8-2008

Reliability Generalization of the Alcohol Use Disorders Identification Test.

Chandni Patel
East Tennessee State University

Reliability Generalization of the Alcohol Use Disorders Identification Test

---

A thesis

presented to

the faculty of the Department of Psychology

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Arts in Psychology

---

by

Chandni Patel

August 2008

Alan Shields, Ph.D., Chair

Jodi Polaha, Ph.D.

Jon Webb, Ph.D.

Keywords: AUDIT, Reliability Generalization, Reliability, Alcohol Use Disorders

ABSTRACT

Reliability Generalization of the Alcohol Use Disorders Identification Test

by

Chandni Patel

The Alcohol Use Disorder Identification Test (AUDIT) is a brief screening instrument for assessing alcohol use problems among adults. This instrument is widely used and continued evaluation of its psychometric performance is needed. Reliability and validity are the primary psychometric characteristics of interest when evaluating psychological instruments. The focus of the present study is on reliability, which reflects the consistency or repeatability of the scores produced by a given instrument. Using meta-analytic methods, results showed that approximately 65% of previously published studies using the AUDIT did not appropriately report reliability estimates. Among the remaining studies, weighted reliability estimate centered on .81 (SD = .07) suggesting that the AUDIT generally produces scores of adequate reliability for most research purposes. Multiple regression equations showed that, among a variety of sample and methodological characteristics, the standard deviation of scores was the only statistically significant predictor of the variability in AUDIT score reliability estimates.

CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The focus of this paper is on the reliability of scores produced by the Alcohol Use Disorders Identification Test (AUDIT) (Saunders, Aasland, Babor, De la Fuente, & Grant, 1993) across various measurement circumstances. Reliability generalization (RG) is a meta-analytic method used to evaluate the reliability of scores produced by an instrument across different studies (Thompson, 1999) and, therefore, provides a framework for evaluating reliability in the present study. A strength of the RG method is that it can evaluate score reliability estimates obtained from a population of studies as opposed to a single study. Because of this, it has become a popular data analytic approach when evaluating the psychometric properties of scores produced by psychological instruments (Vacha-Haase, Henson, & Caruso, 2002).

Using RG, Shields and Caruso (2003) evaluated the reliability of scores produced by the AUDIT. Their analyses, based on 24 studies, indicated that, relative to score reliability, the AUDIT is appropriate for most research purposes (e.g., .80 or higher) with a reported median internal consistency estimate of .81. Because practitioners tend to demand higher reliability estimates (e.g., .90 or higher), however, Shields and Caruso encouraged the cautionary use of the instrument when used in clinical situations and when making clinical decisions on individual patients. Though generally capable of generating reliable scores, considerable variability within AUDIT score reliability estimates was reported in this study (range = .59 to .91) and this indicates that at times the AUDIT can and will produce scores of inferior reliability. In order to better understand this variability in reliability estimates, the predictor variables of score variability, sample age, sample gender, and whether or not the sample was from a clinical

population were related to it. After controlling for score variability, no sample characteristic was a statistically significant predictor of score reliability.

Data analyzed in Shields and Caruso (2003) were based on studies published up to the year 2000. Since that publication, an additional 172 studies using the AUDIT have been published in the empirical literature. The aim of the present study is to replicate the method used in Shields and Caruso to update the AUDIT RG findings. Results will be of immediate use to both clinicians and researchers interested in using the AUDIT.

CHAPTER 2

ALCOHOL USE DISORDERS

Alcohol use disorders are among the most commonly occurring mental disorders in the United States and are associated with significant personal and societal costs (Goetzel, Hawkins, Ozminkowski, & Wang, 2003). Currently, alcohol use disorders rank as the fifth most costly mental health condition among one sample of employers in the United States (Goetzel et al.). Rice (1990) calculated the total economic cost of alcohol abuse and dependence as US$98.6 billion for 1990 and this figure nearly doubled in less than 10 years (Harwood, Fountain, & Livermore, 1998). Though preventable, it is clear that alcohol use disorders impose a staggering economic burden on society.

In addition to this economic cost, people with alcohol use disorders engage in behaviors that have adverse effects on themselves and those around them. For example, the risk of mortality is increased among adolescents and young adults who engage in problem drinking (e.g., from violence, driving under the influence, and suicide; Windle, 2003). Acute physiological problems (e.g., coronary heart disease, stroke, hypertension, liver cirrhosis, and certain cancers) can also arise as a result of these disorders (National Institute on Alcohol Abuse and Alcoholism [NIAAA], 2000).

Alcohol use disorders typically start in adolescence and the intensity of the disorder tends to peak between the ages of 20-30 (Tartar & Vanyukov, 1994). The course of alcohol use disorders is marked by periods of remission and relapse, with longer abstinence rates after treatment (Rohde, Lewinsohn, Kahler, Seeley, & Brown, 2001). However, research indicates that only a small percent of people with the disorder actually seek treatment and among those who do not seek treatment, the tendency to stop drinking is typically preceded by a crisis and followed

by weeks of abstinence before a relapse (Saunders, Zygowicz, & D'Angelo, 2006). Alcohol use disorders can be diagnosed at two levels: abuse and dependence (APA, 2000). Alcohol dependence is more severe than abuse and leads to cognitive, behavioral, and physiological symptoms; alcohol abuse affects one of these areas and, in most cases, is a precursor to dependence (APA).

Alcohol Abuse

According to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition Text Revision (DSM-IV-TR) (American Psychiatric Association, 2000) alcohol abuse is a pattern of problem drinking resulting in significant impairment or distress. There are four criteria for the disorder: failure to fulfill major obligations as a result of excessive drinking; recurrent use of alcohol in physically hazardous situations; legal problems arising from excessive alcohol use; and persistent interpersonal or social problems caused by continued alcohol use.

The DSM-IV-TR states that alcohol abuse is expected to develop in people who have recently started using alcohol, and that the abuse may progress into dependence. The idea that alcohol abuse is an early sign of alcohol dependence has been supported by Ridenour, Cottler, Compton, Spitznagel, and Cunningham-Williams (2002). However, this study used retrospective data and this may have influenced the results. Newcomb, Galaif, and Locke (2001) did not find support for the progression of alcohol abuse to dependence. Participants in the study who had abused alcohol were more likely to be in remission 4 years later than to be diagnosed with alcohol dependence. Some factors that predict poor prognosis of alcohol abuse include younger age at onset of drinking, heavy episodic drinking, family history of alcohol abuse, and being single, divorced, or separated (Grant, Stinson, & Harford, 2001).

Alcohol Dependence

According to the DSM-IV-TR, alcohol dependence is a longstanding pattern of problem drinking that leads to clinically significant impairment or distress. People diagnosed with alcohol dependence meet the criteria for alcohol abuse but must also meet three of the following criteria over a 12month period: tolerance (a need for increased amounts of alcohol to reach intoxication or achieve the desired effect); withdrawal (having physical symptoms as a result of not drinking alcohol for a certain period of time); consumption of larger amounts of alcohol over a longer period of time; persistent desire to cut down or control drinking habits; large amounts of time spent trying to obtain alcohol; and physical and psychological effects due to continued long-term use.

Alcohol Screening Measures

The goal of alcohol screening is to identify people with an alcohol use disorder or to identify people who are at high risk for developing an alcohol use disorder (Carey & Teitelbaum, 1996). In this way, these individuals can be referred directly to treatment or for additional assessment. Self-report is the most common method of assessment used for alcohol use problems, including alcohol screening (Maisto & Connors, 1990). Though alcohol screening can be accomplished via physiological tests and chemical biomarkers (e.g., blood alcohol content and carbohydrate deficient transferrin), studies have demonstrated that self-reports are often better indicators of alcohol use problems and, moreover, are more efficient and cost-effective tests (Alte, Luedemann, Rose, & John, 2004). The most commonly used self-report alcohol screening instruments include measures such as the AUDIT , the Michigan Alcoholism Screening Test (MAST) (Selzer, 1971), the CAGE (Mayfield, McLeod, & Hall, 1974), the Alcohol Use Inventory (AUI) (Horn, Warnberg, & Foster, 1990), and the Substance Abuse

Subtle Screening Inventory-3 (SASSI) (Miller, 1985). A detailed evaluation of all these measures is beyond the scope of this review; only the AUDIT will be discussed in detail.

The AUDIT (Babor, Higgins-Biddle, Saunders, & Monteiro, 1992) is a brief alcohol screening tool that requires approximately 2 minutes to administer and approximately 3 minutes to score (Knight, Sherritt, Harris, Gates, & Chang, 2003) Developed by the World Health Organization for use in primary care settings, the AUDIT was validated on primary health care patients in six countries (Australia, Bulgaria, Kenya, Mexico, Norway, and United States of America) as a screening tool for harmful drinking patterns in the general population (Babor et al.). More specifically, the primary objective of the AUDIT is to identify people with hazardous drinking behaviors (i.e., they have an increased risk of harm to themselves and others), harmful drinking behaviors (i.e., drinking behavior has adverse effects on them physically and mentally), and or alcohol use disorders (as specified in the *DSM-IV-TR*).

Item selection for the AUDIT was based on daily alcohol intake, frequency of consuming six or more drinks per drinking episode, and the ability to discriminate hazardous and harmful drinkers (Saunders et al., 1993). There are 10 questions in the AUDIT and each response receives a score of 0 to 4, resulting in a Total scale score ranging from 0 to 40 where higher scores are more indicative of alcohol use problems (Aertgerts et al., 2000). A score of eight is often used as a cut-off for high risk of an alcohol use disorder (Knight et al., 2003). In addition to the Total scale score, the AUDIT assesses alcohol use across three domains including how much and how often a person drinks (Consumption scale), how dependent he or she is on alcohol (Alcohol Dependence Symptoms scale), and how adverse the consequences are as a result of his or her drinking (Negative Consequences scale) (Aertgerts et al.; Reinert & Allen, 2002).

