Score Reliability of Adolescent Alcohol Screening Measures: A Meta-Analytic Inquiry

Alan L. Shields
Delia C. Campfield
Christopher S. Miller
Ryan T. Howell
Kimberly Wallace
Roger D. Weiss

Alan L. Shields, PhD, is affiliated with Department of Psychology, East Tennessee State University, Johnson City, TN; McLean Hospital, Harvard Medical School, Belmont, MA; and PRO Consulting, Pittsburgh, PA.

Delia C. Campfield, MA, Christopher S. Miller, MA, and Kimberly Wallace, PhD, are affiliated with the Department of Psychology, University of Montana, Missoula, MT.

Ryan T. Howell, PhD, is affiliated with San Francisco State University, San Francisco, CA.

Roger D. Weiss, MD, is affiliated with McLean Hospital/Harvard Medical School, Belmont, MA.

This research was supported by grants from the National Institute on Alcohol Abuse and Alcoholism (R21-AA13423-01; Kim Wallace and Alan Shields) and the National Institute on Drug Abuse (DA-00326 and DA-15831; Roger Weiss). The authors thank Jennifer Gottlieb and Carolee McCall for their project development assistance and Nabil Haddad, PhD, Department of Psychology Chairperson at the University of Montana, for his support of this research.

Address correspondence to Alan L. Shields, PhD, East Tennessee State University, Department of Psychology, Box 70649, Johnson City, TN 37614–1702 (E–mail: ashields@patientreported.com).

Journal of Child & Adolescent Substance Abuse, Vol. 17(4) 2008
Available online at http://jcasa.haworthpress.com

doi: 10.1080/15470650802292855

ABSTRACT. This study describes the reliability reporting practices in empirical studies using eight adolescent alcohol screening tools and characterizes and explores variability in internal consistency estimates across samples. Of 119 observed administrations of these instruments, 40 (34%) reported usable reliability information. The Personal Experience Screening Questionnaire—Problem Severity scale generated average reliability estimates exceeding 0.90 (95% CI = 0.90–0.96) and the Adolescent Alcohol Involvement Scale generated average score reliability estimates below 0.80 (95% CI = 0.67–0.85). Average reliability estimates of the remaining instruments were distributed between these extremes. Sample characteristics were identified as potentially important predictors of variability in the reliability estimates of all the instruments and all instruments under evaluation generated more reliable scores in clinical settings ($M = 0.89$) as opposed to nonclinical settings ($M = 0.82$; r effect size (38) = 0.29, $p < .10$). Clinicians facing instrument selection decisions can use these data to guide their choices and researchers evaluating the performance of these instruments can use these data to inform their future studies.

KEYWORDS. Adolescent alcohol screening, psychometrics, reliability generalization, reliability

The primary goal of alcohol screening is to identify the presence of alcohol use problems among patients and research participants. The prevalence and scope of negative consequences associated with adolescents alcohol use (Johnston, O'Malley, & Bachman, 2003) supports the recommendation for routine alcohol screening among this population (American Academy of Pediatrics, Committee on Substance Abuse, 1993; Elster & Kuznets, 1994). Based on reports distinguishing adolescent alcohol use and problem trajectories from those of adults (Clark, 2004; Deas, Riggs, Langenbucher, Goldman, & Brown, 2000), a number of adolescent-specific alcohol use screening tools have been developed over the past decade (Martin & Winters, 1998).

As the number of available adolescent alcohol screening tools grows, clinicians and researchers are challenged in the task of selecting the tool most appropriate for their needs. To help in this regard, the Substance Abuse and Mental Health Services Administration (SAMHSA, 1999) reconvened an expert panel to provide guidelines for evaluating and selecting adolescent-specific screening and assessment instruments. These guidelines include practical considerations

such as knowing the intended purpose of the instrument, the specific adolescent population for which the instrument was developed, and the settings in which the instrument was normed. Also recommended is that instrument selection be guided by reliability and validity evidence related to the tools. The focus of the present study is on the property of reliability.

