jannarone, R.J. & Roberst J.S. (1984). Reflecting interactions among personality items: Meehl's Paradox revisited. *Journal of Personality and Social Psychology*, 47(3), 621-628.

- *Johnson, J.A. (1997). Seven social performance scales for the California Psychological Inventory. *Human Performance*, 10(1), 1-30.
- *Kadden, R.M., Donovan, D., Litt, M.D., Cooney, N.L. (1996). Psychometric properties of the California Psychological Inventory Socialization Scale in treatment-seeking alcoholics. *Psychology of Addictive Behaviors*, 10(2), 131-146.
- Kennedy, R.S. & Turnage, J.J. (1991). Reliability generalization: A viable key for establishing validity generalization. *Perceptual and Motor Skills*, 72, 297-298.
- *Kosson, D.S., Kelly, J.C., White, J.W. (1997). Psychopathy-related traits predict self-reported sexual aggression among college men. *Journal of Interpersonal Violence*, 12(2), 241-254.
- *Lifton, P.D. (1983). Measures of autonomy. *Journal of Personality Assessment*, 47(5), 514-523.
- *Lilienfield, S.O., Andrews, B.P., Stone-Romero E.F., Stone, D. (1994). The relations between a self-report honesty test and personality measures in prison and college samples. *Journal of Research in Personality*, 28, 154-169.
- *Lindholm, B.W. & Touliatos, J. (1995). Personality traits and curricular choices of undergraduate women in home economics. *Psychological Reports*, 76, 579-586.
- *Loehlin, J.C. (1997). A test of J.R. Harris's theory of peer influences on personality. *Journal of Personality and Social Psychology*, 72(5), 1197-1201.

- *Morris, G.P. & Zoerner Jr., C.E. (1973). Interaction of anxiety and sex-role identification characteristics in American male sex-role behavior. *Psychological Reports*, 32, 1135- 1142.
- Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory*. 3rd Edition. New York NY: McGraw-Hill, Inc.
- *Park, J.H., Meyers, L.S., Czar, G.C. (1998). Religiosity and spirituality: An exploratory analysis using the CPI 3- vector model. *Journal of Social Behavior & Personality*, 13(3) 541-552.
- Pedhazur, E.J. & Schmelkin, L. (1991). *Measurement, Design, and Analysis: An Integrated Approach*. Hillsdale NJ: Lawrence Erlbaum Associates.
- *Peterson, P.L. (1979). Aptitude x treatment interaction effects of teacher structuring and student participation in college instruction. *Journal of Educational Psychology*, 71(4), 521-533.
- *Peterson, P.L., Janicki, T.C., Swing, S.R. (1980). Aptitude-treatment interaction effects of three social studies teaching approaches. *American Educational Research Journal*, 17(3), 339-360.
- *Sandal, G.M. & Endersen, I.M. (2002). The sensitivity of the CPI Good Impression scale for detecting "faking good" among Norwegian students and job applicants. *International Journal of Selection and Assessment*, 10(4), 304-311.
- Sawilowsky, S.S. (2000). Psychometrics versus datametrics: Comment on vacha-haase's "reliability generalization" method

and some EPM editorial policies. *Educational and Psychological Measurement*, 60 (2), 157-173.

*Schmeck, R.R. & Ribich, F.D. (1978). Construct validation of the Inventory of Learning Processes. *Applied Psychological Measurement*, 2(4), 551-562.

Schmidt, F.L., Pearlman, K., Hunter, J.E., & Hirsch, H.R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology*, 38(4), 697-798.

*Schmit, M.J. & Stanard, S.J. (1996). The utility of personality inventories in the employee assistance process: A study of EAP referred police officers. *Employee Assistance Quarterly*, 11(4), 21-42.

*Schuerger, J.M. Tait, E., Tavernelli, M. (1982). Temporal stability of personality by questionnaire. *Journal of Personality and Social Psychology*, 43(1), 176-182.

*Sevig, T.D., Highlen, P.S., Adams, E.M. (2000). Development and validation of the Self-Identity Inventory (SII): A multicultural identity development instrument. *Cultural Diversity and Ethnic Minority Psychology*, 6(2), 162-182.