Since the development of the AUDIT several short versions, modifications for different populations, and translations into other languages have also been created. The instrument is reportedly a preferred measure for assessing use problems because it is brief, focuses on current drinking behaviors, and was derived from an international sample of primary care patients (Allen, Reinert, & Volk, 2001). The AUDIT is used internationally in clinical settings as well as for research purposes (Selin, 2003). Indeed, since its development the AUDIT has been reported in over 200 studies appearing in the empirical literature.

CHAPTER 3

PSYCHOMETRIC PROPERTIES OF SCORES PRODUCED BY AN INSTRUMENT

The psychometric properties of reliability and validity are usually evaluated before psychological instruments, including screening tools, can be endorsed for use and there are many statistical methods available to researchers to ensure that these qualities are satisfactory (e.g., Cook & Campbell, 1979; Grimm & Yarnold, 1995, 2000; Kazdin, 1992; Keppel, 1991). Validity can be briefly described as the extent to which scores from an instrument measure the intended construct, while reliability is the extent to which scores from a measurement are consistent or repeatable. While each can be considered uniquely from the other, they are highly related. For example, test scores can be reliable but not valid but they cannot be valid yet unreliable. That is, reliability is a necessary albeit insufficient prerequisite for validity. The remainder of the discussion focuses on the property of reliability; however, we again acknowledge that validity and reliability are closely related.

Reliability

Reliability plays an important role in the professional lives of psychologists, yet aspects of it may not be well understood (Li, Rosenthal, & Rubin, 1996). Indeed, this notion may contribute to the current trend of underreporting reliability indices in the alcohol screening literature (Miller, Shields, Campfield, Wallace, & Weiss, 2007; Shields & Caruso, 2003, 2004) and larger published literature (e.g., Kieffer, Reese, & Thompson, 2001; Meier & Davis, 1990; Vacha-Haase, Henson, & Caruso, 2002; Whittington, 1998). In order to explain the concept of reliability, the following three sections will define and interpret reliability, describe the reliability indices, and explain the importance of reliability.

Reliability Defined

Reliability is broadly defined as the consistent measurement of a particular construct under different measurement conditions. In other words, measurements are reliable if the scores obtained are not affected by random variables in the testing situation. This idea can be represented through Lord and Novick's (1968) classical measurement theory which posits that observed scores (the respondent's actual score obtained on an instrument) are made up of a true score and a certain amount of random error. The true score is a theoretical score and defined as the average score an examinee would obtain if she or he was assessed on a given test an infinite number of times. Measurement error is defined as the difference between a true score and an observed or actual score and can be a result of random or systematic error. Random error from various sources causes the true score to fluctuate. Sources of random error include testing procedures (e.g., time allotted to each examinee) and test setting (e.g., room temperature) as well as examinee characteristics (e.g. examinee's level of energy or mood). Systematic error, on the other hand, reflects consistent inaccuracies in the test setting or the measurement. An example of systematic measurement error is loud noises or distractions outside the testing room that affect all examinees. The true score and the random error together make up the observed or actual score. Thus,

Observed score = true score + error.

The reliability coefficient represents variability in the observed score that is due to true or actual individual differences in the observed scores (e.g., a reliability coefficient of .75 indicates that 75% of the variability in scores is from actual differences between the examinees, whereas 25% is from random error).

Score reliability estimates will change as a measure is administered to different samples, partially as a function of the score variance in a given group. Variance (σ^2) is a measure of dispersion that represents the difference between each observed score and the average of the observed scores. As mentioned above, the observed score consists of a true score and an error score. Therefore, observed variance consists of a true score variance and an error score variance.

Classical test theory interprets reliability in terms of the ratio of true score variance to observed score variance. The reliability coefficient will typically range from 0 to 1, where the value of 0 indicates an absence of reliability and the value of 1 indicates perfect reliability and no error variance in observed scores. Measures usually generate scores that have a varying degree of reliability and not scores that are completely unreliable or completely reliable. There are no strict guidelines regarding the acceptability or unacceptability of a particular reliability coefficient. Yet, current standards suggest a minimum score reliability of .70 for the early stage of measure development, .80 for basic research purposes, and .90 when important clinical decisions are being made (Nunnally & Bernstein, 1994).

Reliability Indices

The degree of error within the scores of an instrument is measured by a variety of reliability indices. These indices evaluate the unique sources of random error within the scores of a given construct. There are four common estimates of reliability reported in empirical studies: inter-rater reliability, test-retest reliability, parallel-forms reliability, and internal consistency reliability, and each measures a different type of error.

Inter-Rater Reliability

In some instances, two or more researchers are measuring the same construct. Each researcher records information or carries out measurements independently. Random error can

occur in this situation as each researcher may focus on difference aspects of the construct and inter-rater reliability estimates can quantify this random error. The researchers would make observations or measurements according to a pre-established protocol. A correlation coefficient is calculated with the following equation that quantifies the strength of the relationship between each observer's ratings,

Total Agreements / Total Agreements +Total Disagreements

The inter-rater reliability is high if the different raters make similar observations or measurements. Empirical studies generally use the kappa coefficient to estimate inter-rater reliability because it accounts for random chance agreement (i.e., the likelihood that the researchers agree by chance).

Test-Retest Reliability

If a construct is relatively stable (i.e., it does not tend to naturally vary over time), an instrument should produce consistent scores from one measurement occasion to the next. Test-retest reliability estimates evaluate the degree to which an instrument yields similar results from one testing occasion to the next. Constructs such as intelligence and personality are thought to remain stable over time and should produce high test-retest reliability estimates, even over long periods of time (e.g., a year). This reliability index accounts for random error that may arise from the different testing conditions and represents the error of interest in test-retest reliability methods.

Parallel-Forms Reliability

Parallel-forms reliability is estimated by administering a measurement tool to a sample and then on a separate occasion administering a different version of the same tool to the same sample. In other words, this estimate is the correlation of scores obtained from two separate

measures of the same construct that are thought to be equivalent. Random errors due to the characteristics of the measurements can be quantified with this reliability estimate.

Internal Consistency Reliability

Internal consistency reliability coefficients estimate the correlation among different items within the same scale. Therefore, a high internal consistency coefficient indicates that the items are measuring the same construct and a low correlation indicates error within the instrument (i.e., the items may not measure the same construct). Common measures of internal consistency reliability are the split-half method (Spearman, 1910), the Kuder-Richardson method (Kuder & Richardson, 1937), and Cronbach's coefficient alpha (Cronbach, 1951).

Importance of Reliability

Attenuation theory (Lord & Novick, 1968) suggests that the correlation between observed scores is lower than the correlation between corresponding true scores. This is so because the correlation between observed scores is attenuated by unreliability of scores obtained from the measurement. For example, if a particular alcohol screen measure generated score *X* and a measure of depression generated score *Y*, their observed correlation is attenuated or lower than their true score correlation. The difference between an observed correlation and a true score correlation is due to the less than perfect reliability of scores produced by the two measures. This unreliability of measures of alcohol use, *X*, and depression, *Y*, may lead researchers to incorrectly conclude that the two variables are unrelated (Type II error; failing to identify a significant relationship where one exists). Further, the unreliability of scores generated by *X* and *Y* may reduce the observed correlation such that even clinically significant relationships go undetected. Thus, undetected and unsystematic errors in measurement lead to unreliable test scores, and unreliable test scores can significantly interfere with a researcher's power to draw

accurate conclusions about the relationships among events, variables, and concepts. This fact has often led to the conclusion that 'reliability precedes validity.'

Reliability as a Property of Scores

Reliability is recognized as a property of scores because reliability coefficients change as the use of the measurement or testing condition changes (Feldt & Brennan, 1989; Strube, 2000; Thompson, 1994, 1999; Wilkinson, 1999). In other words, reliability is not a property of a test as is often assumed but rather a property of scores generated by a test, which, as noted, can be influenced by multiple and various methodological and subject characteristics. Thompson noted, however, that researchers often fail to recognize this important aspect of reliability. While not a novel concept, its importance is severely understated. It is important to the extent that scores from samples with different characteristics (such as differing diagnostic groups, ages, or gender representation) can and will differ with respect to reliability.

Despite this knowledge, both researchers and practitioners frequently refer to *tests* as being reliable (or unreliable) or that a particular test is more reliable than another. While this may appear to be a semantic issue, Thompson (1994) stresses that referring to a test as reliable or not is simply incorrect and can have deleterious effects on scholarly thinking. For example, researchers often claim that an instrument is reliable without providing any reliability coefficients generated from the current data set (cf. Vacha-Haase et al., 2002). The negative consequence of this is that researchers incorrectly reinforce that reliability is a property of instruments. Furthermore, researchers are unaware of what sample characteristics will alter reliability coefficients and it makes it difficult for other researchers to compare reliability coefficients obtained from their own data sets.

Reporting Reliability

As it becomes clear that reliability is a property of scores and not tests, guidelines on reporting reliability estimates are being created. Thompson (1994) has stated that authors interested in submitting articles to the Educational and Psychological Measurement journal must follow the principle of reliability as a property of scores. He has specifically stated that referring to a test as reliable will be considered unacceptable and that authors should report reliability coefficients of the scores from the given instrument.

More recently, the Task Force on Statistical Inference (TFSI), brought together by the Board of Scientific Affairs of the American Psychological Association (APA), created guidelines on reporting statistics. They state that reliability is not a property of tests but rather a property of scores; furthermore, authors should provide reliability coefficients of scores obtained during research even if the study is not psychometric in nature because these estimates are important for assessing effect size (Wilkinson & the APA/TFSI, 1999).