Referring to consistency in measurement, there are a variety of reasons why clinicians and researchers consider reliability when selecting instruments. First, higher reliability estimates tend to increase our confidence that selected test items assess (a) the same construct (e.g., internal consistency reliability), (b) the same construct at different time points (e.g., test–retest reliability), and (c) the same construct when administered by separate test-givers (e.g., interrater reliability). Further, reliability estimates can be used to inform how close an examinee's score is to his actual or true score (i.e., the test score an examinee would get if it was completely free of measurement error or "noise"; Nunnally & Bernstein, 1994). This is important in research and clinical settings but particularly relevant for health professionals making clinical decisions based on individual test scores. Scores obtained from assessment tools (e.g., an alcohol screening instrument) are imperfect and only approximate what a respondent would actually score if assessed under perfect circumstances. Accounting for this error, confidence intervals offer some index of assurance of how well a respondent's obtained score reflects his actual or true score. Diminishing reliability estimates are associated with wider confidence intervals and wider confidence intervals indicate lower precision in measurement (Charter, 1996; Charter and Feldt, 2001). Thus, low reliability is associated with more diagnostic and respondent classification errors and, ultimately, reduced confidence in clinical decisions.

The importance of considering reliability information in measurement decisions and data interpretation is virtually universal and the benefits of doing so are well documented. Nevertheless, difficulties are often experienced when interpreting specific reliability estimates and values (Charter & Feldt, 2002). One apparently troublesome reliability concept is that reliability is a property of scores and not tests and, therefore, fluctuates across administrations (Thompson & Vacha-Haase, 2000). In other words, tests, surveys, or questionnaires (i.e., the actual instruments) are neither reliable nor unreliable; instead, it is the scores or results they produce that are subject

to reliability. Therefore, a given test manual's reported reliability coefficient cannot be understood as *the* reliability of the test as it is dependent on the sample from which it was derived. Later samples may have characteristics that cause reliability to be lower or higher than in the normative sample, which, as noted, impacts the usefulness of the instrument. Because of this, current standards recommend that reliability data be included in all empirical research, not just psychometric studies, and be based on the data gathered for a given study (Wilkinson & the American Psychological Association's Task Force on Statistical Inference, 1999).

To date, both primary studies and reviews of adolescent alcohol screening instruments have reflected on reliability as a static estimate and report, for example, that alcohol screening tool X is reliable based on one or more previous studies. This is important information for test users selecting instruments yet is limited. By not taking a "score reliability" perspective (i.e., acknowledging reliability as a property of scores and not tests, per se), these studies do not take into account the variable nature of reliability and the contextual information necessary to inform whether the reported reliability might generalize to the circumstances under which test users plan to administer it. As noted, reliability estimates fluctuate, often substantially, across test administrations. This fluctuation begs the question, what is influencing those estimates? An improved understanding of the typical score reliability as well as the determinants of reliability fluctuations among adolescent alcohol screening tools can be useful for those who are deciding whether or not to use a given adolescent alcohol screening tool.

THE PRESENT STUDY

The primary goal of this article was to meta-analytically synthesize the adolescent alcohol screening literature. This synthesis will provide more comprehensive, instrument-specific reliability information available to clinicians and researchers. More specifically, this study used meta-analytic methods to examine score reliability of eight commonly used adolescent alcohol screening measures. This study had two main objectives: (a) to provide a characterization of the reliability of scores for each measure across studies and (b) to explore the relationship among sample characteristics and score reliability within each instrument. The archival data collection approach also

allows for the evaluation of reliability reporting practices in the adolescent alcohol screening literature.