*Sheppy, M.I., Friesen, J.D., Hakstian, A.R. (1988). Eco-systemic analysis of anorexia nervosa. *Journal of Adolescence*, 11, 373-391.

Shields, A.l. & Caruso, J.C. (2003). Reliability generalization of the Alcohol Use Disorders Identification Test. *Educational and Psychological Measurement*, 63 (3), 404-413

- Shields, A.L. & Caruso, J.C. (2004). Reliability induction and reliability generalization study of the CAGE. *Educational and Psychological Measurement*, 64 (2), 254-270.
- Silver, N.C. & Dunlap, W.P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology*, 72 (1), 146-148.
- Silver, N.C., Ross, S.A., Alvarez, K.M, Bensaheb, A., Karwoski, J.E., LaSota, M., Silverman, J. & Zaikina, H. (2004). Reliability generalization of the Child Behavior Checklist. Paper presented at the meeting of the Southeastern Psychological Association, Atlanta, GA.
- *Sinha, A.K. & Sinha, U.P. (1977). Some socialisational correlates of worldmindedness. *Asian Journal of Psychology & Education*, 2(3), 27-30.
- *Swan, G.E., Carmelli, D., Rosenman, R.H. (1986). Spouse-pair similarity on the California Psychological Inventory with reference to husband's coronary heart disease. *Psychosomatic Medicine*, 48(3-4), 172-186.
- Thompson, B (1994). Guidelines for authors. *Educational and Psychological Measurement*, 54, 834-847.
- Thompson, B. (2002). Score Reliability: Contemporary Thinking on Reliability Issues. Thousand Oaks: Sage Publications.
- Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, 58 (1), 6-20.

- Vacha-Haase, T., Henson, R.K., & Caruso, J.C. (2002). Reliability Generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, 62 (4), 562-569.
- Viswesvaran, C. & Ones, D.S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement*, 60 (2), 224-235.
- *Weiser, N.L. & Meyers, L.S. (1993). Validity and reliability of the revised California Psychological Inventory's vector 3 scale. *Educational and Psychological Measurement*, 53, 1045-1054.
- Wilkinson, L. & American Psychological Association (APA) Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54 (8), 594-604.
- Yang, J., McCrae, R.R., Costa, Jr., P.T. (1998). Adult age differences in personality traits in the United States and the People's Republic of China. *Journal of Gerontology: Psychological Sciences*, 53B(6), 375-383.
- Yin, P. & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, 60 (2), 201-223.

- *Ying, Y. (1990). Use of the CPI structural scales in Taiwan college graduates. *The International Journal of Social Psychiatry*, 36(1), 49-57.
- *Zebb, B.J. & Meyers, L.S. (1993). Reliability and validity of the revised California Psychological Inventory's vector 1 scale. *Educational and Psychological Measurement*, 53, 271-280.

Table 1 *Descriptive Statistics and Number of Occurrences of the Coefficients for Each Scale of the CPI*

<u>Scale</u>	<u>Internal Consistency (α)</u>				<u>Test-Retest</u>			
	<u>Mean</u>	<u>SD</u>	<u>n</u>	<u>Range (min-max)</u>	<u>Mean</u>	<u>SD</u>	<u>n</u>	<u>Range (min-max)</u>
Dominance	.73	.19	5	.52-.96	.77	n/a	1	n/a
Capacity for Status	.59	.01	3	.58-.59	.75	.04	2	.72-.77
Sociability	.56	.30	4	.13-.80	.79	.05	2	.76-.83
Social Presence	.57	.09	2	.50-.63	.67	n/a	1	n/a
Self-acceptance	.52	.05	3	.48-.57	.64	n/a	1	n/a
Well-being	.43	.06	2	.38-.47	.77	n/a	1	n/a
Responsibility	.58	.13	2	.48-.67	.72	.06	2	.67-.76
Socialization	.71	.10	16	.42-.88	.84	.04	3	.81-.89
Self-control	.68	.13	3	.55-.81	.78	n/a	1	n/a
Tolerance	.78	.09	5	.68-.87	.69	n/a	1	n/a
Good Impression	.71	.17	2	.59-.83	n/a	n/a	0	n/a
Communality	n/a	n/a	0	n/a	n/a	n/a	0	n/a
Achievement via Conformance	.54	.07	5	.48-.65	.65	.06	2	.60-.69
Achievement via Independence	.66	.15	6	.47-.88	.72	.21	2	.57-.86
Intellectual Efficiency	.64	.09	3	.54-.72	.71	.14	3	.55-.83
Flexibility	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Femininity	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Table 2 Means and Standard deviations of the internal consistency and test retest coefficients of the five factors of the CPI