Even though the importance of reporting reliability estimates in research has been stressed, the empirical literature to date appears to lack these data. For example, Vacha-Haase et al., (2002) reviewed over 20 studies that independently evaluated score reliability reporting practices for various measures and found that reliability was underreported in these studies. Among all the measures under evaluation (n =28), less than 25% of the studies on specific measures reported reliability information based on current data and there was great variability in the status of reliability reporting between studies (range = 0%-71%). Meier and Davis (1990) have reported similar findings on reliability reporting in the *Journal of Counseling Psychology* (*JCP).* In 1967, only 5% of studies reported reliability measures. Although, this percentage increased to 23% in 1987, reporting practices are still low. Kieffer et al. (2001) reported that

among articles published in JCP from 1988 to 1997, 43.9% of studies adequately reported reliability, 15.9% cited previously reported reliability estimates, and 40.2% did not mention reliability at all. Reviews of other journals have yielded similar results (Thompson & Snyder, 1998; Whittington, 1998).

However, more recent reviews have indicated some improvement in reliability reporting practices. O'Rourke (2004) found that approximately 68% of studies on the Center for Epidemiologic Studies Depression Scale reported reliability coefficients for the current data set. Ross, Blackburn, and Forbes (2005) estimate nearly 50% of articles published on the Patterns of Adaptive Learning Survey Goal Orientation Scales contain reliability data from the current samples. Hellman, Fuqua, and Worley (2006) reported that approximately 90% of articles discussing the Survey of Perceived Organizational Support provided reliability estimates from current data. These later findings are perhaps indicative of improved reliability reporting practices in the empirical literature.

Reliability Generalization

The term Reliability Generalization (RG) was introduced by Vacha-Haase (1998) as a meta-analytic procedure and is an extension of the validity generalization technique proposed by Schmidt and Hunter (1977) and Hunter and Schmidt (1990). There are two main objectives of RG studies: 1) to identify the typical reliability coefficients of scores obtained from a given instrument's measurements; 2) to recognize the variability among reliability scores and characterize the sources of this variance that predict or explain variation across samples and administration protocols.

Like other meta-analytic procedures, published studies are used as the units of analysis in an RG and means, standard deviations, and other descriptive statistics are computed for the

reliability coefficients across studies. Within RG studies, reliability coefficients are the dependent variable of interest and features of the studies and samples that best predict variation in these reliability coefficients are investigated.

Vacha-Haase (1998) also provided the first example of an RG study through examining score reliability of the Bem Sex Role Inventory. In this evaluation, it was reported that only 13.8% of studies provided reliability estimates from the data at hand, while 18.9 % of studies induced reliability by report of previous estimates and 67.3% made no mention of reliability. Furthermore, this evaluation of studies demonstrated that the reliability of scores can be predicted by various sample and method characteristics (Vacha-Haase). Vacha-Haase, through this RG, demonstrated the importance of recognizing reliability as a property of scores rather than measures, as reliability coefficients tended to widely vary across the 87 samples.

Once Vacha-Haase (1998) established the usefulness of RG studies, Educational and Psychological Measurement (Thompson, 2000) dedicated a section to RG. In this special section three RG studies were highlighted: the reliability of the Beck Depression Inventory (Yin & Fan, 2000), the "Big Five Factors" of personality (Viswesvaran & Ones, 2000), and the NEO Personality Scales (Caruso, 2000). These studies further supported the usefulness of viewing reliability in terms of scores by recording significant variability in score reliability estimates, and they also stressed the increased need for reporting reliability estimates. Since that time, we have identified a total of 43 studies using RG methods to evaluate literally dozens of instruments. These studies are listed in Table 1.

Table 1

Reliability Generalization Studies in the Published Empirical Literature by Year of Publication

	Reference	Year	Instrument Assessed
1	Vaacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58,* 6-20.	1998	Bem Sex-Role Inventory
2	Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement, 60,* 236-254.	2000	NEO- Personality Scales
3	Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60,* 201-223.	2000	Beck Depression Inventory
4	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,* 224-235.	2000	Big Five Factors of Personality
5	Vacha-Haase, T., Kogan, L. R., Tani, C. R., & Woodall, R. A. (2001). Reliability generalization: Exploring variation of reliability coefficients of MMPI clinical scales scores. *Educational and Psychological Measurement, 61,* 45-59.	2001	Minnesota Multiphasic Personality Inventory
6	Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement, 61,* 404-420.	2001	Teacher Efficacy Scale
7	Capraro, M. M., Capraro, R. M., & Henson, R. K. (2001). Measurement error of scores on the Mathematics Anxiety Rating Scale across studies. *Educational and Psychological Measurement, 61,* 373-386.	2001	Mathematics Anxiety Rating Scale
8	Caruso, J. C., & Edwards, S. (2001). *Reliability generalization* of the Junior Eysenck Personality Questionnaire. *Personality and Individual Differences, 31,* 1733-184.	2001	Junior Eysenck Personality Quesionnaire
9	Caruso, J. C., Witkiewitz, K., Belcourt-Dittloff, A., & Gottlieb, J.D. (2001). Reliability of scores from the Eysenck Personality Questionnaire: A reliability generalization study. *Educational and Psychological Measurement, 61,* 675-689.	2001	Eysenck Personality Questionnaire
10	Vacha-Haase, T., Tani, C. R., Kogan, L. R., Woodall, R. A., & Thompson, B. (2001). Reliability generalization: Exploring reliability variations on MMPI/MMPI-2 validity scale scores. *Assessment, 8,* 391-401.	2001	Minnesota Multiphasic Personality Inventory
11	Thompson, B., & Cook, C. (2002). Stability of the reliability of LibQUAL+-super(TM) scores: A reliability generalization meta-analysis study. *Educational and Psychological Measurement, 62,* 735-743.	2002	LibQUAL
12	Deditius-Island, H. K., & Caruso, J. C. (2002). An examination of the reliability of scores from Zuckerman's	2002	Sensation Seeking Scales, Form V

Sensation Seeking Scales, Form V. *Educational and Psychological Measurement, 62,* 728-734.

Table 1 (continued)

13	Henson, R. K., & Hwang, D. Y. (2002). Variability and prediction of measurement error in a Kolb's Learning Style Inventory Scores: A reliability generalization study. *Educational and Psychological Measurement, 62,* 712-727.	2002	Kolb's Learning Style Inventory
14	Lane, G. G., White, A. E., & Henson, R. K. (2002). Expanding reliability generalization methods with KR-21 estimates: An RG study of the Coopersmith Self-Esteem Inventory. *Educational and Psychological Measurement, 62,* 685-711.	2002	Coopersmith Self-Esteem Inventory
15	Wallace, K. A., & Wheeler, A. J. (2002). Reliability generalization of the Life Satisfaction Index. *Educational and Psychological Measurement, 62,* 674-684.	2002	Life Satisfaction Index
16	Hanson, W. E., Curry, K. T., & Bandalos, D. L. (2002). Reliability generalization of Working Alliance Inventory scale scores. *Educational and Psychological Measurement, 62,* 659-673.	2002	Working Alliance Inventory
17	Nilsson, J. E., Schmidt, C. K., & Meek, W. D. (2002). Reliability generalization: An examination of the Career Decision-Making Self-Efficacy Scale. *Educational and Psychological Measurement, 62,* 647-658.	2002	Career Decision-Making Self-Efficacy Scale
18	Reese, R. J., Kieffer, K. M., & Briggs, B. K. (2002). A reliability generalization study of select measures of adult attachment style. *Educational and Psychological Measurement, 62,* 619-646.	2002	Adult Attachment Style
19	Barnes, L. L. B., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement, 62,* 603-618.	2002	Spielberger State-Trait Anxiety Inventory
20	Capraro, R., & Capraro, M. M. (2002). Myers-Briggs Type Indicator score reliability across studies: A meta-analytic reliability generalization study. *Educational and Psychological Measurement, 62,* 590-602.	2002	Myers-Briggs Type Indicator
21	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,* 570-589.	2002	Marlowe-Crowne Social Desirability Scale
22	Kieffer, K. M., & Reese, R. J. (2002). A reliability generalization study of the Geriatric Depression Scale. *Educational and Psychological Measurement, 62,* 969-994.	2002	Geriatric Depression Scale
23	Shields, A. L. (2003). Reliability generalizations of three alcohol screening measures: The Alcohol use Disorders Identification Test, the CAGE Questionnaire, and the Michigan Alcoholism Screening Test. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 64,* 2406.	2003	Alcohol use Disorders Identification Test CAGE Questionnaire Michigan Alcoholism Screening Test
24	Campbell, J. S. (2003). Meta-analysis and reliability generalization of the psychopathy checklist applied to youth. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 64,* 1544.	2003	Psychopathy Checklist
25	Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement, 63,* 75-95.	2003	Beck Depression Inventory

26	Youngstrom, E. A. & Green, K. W. (2003). Reliability generalization of self-report of emotions when using the	2003	Differential Emotions Scale

Table 1 (continued)