METHOD

Measures

The National Institute for Alcohol Abuse and Alcoholism (NIAAA) published a comprehensive guidebook for assessing alcohol problems (NIAAA, 1995), which identified over 20 alcohol screening tools, of which 12 could be used with adolescents. Four of these were developed for use among adults and therefore were not selected for these analyses. The screening measures used in the present study include the Adolescent Alcohol Involvement Scale (AAIS; Mayer & Filstead, 1979), the Adolescent Drinking Index (ADI; Harrell & Wirtz, 1989), the Drug Use Screening Inventory (DUSI; Tarter, 1990), the Perceived Benefit of Drinking Scale (PBDS; Petchers & Singer, 1987), the Personal Experience Screening Questionnaire (PESQ; Winters, 1992), the Problem Oriented Screening Instrument for Teenagers (POSIT; Rahdert, 1991), the Rutgers Alcohol Problem Index (RAPI; White & Labouvie, 1989), and Young Adult Alcohol Problem Screening Test (YAAPST; Hurlbut & Sher, 1992). A brief description of each can be found in Table 1.

Most of the identified tools are generally thought to support a unidimensional factor structure. However, several of the identified instruments are multifactored and thought to assess constructs both within and outside the domain of alcohol use. Therefore, for multifactored instruments, we focus on the unidimensional scale that is used for alcohol screening purposes. This includes the POSIT Alcohol and Drug Use/Abuse Scale (POSIT-ADS), the DUSI Substance Use Scale (DUSI-SUS), and the PESQ Problem Severity scale (PESQ-PS).

Methodological Framework

Reliability Induction

Vacha-Haase, Kogan, and Thompson (2000) refer to the practice of ''referencing the reliability coefficients from prior reports as the sole warrant for presuming the score integrity of entirely new data''

TABLE 1. List and Brief Description of the Commonly Used Adolescent Alcohol Screening Tools

Public Domain Instruments	Brief Description and Construct Assessed
Perceived Benefit of Drinking Scale (PBDS)	Description: This 5-item tool designed particularly for high school juniors and seniors can be completed and scored in about five minutes. Each item is endorsed as true or false by the respondent and summing items creates a single composite score ranging from 0 to 5 with higher scores more indicative of alcohol problems. Conceptualization of adolescent alcohol problems: The potential for problem drinking is conceptualized by assessing drinking motives. Increased perceived benefits from drinking will be associated with increased drinking behavior.
Adolescent Alcohol Involvement Scale (AAIS)	Description: This 14-item tool can be completed and scored in about five minutes. Respondents choose from multiple answers per item and each item is assigned a point value based on the highest scoring answer per item. Summed scores reflect a unidimensional factor structure and categorize respondents as abstainer/occasional drinker (0–19), normal user/ relative abstainer (20–41), misuser (42–57), and abuser/dependent (58–79). Conceptualization of adolescent alcohol problems: In addition to quantity and frequency, alcohol problems are reflected by the degree that it interferes with psychological functioning, social relations, and family living.
Rutgers Alcohol Problem Index (RAPI)	Description: This 23-item tool can be completed and scored in about 10 minutes. Items are endorsed on a 0–4 scale and summed to produce a composite score (0 to 69) reflecting a unidimensional factor structure. Higher scores represent increased frequency of negative consequences due to alcohol use. The RAPI was designed to evaluate the three years of use prior to assessment but can be modified for other time frames (e.g., the past six months). Conceptualization of adolescent alcohol problems: In addition to quantity and frequency and drinking patterns, alcohol problems are reflected by negative consequences associated with problem drinking.
Young Adult Alcohol Problem Screening Test (YAAPST)	Description: This 36-item tool assesses lifetime, past-year, and past-year's frequency of negative consequences of alcohol use among college students. It can be completed and scored in about 20 minutes. Items are differentially weighted and, when summed, produce a composite score reflecting a unidimensional factor structure. The YAAPST was designed for older adolescents attending college, but it can also be used with other adolescents.