<u>Factor (subscales)</u>	<u>Internal Consistency</u>			<u>Test-retest</u>		
	<u>Mean</u>	<u>SD</u>	<u>n</u>	<u>Mean</u>	<u>SD</u>	<u>n</u>
1: Dominance, Capacity for Status, Sociability, Social Presence, Self-acceptance, Independence, Empathy	.601	.19	12	.724	.065	6
2: Responsibility, Socialization, Self-control, Tolerance, Good Impression, Communality, Sense of Well Being	.645	.128	29	.759	.079	8
3: Achievement via Independence, Flexibility, Tolerance, Intellectual Efficiency, Psychological Mindedness	.646	.131	18	.708	.077	4
4: Communality, Responsibility, Socialization, Sense of Well-being	.665	.131	19	.776	.074	6
5: Femininity/Masculinity	.602	.132	4	.89	n/a	1

Figure 1. Box-and-whisker plots of internal consistency coefficients by scale.

Note. The two letter abbreviations refer to the names of the subscales, where: Do = Dominance; Cs = Capacity for Status; Sy = Sociability; Sp = Social Presence; Sa = Self-acceptance; Wb = Sense of Well-being; Re = Responsibility; So = Socialization; Sc = Self-control; To = Tolerance; Gi = Good Impression; Ac = Achievement via Conformance; Ai = Achievement via Independence; Ie = Intellectual Efficiency; Fx = Flexibility; Fe = Femininity/Masculinity; n/a = there were no coefficients available for this subscale. O = outlier, * = extreme value.

Figure 2. Box-and-whisker plots of test retest coefficients by scale.

Note. The two letter abbreviations refer to the names of the subscales, where: Do = Dominance; Cs = Capacity for Status; Sy = Sociability; Sp = Social Presence; Sa = Self-acceptance; Wb = Sense of Well-being; Re = Responsibility; So = Socialization; Sc = Self-control; To = Tolerance; Gi = Good Impression; Ac = Achievement via Conformance; Ai = Achievement via Independence; Ie = Intellectual Efficiency; Fx = Flexibility; Fe = Femininity/Masculinity; n/a = there were no coefficients available for this subscale.

Figure 3. Box-and-whisker plot of internal consistency coefficients by factor.

Note. F1 = Factor 1; F2 = Factor 2; F3 = Factor 3; F4 = Factor 4; F5 = Factor 5. O = outlier, * = extreme value.

Figure 4. Box-and-whisker plot of test retest coefficients by factor.

Note. F1 = Factor 1; F2 = Factor 2; F3 = Factor 3; F4 = Factor 4; No test retest coefficients were available for Factor 5.

VITA

Graduate College
University of Nevada, Las Vegas

Helen Zaikina-Montgomery

Home Address:

5141 Lindell Road, Unit 202
Las Vegas, NV 89118

Degrees:

Bachelor of Arts, 2001
California State University, San Marcos

Publications:

Wogalter, M.S., Silver, N.C., Leonard, S.D., & Zaikina, H.
Warning Symbols. In M.S. Wogalter (Ed.) Handbook of
Warnings. Mahwah, NJ: Lawrence Erlbaum.

Thesis Title: Reliability Generalization of the California
Psychological Inventory

Thesis Examination Committee:

Chairperson, Dr. N. Clayton Silver, Ph.D.
Committee Member, Dr. Murray G. Millar, Ph.D.
Committee Member, Dr. Karen A. Kemtes, Ph.D.
Graduate Faculty Representative, Dr. Alice J. Corkill,
Ph.D.