	Differential Emotions Scale. *Educational and Psychological Measurement, 63,* 279-295.		
27	Shields, A. L., & Caruso, J. C. (2003). Reliability generalization of the Alcohol Use Disorders Identification Test. *Educational and Psychological Measurement, 63,* 404-413.	2003	Alcohol Use Disorders Identification Test
28	Shields, A. L., & Caruso, J. C. (2004). A Reliability Induction and Reliability Generalization Study of the CAGE Questionnaire. *Educational and Psychological Measurement, 64,* 254-270.	2004	CAGE
29	Kieffer, K. M., Cronin, C., & Fister, M. C. (2004). Exploring Variability and Sources of Measurement Error in Alcohol Expectancy Questionnaire Reliability Coefficients: A Meta-Analytic Reliability Generalization Study. *Journal of Studies on Alcohol, 65,* 663-671.	2004	Alcohol Expectancy Questionnaire
30	O'Rourke, N. (2004). Reliability Generalization of Responses by Care Providers to the Center for Epidemiologic Studies-Depression Scale. *Educational and Psychological Measurement, 64,* 973-990.	2004	CES-Depression Scale
31	Rexrode, K. R. (2005). The Ways of Coping Questionnaire: A reliability generalization study. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 66,* 3424.	2005	Ways of Coping Questionnaire
32	Zangaro, G. A., & Soeken, K. L. (2005). Meta-analysis of the reliability and validity of Part B of the Index of Work Satisfaction across studies. *Journal of Nursing Measurement, 13,* 7-22.	2005	Part B of the Index of Work Satisfaction
33	Ryngala, D. J., Shields, A. L., & Caruso, J. C. (2005). Reliability Generalization of the Revised Children's Manifest Anxiety Scale. *Educational and Psychological Measurement, 65,* 259-271.	2005	Revised Children's Manifest Anxiety Scale.
34	Mji, A., & Alkhateeb, H. M. (2005). Combining Reliability Coefficients: Toward Reliability Generalization of the Conceptions of Mathematics Questionnaire. *Psychological Reports, 96,* 627-634.	2005	Conceptions of Mathematics Questionnaire
35	Ross, M. E., Blackburn, M., & Forbes, S. (2005). Reliability Generalization of the Patterns of Adaptive Learning Survey Goal Orientation Scales. *Educational and Psychological Measurement, 65,* 451-464.	2005	Patterns of Adaptive Learning Survey Goal Orientation Scales
36	Campbell, J. S., Pulos, S., Hogan, M., & Murry, F. (2005). Reliability Generalization of the Psychopathy Checklist Applied in Youthful Samples. *Educational and Psychological Measurement, 65,* 639-656.	2005	Psychopathy Checklist
37	Leach, L. F., Henson, R. K. Odom, L. R., & Cagle, L. S. (2006). A Reliability Generalization Study of the Self-Description Questionnaire. *Educational and Psychological Measurement, 66,* 285-304.	2006	Self-Description Questionnaire
38	Hellman, C. M., Fuqua, D. R., & Worley, J. (2006). A Reliability Generalization Study on the Survey of Perceived Organizational Support: The Effects of Mean Age and Number of Items on Score Reliability. *Educational and Psychological Measurement, 66,* 631-642.	2006	Survey of Perceived Organizational Support

39	Graham, J. M., Liu, Y. J., & Jeziorski, J. L. (2006). The Dyadic Adjustment Scale: A Reliability Generalization Meta-Analysis.	2006	Dyadic Adjustment Scale

Table 1 (continued)

40	Rouse, S. V. (2007). Using reliability generalization methods to explore measurement error: An illustration using the MMPI-2 PSY-5 Scales. *Journal of Personality Assessment, 88,* 264-275.	2007	MMPI-2 PSY-5 Scales
41	Li, A., & Bagger, J. (2007). The balanced inventory of desirable responding (BIDR): A reliability generalization study. *Educational and Psychological Measurement, 67,* 525-544.	2007	Balanced Inventory of Desirable Responding
42	Shields, A. L., Howell, R. T., Potter, J. S., & Weiss, R. D. (2007). The Michigan Alcoholism Screening Test and its shortened forms: A meta-analytic inquiry into score reliability. *Substance Use and Misuse, 42,* 1783-1800.	2007	The Michigan Alcoholism Screening Test and its shortened forms
43	Miller, C. S., Shields, A. L., Campfield, D., Wallace, K. A., Weiss, R. D. (2007). Substance use scales of the Minnesota Multiphasic Personality Inventory: An exploration of score reliability via meta-analyses. *Educational and Psychological Measurement, 67,* 1052-1065.	2007	MAC/MAC-R

CHAPTER 4

THE PRESENT STUDY

As discussed above, it is clear that research on the reliability estimates of scores produced by alcohol screening instruments is generally lacking. In part, this may be attributed to the lack of reliability findings appropriately reported in the empirical literature. The AUDIT is an increasingly common tool and used with many different samples and this makes it important to establish the reliability of the scores taken from these diverse samples. The primary aim of this project is to improve the psychometric understanding of the AUDIT, and the results can be used to give healthcare providers an enhanced empirical foundation to refer to when using these measures. This was accomplished by applying the meta-analytic technique of RG to the AUDIT. The primary objectives of this study were to: (1) describe the typical reliability of scores from the AUDIT and (2) explore the methodological and sample characteristics associated with AUDIT score reliability. We also characterize the reliability reporting practices among these published studies using the AUDIT.

CHAPTER 5

METHOD

Data Collection

A literature search was conducted among published studies found in the American Psychological Association's PsychINFO and the National Library of Medicine's PubMed databases. These searches were not limited to a specific year, but rather spanned from earliest to latest available; however, the search was limited to articles published in peer-reviewed journals in English. The key words searched for included *alcohol use disorder identification test* and *AUDIT*. A manual bibliography search was also conducted among all the obtained articles. Articles included in this review span the years 1991 through February 2008.

Criteria for Including Studies in the Literature Review

Only studies using the AUDIT, or one of its revised versions, were included in the analysis. Studies that report reliability coefficients from their present sample were considered for inclusion in the RG.

Criteria for Excluding Studies in the RG Data Analysis

Studies that did not report AUDIT score reliability data based on the data at hand were excluded from the RG analysis. Additionally, studies found in more than one database (e.g., PUBMED and PsycINFO) were included in these analysis only once.

Coding Studies

Data coding was based on a scale from one to five. A study is coded as a one if the abstract mentions the AUDIT but the study cannot be located (e.g., it is a rare publication). When a study is coded as a two, it indicates that the study is not an empirical article (e.g., review article, meta-analysis, or commentary). Studies that are coded as three do not contain reliability

information of any kind (e.g., reliability induction by omission). Similarly, a study is coded as a four if the study mentions reliability but provides no direct empirical support (e.g., the article mentions that the "AUDIT is reliable" but provides no reliability estimates to support this claim) or if the study mentions reliability and provides a reliability estimate from previous research (e.g., reliability induction by report). Lastly, a study is scored as a five when the study reports an AUDIT reliability coefficient that is based on the data obtained in that study. It is the studies coded as a five that are analyzed in the RG study.

Criterion Variable

The criterion variable or dependent variable of interest in reliability generalization studies is typically a numeric index of reliability. Internal consistency reliability estimates are commonly reported in studies; therefore, this study used these estimates as the criterion variable. Coefficients of internal consistency were recorded as reported in the obtained empirical study (e.g., .85).

Predictor Variables

This study also examined methodological and sample characteristics possibly influencing the reliability of scores produced by the AUDIT. Although there is a wide range of potential factors that could contribute to the variance found within score reliability estimates, the number and type of predictor variables that can be coded are limited by their availability within the literature. Consistent with previous RGs of alcohol screening tools, the following predictor variables were investigated in the present study and more fully described below: score variability, sample age, sample gender composition, ethnicity of the sample, and sample type.

Score variability can demonstrate tremendous predictive power as regards score reliability (e.g., Caruso, Witkiewitz, Belcourt-Dittloff, & Gottlieb, 2001). This is consistent with

classical test theory (e.g., Lord & Novick, 1968) which can show that if error variance remains constant and observed score variance increases, then true score variance must increase to exactly the same extent (cf., Henson, 2001). Therefore, the ratio of true score variance to total variance, the reliability, also must increase. Without its inclusion as a predictor variable, variability in score reliability that is really due to increases in observed score variability may be erroneously attributed to other predictor variables. For this reason, the standard deviation of measure scores are included as a predictor variable.

Research has demonstrated that alcohol use disorders are prevalent and an important source of social and health care related problems and mortality in both the elderly (Adams, Magruder-Habib, Trued, & Broome, 1992; Adams, Yuan, Barboriak, & Rimm, 1993; Smith-Black, Rabins, & McGuire, 1998) and children (Reinherz, Giaconia, Rose, Lefkowitz, & Pakiz, 1993). Brief alcohol screening measures have been used in studies of older adults (e.g., Joseph, Rasmussen, Ganzini, & Atkinson, 1997) and are becoming more widely used in studies of younger populations (e.g., Chung et al., 2000; Thom, Herring, & Judd, 1999) because most adolescent specific alcohol screening measures are long and relatively cumbersome. Given the wide age-range in populations that have been administered the AUDIT, it is important to determine if the instrument can generate reliable scores among different age groups. To this end, average sample age will be used as a predictor variable within this RG.

A review of the literature pointed to the paucity of gender-specific analyses found in alcohol screening validation studies (Bradley, Boyd-Wickizer, Powell, & Burman, 1998). A meta-analysis performed in the same report found that alcohol screening measures can perform differently among women and men (Bradley et al.). Therefore, gender is included as a predictor variable and coded as proportion of the sample that is female.

A gap exists in our knowledge of the performance of alcohol screening measures among minority populations (Steinbauer, Cantor, Holzer, & Volk, 1998). Researchers have made great strides in addressing this problem (e.g., Cherpitel, 1998; Cherpitel & Clark, 1995; Volk, Steinbauer, Cantor, & Holzer, 1997), but little is known about the overall reliability of scores on alcohol screening measures among these populations. Therefore, ethnicity will be used as predictor variable to begin drawing conclusions about the usefulness of the AUDIT across ethnic minority groups. This variable will be coded as proportion of the sample that is identified as non-white.

Alcohol use disorders and alcohol use are common problems in a wide variety of psychiatric and health care environments (Galanter, Castaneda, & Ferman, 1988; Moore et al., 1989). These settings can provide vital opportunities for health care professionals to intervene with alcohol problems at potentially early stages of development. However, the empirical literature suggests that alcohol screening is not routinely applied and alcohol use disorders go under-identified (Duszynski, Nicto, & Valente, 1995; Wolford et al., 1999). One possible explanation for this lack of use is that health care professionals are not aware of which measures to use with certain populations. The AUDIT was designed for use among primary care patients though is used among clinical and non-clinical samples alike. In order to better understand how the AUDIT performs among samples of different populations, sample type will also be used as a predictor variable. It will be dummy-coded as clinical (1) and non-clinical (2) sample type. Clinical sample type represents any sample that was assessed in a mental or medical health treatment location (e.g., primary care, emergency room, psychiatric facility, substance use treatment center).

Data Analysis

Characterizing the typical reliability of AUDIT scores can be accomplished via traditional descriptive statistics. We reported the mean, median, standard deviation, and range of internal consistency reliability estimates obtained from the empirical literature. Data analytic methods within the general linear model are often employed to accomplish the second goal of RG, to investigate associations between study and sample characteristics and score reliability. In this way, variation in reliability estimates can be systematically explored for their controlling factors. Specifically, simple correlations and multiple regression and correlation equations were used to evaluate the relationship between the predictor variables and AUDIT score reliability. Because score variability is directly related to score reliability under certain basic assumptions, a hierarchical analysis with score variability entered in Model 1, and the other predictor variables entered as a block in Model 2 was employed within the multiple regression approach.

CHAPTER 6

RESULTS

The literature search resulted in 295 studies reporting use of the AUDIT. This suggests that the AUDIT is a widely used AUD screening instrument. Fourteen articles (4.8%) were review papers, letters to the editor, or nonempirical reports. The majority of articles (n = 194, 65.7%) did not report reliability at all. Twenty-eight (9.5%) articles reported reliability from a past study. Fifty-nine (20%) studies reported reliability based on the sample used in the report.

Some studies reported more than one reliability coefficient. For example, they reported reliability coefficients for different age groups, modes of administration, and males and females. This resulted in 94 samples available for further analysis. Some studies reported reliability using a different version of the AUDIT (e.g. AUDIT-C) and some significantly altered the wording of the instrument– these were not included in the sample for analysis, and the sample size was reduced to 68 studies. Another cut was made for studies that reported a reliability coefficient other than internal consistency (the focus of this study) and this resulted in 57 samples that were used in the present RG analyses.

Table 2 provides descriptive statistics for all the variables examined, of particular interest are the statistics for score reliability. The unweighted mean reliability estimate for scores from the AUDIT was .82 (SD = .07) and the weighted mean reliability estimate was .81 (SD = .05) indicating that the AUDIT generally produces scores with adequate reliability. When compared to the conventional cut-off values for reliability estimates (Nunnally & Bernstein, 1994), the mean reliability estimates here meet the cut-off for research purposes (.80) but do not meet the cut-off for clinical purposes (.90). The median is slightly higher than the mean due to the modest

negative skew of the distribution of reliability coefficients. The range (.59 to .97) indicates that there is a substantial degree of variability in the reliability of AUDIT scores.

The factors that may predict variability in reliability coefficients were first analyzed using a bivariate correlation. These results are also presented in Table 2. The relationship between the variability of scores obtained on the AUDIT (standard deviation) and score reliability was statistically significant, (r (43) = .688, p = .000). This suggests that as score variability increases score reliability also increases. According to classical test theory, if observed score variance increases the true score variance should also increase by the same amount. As a result the ratio of true score variance to total score variance (reliability) must also increase. The relationship between the type of sample (clinical or not clinical) and score reliability was also statistically significant, (r (57) = .447, p = .000). Score reliability increases when the AUDIT is used in clinical samples. This supports the intended use of the AUDIT which was as an AUD screening tool for use in primary care settings.

Table 2

Descriptive Statistics for Reliability Coefficients and Predictor Variables (total N of 57 samples yielded 44,356 respondents)

Variable	*n*	*M*	*Mdn*	*SD*	Range	r	*p*
Reliability coefficients							
Unweighted score reliability (α)	57	.819	.830	.07	.59 - .97		
Weighted score reliability (α)[a]	57	.805	.810	.05	.59 - .97		
Predictor variables							
Average Scores	52	7.41	6.32	4.10	2.00 – 23.4	.259	.064
SD of Scores	43	5.73	5.80	2.35	.30 – 12.32	.688	.000
Average age	43	33.76	36.4	9.93	16.4 – 53.4	.122	.435
Proportion non-white	26	.35	.20	.34	.02 – 1.00	.345	.085
Proportion female	55	.40	.44	.25	0 – 1.00	-.121	.378
Total sample size	57	778.18	287	1492.82	33 – 8686	-.102	.452
Sample type[b]	57	.56	-	.50	0 – 1.00	.447	.000

r = PCC between unweighted score reliability and identified predictor variables.
[a]Weighted by sample size.
[b]Sample type was coded as clinical (1) and non-clinical (0).

The factors that may predict variability in reliability coefficients were then analyzed using multiple regression. Table 3 presents these results. Reliability coefficients were regressed onto the predictor variables using a hierarchical regression. Score variance was entered into the model first. Mean scores, mean age, proportion non-white, proportion of females, and whether or

not the sample was clinical were entered as a block into the second model. In model 1, as expected, score variability predicted a statistically significant and large amount of this variance in score reliability: R^2 = .22 (F (1, 19) = 5.34, p < .05), Adjusted R^2 = .178. When the other predictor variables were entered in Model 2, none of them was statistically significant: R^2 = .382 (R^2 change = .163 (F change (5, 14) = .738, p = .607, Adjusted R^2 = .118)). Results from model 2 suggest that, taken together, factors such as age, ethnicity, and gender do not appreciably affect the reliability of scores from the AUDIT and that the AUDIT may be used with a sample of men and women with different ethnicities who are from different age groups. This is also consistent with the original design of the instrument.

Table 3

Summary Results from Hierarchical Regression Analyses

	B	β	SE	*t*	*p*
Model 1[a]					
SD of scores	.011	.469	.005	2.312	.032
Model 2[b]					
SD of scores	.010	.433	.010	1.004	.333
Average scores	-.006	-.471	.005	-1.193	.253
Average age	.001	.201	.002	.639	.533
Proportion nonwhite	.022	.141	.039	.569	.579
Proportion female	-.099	-.320	.127	-.778	.450
Sample type[c]	-.006	-.053	.035	-.183	.858

[a]Model I variance explained in reliability coefficients: R^2 = .22; F (1, 19) = 5.34, p = .032.
[b]Model II variance explained in reliability coefficients: R^2 = .38; F (6, 14) = 1.44, p = .266.
[c]Sample type was coded as clinical (1) and non-clinical (0).

CHAPTER 7

DISCUSSION

The use of the AUDIT has been increasing as can be seen from the number of studies found in the present literature search. Indeed, the number of published studies using the AUDIT has more than doubled since the previous RG on the instrument (Shields & Caruso, 2003). The AUDIT is a widely used alcohol screening instrument and the descriptive statistics from this RG indicate that scores on the AUDIT are generally reliable. The values of .82 and .83 for the unweighted mean and median score reliability indicate adequate proportions of true score variance for research purposes. However, the variability of internal consistency estimates (range = .59 to .97) shows that the AUDIT will not always produce reliable scores. The reliability coefficients found here are similar to those found earlier in Shields and Caruso (.79 and .81 respectively). Because of this, conclusions drawn from the previous study can be supported. More specifically, in situations where important clinical decisions or diagnoses have to be made the AUDIT should be used with caution and preferably in addition to other instruments. However, researchers using the AUDIT among groups in a research setting can be confident in its use among a variety of samples.

In general, the descriptive statistics on the predictor variables in the present study are comparable to those from Shields and Caruso (2003). Score variability across the two studies is similar (5.73 and 5.3), with a broader range found in the present study (12.02 versus 9.9). This broader range of scores may also impact the score reliability and may lead to the increase in reliability coefficients found in the present study compared to the previous study. The average age of participants in the present analysis is slightly lower in the current study (33.7 versus 36.1) but the range of average age is similar across the two studies (16.4 to 53.4 versus 16.5 to 53.0).

Sample characteristics of gender (40% women versus 38% women) and sample type (clinical or nonclinical; 56% clinical versus 58% clinical) are also similar across the two studies. This indicates that the AUDIT, despite being designed for use among adult clinical samples is being administered in non-clinical settings and among adolescents as well.

As expected, the results of the regression analyses replicated what was found in Shields and Caruso (2003) and clearly indicated that score variability was most predictive of score reliability. As discussed, classical test theory (Lord & Novick, 1968) does explain this relationship between score variability and variance in score reliability; if observed score variance increases then true score variance must increase by to the same extent. Reliability is the ratio of true score variance to total score variance and, therefore, must also increase.

In contrast to AUDIT score variability, the other predictor variables evaluated in the present study did not significantly predict variability in AUDIT score reliability. While this suggests that these variables do not impact AUDIT score reliability, it is possible that the variability in score reliability can be predicted by these variables indirectly. For example, it may be the case that women produce less variable AUDIT mean scores than do men. In this way, score variability moderates the impact that sex has on AUDIT score reliability. These potential moderating effects reflect one interesting area of future research. Nevertheless, the fact that no other predictor variable examined here accounted for variance in score reliability above and beyond that accounted for by score variability alone can tentatively be considered a positive characteristic of the AUDIT. This pattern of results also replicates what was found in Shields and Caruso (2003) and suggests that the AUDIT apparently produces equally reliable scores across different age groups, genders, ethnicities, and sample types, once score variability is accounted for. In addition to the predictor variables analyzed in Shields and Caruso, the present study

included ethnicity in the regression analyses. Ethnicity is not a statistically significant predictor of reliability of scores obtained from the AUDIT and, as mentioned above, is indicative of the use of the AUDIT in a diverse sample.

Limitations

Whenever a meta-analytic study is conducted the results must be considered in a light of potential bias. A "file-drawer" problem (Rosenthal, 1979) exists because studies of a higher quality are more likely to be written, submitted, and accepted for publication compared to studies of a poorer quality. File-drawer studies are a problem because the psychometric properties of unpublished studies may differ from the psychometric properties of published studies. Therefore, it is possible that the reliability coefficients examined here are slightly higher than those that would have been found if all studies using the AUDIT could be obtained. This conclusion is drawn under the assumption that non-published studies contain AUDIT scores of lower reliability than those of published reports.

Fifty-nine (20%) of the 295 studies using the AUDIT reported a reliability estimate based on the data gathered for that particular study. This shows a 4% increase in psychometric reporting practices since Shields and Caruso (2003), in which 16.3% of studies using the AUDIT reported reliability estimates based on the data gathered for that particular study. This reflects an overall improvement in the reliability reporting practices in the AUDIT literature. However, even with this increase in reliability reporting over the last few years, it is still significant to note that 80% of the studies did not appropriately report reliability estimates and this may limit the generalizability of these findings. If the missing reliability estimates were not reported because of low values, the present results might be an overestimate of AUDIT score reliability across samples. However, because the studies were published, we can speculate that the scores were of

adequate reliability, and the present results may provide a good estimate of the average reliability across samples.

Specific to RG, the only information that was needed for a study to be included in the analyses was reliability of scores, variability in scores, and sample characteristics such as average age, gender, ethnicity, and whether the sample consisted of a patient group. The fact that 80% of the studies using the AUDIT did not report this information indicates that researchers can improve their reporting of psychometric information. Despite explicit recommendations to report these data (Thompson, 1994; Wilkinson & APA/TFSI, 1999), published reports still frequently do not contain psychometric estimates of instruments or other relevant descriptive data. Nonetheless, as recommended by Wilkinson and the APA/TFSI "authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric" (p. 596).

CHAPTER 8

CONCLUSION

The results obtained here should not be interpreted as *the* reliability of scores obtained from the AUDIT. Instead, with regards to AUDIT score reliability, the findings support the use of the instrument for most research purposes. However, the authors encourage the use of additional instruments and caution when using it for clinical diagnosis or other important clinical decisions. Table 4 compares the mean reliability coefficients of several alcohol screening measures obtained through RG studies. Again, with regards to score reliability, the AUDIT performs as well or better than the other commonly used tools. The one exception is the Michigan Alcoholism Screening Test (MAST) which produced scores of similar reliability. Nevertheless, the AUDIT is a 10item instrument while the MAST contains 25 items. The reduced burden of answering 15 fewer items on the AUDIT with the expectation that the scores will be equally reliable is a compelling reason to use the AUDIT. In conclusion, the AUDIT continues to meet its objective as a widely useful alcohol screening tool and these findings support the use of the AUDIT with a wide variety of populations regardless of age, gender, or ethnicity. Furthermore, the AUDIT was created for use in primary care settings and the results here indicate that it can and should continue to be used in this way.

Table 4

Comparison of RG Reliability of Alcohol Screening Instruments

Measure	Objective	No. Items	No. of Studies in RG	RG Reliability	SD
AUDIT[a]	Screen for excessive drinking and in particular to help practitioners identify people who would benefit from reducing or ceasing drinking.	10	24	.79	.1
AUDIT[b]		10	59	.82	.07
CAGE[c]	Lifetime alcohol use screening instrument for clinical populations.	4	13	.74	.09
MAC/MAC-R[d]	A personality based indicator of substance abuse problems.	49	23	.47	.1
BMAST[e]	Brief version of the original MAST, used to assess alcohol problems in the general population.	10			
MAST[f]	Assess alcohol problems in the general population.	25	32	.82	.1
SMAST[g]	Short version of the original MAST, used to assess alcohol problems in the general population.	13	16	.79	.1

[a]AUDIT: Shields and Caruso (2003).
[b]AUDIT: reliability from current study.
[c] CAGE: Shields and Caruso (2004).
[d] MacAndrew Alcohol Scale/MacAndrew Alcohol Scale-Revised (MAC/MAC-R): Miller, Shields, Campfield, Wallace, Weiss (2007).
[e] Brief Michigan Alcoholism Screening Test (BMAST): Shields, Howell, Potter, Weiss (2007)an independent reliability estimate for the BMAST was not reported due to insufficient reliability reporting in the empirical literature.
[f] Michigan Alcoholism Screening Test (MAST): Shields, Howell, Potter, Weiss (2007).
[g]Short Michigan Alcoholism Screening Test (SMAST): Shields, Howell, Potter, Weiss (2007).

REFERENCES

References marked with an asterisk (*) indicate studies included in the meta-analysis

Adams, W. L., Magruder-Habib, K., Trued, S., & Broome, H. L., (1992). Alcohol abuse in elderly emergency department patients. *Journal of the American Geriatric Society, 40,* 1236–1240.

Adams, W. L., Yuan, Z., Barboriak, J. J., & Rimm, A. A. (1993). Alcohol related hospitalizations of elderly people: Prevalence and geographic variation in the United States. *Journal of the American Medical Association, 270*, 1222–1225.

Aertgeerts, B., Buntinx, F., Bande-Knops, J., Vandermeulen, C., Roelants, M., Ansoms, S. et al. (2000). The value of CAGE, CUGE, and AUDIT in screening for alcohol abuse and dependence among college freshmen. *Alcoholism: Clinical and Experimental Research, 24*, 53–57.

*Alati, R., Kinner, S., Najman, J. M., Fowler, G., Watt, K., & Green, D. (2004). Gender differences in the relationships between alcohol, tobacco, and mental health in patients attending an emergency department. *Alcohol & Alcoholism, 39,* 463–469.

Allen, J. P., Reinert, D. F., & Volk, R. J. (2001). The alcohol use disorders identification test: An aid to recognition of alcohol problems in primary care patients. *Preventive Medicine, 33,* 428–433.

American Psychiatric Association . (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: American Psychiatric Association.

Babor, T. F., Higgins-Biddle, J. C., Saunders, J. B., & Monteiro, M. G. (2001). *The alcohol use disorders identification test: Guidelines for use in primary care.* Switzerland: World Health Organization.

*Barry, K. L., & Fleming, M. F. (1993). The alcohol use disorders identification test (AUDIT) and the SMAST-13: Predictive validity in a rural primary care sample. *Alcohol & Alcoholism, 28,* 33–42.

*Bergman, H., & Kallmen, H. (2002). Alcohol use among Swedes and a psychometric evaluation of the alcohol use disorders identification test. *Alcohol & Alcoholism, 37,* 245–251.

*Bischof, G., Reinhardt, S., Grothues, J., Dybek, I., Meyer, C., Hapke, U., et al. (2005). Effects of item sequence on the performance of the AUDIT in general practices. *Drug and Alcohol Dependence, 79,* 373–377.

Bradley, K. A., Boyd-Wickizer, J., Powell, S. H., & Burman, M. L. (1998). Alcohol screening questionnaires in women. *Journal of the American Medical Association, 280,* 166–171.

*Carey, K. B., Carey, M. P., & Chandra, P. S. (2003). Psychometric evaluation of the alcohol use disorders identification test and short drug abuse screening test with psychiatric patients in India. *Journal of Clinical Psychiatry, 64,* 767–773.

Carey, K. B., & Teitelbaum, L. M. (1996). Goals and methods of alcohol assessment. *Professional Psychology: Research and Practice, 27,* 460–466.

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement, 60,* 236–254.

Caruso, J. C., Witkiewitz, K., Belcourt-Dittloff, A., & Gottlieb, J. D. (2001). Reliability of Scores from the Eysenck Personality Questionnaire: A Reliability Generalization Study. *Educational and Psychological Measurement, 61,* 675–689.

Cherpitel, C. J. (1998). Differences in performance of screening instruments for problem drinking among Blacks, Whites, and Hispanics in an emergency room population. *Journal of Studies on Alcohol, 59,* 420–426.

Cherpitel, C. J., & Clark, W. B. (1995). Ethnic differences in performance of screening instruments for identifying harmful drinking and alcohol dependence in the emergency room. *Alcoholism: Clinical and Experimental Research, 19,* 628–634.

*Chung, T., Colby, S. M., Barnett, N P., & Monti, P. M. (2002). Alcohol use disorders identification test: Factor structure in an adolescent emergency department sample. *Alcoholism: Clinical and Experimental Research, 26,* 223–231.

Chung, T., Colby, S. M., Barnett, N. P., Rohsenow, D. J., Spirito, A., & Monti, P. M. (2000). Screening adolescents for problem drinking: Performance of brief screens against DSM-IV alcohol diagnoses. *Journal of Studies on Alcohol, 61,* 579–587.

*Clements, R. (1998). A critical evaluation of several alcohol screening instruments using the CIDI-SAM as a criterion measure. *Alcoholism: Clinical and Experimental Research, 22,* 985–993.

*Conley, T. B. (2001). Construct validity of the MAST and AUDIT with multiple offender drunk drivers. *Journal of Substance Abuse Treatment Special, 20,* 287–295.

*Conley, T. B. (2006). Court ordered multiple offender drunk drivers: Validity and reliability of rapid assessment. *Journal of Social Work Practice in the Additions, 6,* 37–51.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

*Daeppen, J. B., Yersin, B., Landry, U., Pécoud, A., & Decrey, H. (2000). Reliability and validity of the alcohol use disorders identification test (AUDIT) imbedded within a general health risk screening questionnaire: Results of a survey in 332 primary care patients. *Alcoholism: Clinical and Experimental Research, 24,* 659–665.

*Davey, J. D., Obst, P. L., & Sheehan, M. C. (2000). The use of AUDIT as a screening tool for alcohol use in the police work-place. *Drug and Alcohol Review, 19,* 49–54.

*Dawe, S., Seinen, A., & Kavanagh, D. (2000). An examination of the utility of the AUDIT in people with schizophrenia. *Journal of Studies on Alcohol, 61,* 744–750.

Duszynski, K. R., Nicto, F. J., & Valent, C. M. (1995). Repeated practices, attitudes, and confidence levels of primary care physicians regarding patients who abuse alcohol and other drugs. *Maryland Medical Journal, 44,* 439–446.

*Fear, N. T., Iversen, A., Meltzer, H., Workman, L., Hull, L., Greenberg, N., et al. (2007). Patterns of drinking in the UK armed forces. *Addiction, 102,* 1749–1759.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. H. Linn (Ed.), *Educational measurement (3rd ed.*). New York: Macmillan.

*Fleming, M. F., Barry, K. L., & MacDonald, R. (1991). The alcohol use disorders identification test (AUDIT) in a college sample. *Intnl Journal of the Addictions, 26,* 1173–1185.

*Gache, P., Michaud, P., Landry, U., Accietto, C., Arfaoui, S., Wenger, O., et al. (2005). The alcohol use disorders identification test (AUDIT) as a screening tool for excessive drinking in primary care: Reliability and validity of a French version. *Alcoholism: Clinical and Experimental Research, 29,* 2001–2007.

Gallanter, M., Castaneda, R., & Ferman, J. (1988). Substance abuse among general psychiatric patients. : Place of presentation, diagnosis, and treatment. *American Journal of Drug and Alcohol Abuse, 14*, 211–235.

Goetzel, R. Z., Hawkins, K., Ozminkowski, R. J., & Wang, S. (2003). The health and productivity cost burden of the "top 10" physical and mental health conditions affecting six large U.S. employers in 1999. *Journal of Occupational and Environmental Medicine, 45*, 5–14.

*Gomez, A., Conde, A., Santana, J. M., & Jorrin, A. (2005). Diagnostic usefulness of brief versions of alcohol use disorders identification test (AUDIT) for detecting hazardous drinkers in primary care settings. *Journal of Studies on Alcohol, 66*, 305–308.

Grant, B. F., Stinson, F. S., & Harford, T. (2001). The 5-year course of alcohol abuse among young adults. *Journal of Substance Abuse, 13*, 229–238.

Grimm, L. G., & Yarnold, P. R. (Eds.). (2000). Introduction to multivariate statistics. *Reading and Understanding MORE Multivariate Statistics* (pp. 437). Washington, DC: American Psychological Association.

*Hansson, H., Rundberg, J., Zetterlind, U., Johnsson, K. O., & Berglund, M. (2007). Two-year outcome of an intervention program for university students who have parents with alcohol problems: A randomized controlled trial. *Alcoholism: Clinical and Experimental Research, 31*, 1927–1933.

Harwood, H. J., Fountain, D., & Livermore, G. (1998). Economic costs of alcohol abuse and alcoholism. In M. Galanter (Ed.), *Recent developments in alcoholism, vol. 14: The consequences of alcoholism: Medical neuropsychiatric economic cross-cultural.* (pp. 307-330). New York: Plenum Press.

*Hays, R. D., Hill, L., Gillogly, J. J., Lewis, M. W., Bell, R. M., & Nicholas, R. (1993). Response times for the CAGE, short-MAST, AUDIT, and JELLINEK alcohol scales. *Behavior Research Methods, Instruments, & Computers, 25,* 304–307.

*Hays, R. D., Merz, J. F., & Nicholas, R. (1995). Response burden, reliability, and validity of the CAGE, short MAST, and AUDIT alcohol screening measures. *Behavior Research Methods, Instruments, Computers, 27,* 277–280.

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.

*Hermansson, U., Helander, A., Huss, A., Brandt, L., & Ronnberg, S. (2000). The alcohol use disorders identification test (AUDIT) and carbohydrate-deficient transferrin (CDT) ina routine workplace health examination. *Alcoholism: Clinical and Experimental Research, 24,* 180–187.

Horn, J. L., Wanberg, K. H., & Foster, F. M. (1990). Guide to the Alcohol Use Inventory (AUI). Minneapolis, MN: National Computer Systems.

*Ivis, F. J., Adlaf, E. M., & Rehm, J. (2000). Incorporating the AUDIT into a general population telephone survey: A methodological experiment. *Drug and Alcohol Dependence, 60,* 97–104.

Joseph, C. L., Rasmussen, J., Ganzini, L., & Atkinson, R. M. (1997). Outcome of nursing home care for residents with alcohol use disorders. *International Journal of Geriatric Psychiatry, 12,* 767–772.

*Karno, M., Granholm, E., & Lin, A. (2000). Factor structure of the Alcohol Use Disorders Identification Test (AUDIT) in a mental health clinic sample. *Journal of Studies on Alcohol, 61,* 751–758.

Kazdin, A. E. (Ed.). (1992). *Methodological issues & strategies in clinical research.* Washington, DC: American Psychological Association.

*Kelly, T. M., & Donovan, J. E. (2001). Confirmatory factor analyses of the alcohol use disorders identification test (AUDIT) among adolescents treated in emergency departments. *J. Stud. Alcohol, 62,* 838–842.

*Kelly, T. M., Donovan, J. E., Chung, T., Cook, R. L., & Delbridge, T. R. (2004). Alcohol use disorders among emergency department treated older adolescents: A new brief screen (RUFT-Cut) using the AUDIT, CAGE, CRAFFT, an RAPS-QF. *Alcoholism: Clinical and Experimental Research, 28,* 746–753.

*Kelly, T. M., Donovan, J. E., Kinnane, J. M., & Taylor, D. (2002). A comparison of alcohol screening instruments among under-aged drinkers treated in emergency departments. *Alcohol and Alcoholism, 37,* 444–450.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd. ed.). Englewood Cliffs: Prentice Hall.

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *Journal of Experimental Education, 69,* 280–309.

Knight, J. R., Sherritt, L., Harris, S. K., Gates, E. C., & Chang, G. (2003). Validity of brief alcohol screening tests among adolescents: A comparison of the AUDIT, POSIT, CAGE, and CRAFFT. *Alcoholism: Clinical and Experimental Research, 27,* 67–73.

*Kokotailo, P. K., Egan, J., Gangnon, R., Brown, D., Mundt, M., & Fleming, M. (2004). Validity of the alcohol use disorders identification test in college students. *Alcoholism: Clinical and Experimental Research, 28,* 914–920.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

*Lapham, S. C., Brown, P., Suriyawongpalsal, P., Skipper, B. J., Chadbunchachai, W., & Palsarnsilp, S. (1999). Use of AUDIT for alcohol screening among emergency room patients in Thailand. *Substance Use & Misuse, 34,* 1881–1895.

*Lennings, C. J. (1999). An evaluation of the Leeds dependence questionnaire. *J Child & Adolescent Substance Abuse, 8,* 73–87.

*Lennings, C. J., Feeney, G. F., Sheehan, M., Young, R., McPherson, A., & Tucker, J. (1997). Work-place screening of mine employees using the alcohol use disorders identification test (AUDIT) and alcohol breathalyzation. *Drug and Alcohol Review, 16,* 357–363.

*Leonardson, G. R., Kemper, E., Ness, F. K., Koplin, B. A., Daniels, M. C., & Leonardson, G. A. (2005). Validity and reliability of the AUDIT and CAGE aid in Northern Plains American Indians. *Psychological Reports, 97,* 161–166.

Li, H., Rostenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods, 1,* 98–107.

*Lima, C. T., Freire, A. C., Silva, A. P., Teixeira, R. M., Farrell, M., & Prince, M. (2005). Concurrent and construct validity of the AUDIT in an urban Brazilian sample. *Alcohol & Alcoholism, 40,* 584–589.

Lord, F. M., & Novick, M. R. (1968). *Statistical theory of mental test scores.* Reading, MA: Addison-Wesley.

Mayfield, D., McLeod, G., & Hall, P. (1974). The CAGE questionnaire: Validation of a new alcoholism instrument. *American Journal of Psychiatry, 131*, 1121–1123.

*Maisto, S. A., Conigliaro, J., McNeil, M., Kraemer, K., & Kelley, M. E. (2000). An empirical investigation of the factor structure of the AUDIT. *Psychological Assessment, 12,* 346–353.

Maisto, S. A., & Connors, G. J. (1990). Clinical diagnostic techniques and assessment tools in alcohol research. *Alcohol Health & Research World.Special Issue: Tools of Alcohol Research, 14,* 232–238.

*McCann, B. S., Simpson, T. L., Ries, R., & Roy-Byrne, P. (2000). Reliability and validity of screening instruments for drug and alcohol abuse in adults seeking evaluation for attention-deficit/hyperactivity disorder. *American Journal on Addictions, 9,* 1– 9.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37,* 113–115.

Miller, G. A. (1985). *The Substance Abuse Subtle Screening Inventory manual.* Bloomington, IN: Glenn A. Miller.

Miller, C. A., Shields, A. L., Campfield, D., Wallace, K. A., & Weiss, R. D. (2007). Substance use scales of the Minnesota Mulitphasic Personality Inventory: An exploration of score reliability via meta-analysis. *Educational and Psychological Measurement, 67,* 1052–1065.

Moore, R. D., Bone, L. R., Geller, G., Mamon, J. A., Stokes, E. J., & Levine, D. M. (1989). Prevalence, detection, and treatment of alcoholism in hospitalized patients. *Journal of the American Medical Association, 261,* 403–407.

National Institute on Alcohol Abuse and Alcoholism (2000, June). *Tenth special report to the U.S. congress on alcohol and health*. Bethesda, MD: National Institute on Drug Abuse.

*Neumann, T., Neuner, B., Gentilello, L. M., Weiss-Gerlach, E., Mentz, H., Rettig, J. S. et al. (2004). Gender differences in the performance of a computerized version of the alcohol use disorders identification test in subcritically injured patients who are to be admitted to the emergency department. *Alcoholism: Clinical and Experimental Research, 28,* 1693–1701.

Newcomb, M. D., Galaif, E. R., & Locke, T. F. (2001). Substance use diagnoses within a community sample of adults: Distinction, comorbidity, and progression over time. *Professional Psychology: Research and Practice, 32*, 239–247.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd ed.)*. New York: McGraw-Hill.

*O'Hare, T. (2005). Comparing the AUDIT and 3 drinking indices as predictors of personal and social drinking problems in freshman first offenders. *Journal of Alcohol and Drug Education, 49,* 37–61.

*O'Hare, T., & Sherrer, M. V. (1999). Validating the alcohol use disorder identification test with college first-offenders. *Journal of Substance Abuse Treatment, 17*, 113–119.

*O'Hare, T., Sherrer, M. V., LaButti, A., & Emrick, K. (2004). Validating the alcohol use disorders identification test with persons who have a serious mental illness. *Research on Social Work Practice, 14,* 36–42.

*Pal, H. R., Jena, R., & Yadav, D. (2004). Validation of the alcohol use disorders identification test (AUDIT) in urban community outreach and de-addiction center samples in North India. *J. Stud. Alcohol, 65,* 794–800.

Piccinelli, M., Tessari, E., Bortolomasi, M., Piasere, O., Semenzin, M., Garzotto, N., et al. (1997). Efficacy of the alcohol use disorders identification test as a screening tool for hazardous alcohol intake and related disorders in primary care: A validity study. *British Medical Journal, 314,* 420–424.

Reinert, D. F., & Allen, J. P. (2002). The alcohol use disorders identification test (AUDIT): A review of recent research. *Alcoholism: Clinical and Experimental Research, 26,* 272–279.

Reinherz, H. Z., Giaconia, R. M., Lefkowitz, E. S., & Pakiz, B. (1993). Prevalence of psychiatric disorders in a community population of older adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry, 32,* 369–377.

Rice, D. P., Kelman, S., & Miller, L. S. (1990). The economic costs of alcohol and drug abuse and mental illness: 1985. Report Submitted to the Alcohol, Drug Abuse and Mental Health Administration, U.S. Department of Human Services. DHHS Publication No. (ADM) 90-1694. San Francisco: Institute for Health & Aging, University of California.

Ridenour, T. A., Cottler, L. B., Compton, W. M., Spitznagel, E. L., & Cunningham-Williams, R. M. (2002). Is there a progression from abuse disorders to dependence disorders? *Addiction, 98,* 635–644.

Rohde, P., Lewinsohn, P. M., Kahler, C. W., Seeley, J. R., & Brown, R. A. (2001). Natural course of alcohol use disorders from adolescence to young adulthood. *Journal of American Academic Child and Adolescent Psychiatry, 40,* 83–90.

*Rumpf, H. J., Hapke, U., Meyer, C. & John, U. (2002). Screening for alcohol use disorders and at-risk drinking in the general population: Psychometric performance of three questionnaires. *Alcohol & Alcoholism, 37,* 261–268.

Saunders, J. B., Aasland, O. G., Babor, T. F., De la Fuente, J. R., & Grant, M. (1993). Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction, 88*, 791–803.

Saunders, S. M., Zygowicz, K. M., & D'Angelo, B. R. (2006). Person-related and treatment-related barriers to alcohol treatment. *Journal of Substance Abuse Treatment, 30,* 261–270.

*Schmidt, A., Barry, K. L., & Fleming, M. F. (1995). Detection of problem drinkers: The alcohol use disorders identification test (AUDIT). *Southern Medical Journal, 88*, 52–59.

*Selin, K. H. (2003). Test-retest reliability of the alcohol use disorder identification test in a general population sample. *Alcoholism: Clinical and Experimental Research, 27*, 1428–1435.

Selzer, M. L. (1971). The Michigan Alcohol Screening Test: The quest for a new diagnostic instrument. *American Journal of Psychiatry, 127*, 1653–1658.

Shields, A. L., & Caruso, J. C. (2003). Reliability generalization of the alcohol use disorder identification test. *Educational and Psychological Measurement, 63*, 404–413.

Shields, A. L. & Caruso, J.C. (2004). A reliability induction and reliability generalization study of the CAGE questionnaire. *Educational and Psychological Measurement, 63*, 1–17.

*Shields, A. L., Guttmannova, K., & Caruso, J. C. (2004). An examination of the factor structure of the alcohol use disorders identification test in two high-risk samples. *Substance Use & Misuse, 39,* 1161–1182.

*Skipsey, K., Burleson, J. A., & Kranzler, H. R. (1997). Utility of the AUDIT for identification of hazardous or harmful drinking in drug-dependent patients. *Drug and Alcohol Dependence, 45,* 157–163.

Smith-Black, B. Rabins, P. V., & McGuire, M. H. (1998). Alcohol and mortality: Alcohol use disorder is a risk factor for mortality among older public housing residents. *International Psychogeriatrics, 10,* 309–327.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3,* 271–295.

*Stalhbrandt, H., Johnsson, K. O., & Berglund, M. (2007). Two-year outcome of alcohol interventions in Swedish university halls of residence: A cluster randomized trial of a brief skills training program, twelve step-influenced intervention, and controls. *Alcoholism: Clinical and Experimental Research, 31,* 458–466.

Steinbauer, J. R., Cantor, S. B., Holzer, C. E., & Volk (1998). Ethnic and sex bias in primary care screening tests for alcohol use disorders. *Annals of Internal Medicine, 129,* 353–362.

Tarter, R. E., & Vanyukov, M. (1994). Alcoholism: A developmental disorder. *Journal of Consulting and Clinical Psychology, 62,* 1096–1107.

Thom, B., Herring, R., & Judd, A. (1999). Identifying alcohol-related harm in young drinkers: The role of accident and emergency departments. *Alcohol and Alcoholism, 34,* 910–915.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837–847.

Thompson, B. (2000). Reliability generalization [Special Section]. *Educational and Psychological Measurement, 60,* 157–254.

Thompson, R. L. (1999, January). *Reliability generalization: An important meta-analytic method, because it is incorrect to say, "The test is unreliable."* Paper presented as the Annual Meeting of the Southwest Educational Research Association, San Antonio, TX.

*Tsai, M. C., Tsai, Y. F., Chen, C. Y., & Liu, C. Y. (2005). Alcohol use disorders identification test (AUDIT): Establishment of cut-off scores in a hospitalized Chinese population. *Alcoholism: Clinical and Experimental Research, 29,* 53–57.

Vacha-Haase, T. (1998). Reliability Generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6–20.

Vacha-Haase, T., Henson, R., & Caruso, J. C. (2002). Reliability Generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement, 62,* 562–569.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology, 10*, 413–425.

*Varvasovsky, Z., & McKee, M. (2000). Problem drinking among hospitalized patients in Hungary. *Alcohol & Alcoholism, 35,* 574-579.

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*, 224–235.

Volk, R. J., Steinbauer, J. R., Cantor, S. B., & Holzer, C. E. (1997). The Alcohol Use Disorders Identification Test (AUDIT) as a screen for at-risk drinking in primary care patients of different racial/ethnic backgrounds. *Addiction, 92,* 197–206.

*Wennberg, P. (1996). The alcohol use disorders identification test (AUDIT): A psychometric evaluation. *Reports from the Dept of Psych, Stockholm University, 811,* 1–14.

*West, S. L., & Graham, C. W. (2001). Assessing the validity of college alcohol problem scale with African American undergraduates: Results of preliminary investigation. *Journal of Alcohol and Drug Education, 46,* 4–17.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58,* 21–37.

Wilkinson, L. & the American Psychological Association Task Force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist, 54,* 594–604.

Willson, V. L. (1980). Research techniques in AERJ articles: 1969 to 1978. *Educational Researcher, 9,* 5–10.

Windle, M. (2003). Alcohol use among adolescents and young adults. *Alcohol Research & Health, 27*(1), 79–85.

Wolford, G. L., Rosenberg, S. D., Drake, R. E., Mueser, K. T., Oxman, T. E., Hoffman, D., et al. (1999). Evaluation of methods for detecting substance use disorder in persons with severe mental illness. *Psychology of Addictive Behaviors, 13,* 313–326.

Yin, P. & Fan, X. (2000). Assessing the reliability of Beck depression inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60,* 201–223.

VITA

CHANDNI PATEL

Personal Data:	Date of Birth: July 19, 1984
	Place of Birth: Gaborone, Botswana
	Marital Status: Single
Education:	B.S. Psychology, Kennesaw State University, Kennesaw, Georgia 2006
	M.A. Clinical Psychology, East Tennessee State University, Johnson City, Tennessee 2008
Professional Experience:	Graduate Assistant, East Tennessee State University, Department of Psychology, 2006 – 2008
	Intern, Frontier Health, Fall 2007
	Intern, Sullivan County Sheriff's Office, Spring 2008