(Continued)

TABLE 1. Continued

	Conceptualization of adolescent alcohol problems: Alcohol problems are considered primarily in terms of negative consequences of alcohol use. In addition to conventional dependence symptoms, age-specific consequences of alcohol use like poor grades and/or getting into regrettable sexual situations are considered.
Problem-Oriented Screening Instrument for Teenagers (POSIT)	Description: This 139-item instrument, designed for adolescents 12–19 years old, screens the potential need for service in 10 problem areas. The POSIT takes about 25 minutes to complete and items are dichotomously scored as either "Yes" (1) or "No" (0). The 17-item alcohol and drug use/abuse scale (POSIT-ADS) has a range of scores from 0 to 17, screens for drug and alcohol use simultaneously, and is the focus of the present study. One of the POSIT-ADS items specifies "past month" as the time period under evaluation while the remaining items tap lifetime experiences.
	Conceptualization of adolescent alcohol problems: Alcohol problems are considered primarily in terms of negative consequences of alcohol use across a variety of the life areas that may be affected by adolescent alcohol use.
Proprietary Instruments	
Adolescent Drinking Index (ADI)	Description: This 24-item tool can be completed and scored in about 10 minutes. It was developed to support a multifactor structure but a single composite score can be obtained by summing all 24 items. Items are endorsed on either a 3-point or a 4-point scale and the total score range is 0–62, with higher scores indicative of a greater likelihood of alcohol problems. ADI items focus on the problems that arise from alcohol use, particularly in the 12 months prior to evaluation. This instrument was developed especially for adolescents referred for emotional and behavioral problems.
	Conceptualization of adolescent alcohol problems: Test developers approach alcohol-related dysfunction within a multidimensional framework. Dysfunction is identified and understood by evaluating loss of drinking control and the social, psychological, and physical indicators of alcohol abuse.
Personal Experience Screening Questionnaire (PESQ)	Description: This 40-item self-report takes about 10 minutes to complete and supports a multiscale factor structure. The Problem Severity Scale consists of 18 items, each of which contains four scoring options test-takers endorse in response to questions about attitudes, behaviors, and consequences of substance abuse. The PESQ was designed for a general adolescent population and assesses alcohol and drug use simultaneously. It evaluates lifetime use with special emphasis on the 12 months prior to assessment.

(Continued)

TABLE 1. Continued

	Conceptualization of adolescent alcohol problems: The test developer approaches adolescent alcohol and substance use problems within a multidimensional framework. These problems are identified and understood by evaluating dependence symptoms, benefits of use, and behaviors related to and consequential of substance abuse.
Drug Use Screening Inventory (DUSI)	Description: This original and its revised 159-item tool evaluate the severity of past year consequences related to alcohol and other drug use among adolescents age 16 or older. This instrument takes approximately 20 minutes to complete. Aside from frequency questions, items are endorsed as "mostly yes" or "mostly no" as applied to the respondent. Endorsed responses produce three scores describing problem severity, relative problem severity, and overall problem severity. The 15-item Substance Use Scale (DUSI-SUS) screens for alcohol and drug use simultaneously and is the focus in the present study.
	Conceptualization of adolescent alcohol problems: Drug and alcohol use cannot be disentangled from other life experiences. Test developers conceptualize alcohol and drug use in terms of quantity and frequency of use and level of cooccurring medical, psychiatric, and psychosocial maladjustment.

(p. 512) as reliability induction (RI) and concluded that rarely is it an empirically defensible procedure. Shields and Caruso (2004) further delineated RI practices by designating the practice of not including reliability estimates as RI by omission and the reporting of reliability estimates from previous research as RI by report. Along with accurate reliability reporting, these two RI designations provide a framework around which the reliability reporting practices in the adolescent alcohol screening literature can be documented.

Reliability Generalization

The meta-analytic method of reliability generalization (RG) was introduced by Vacha-Haase (1998) to examine score reliability associated with a given measure, the variability in score reliability across test administrations, and the determinants of that variability (see Henson & Thompson, 2002, for a review of RG methods).

Archival Data Collection

A literature search using the American Psychological Association's PsychINFO and the National Library of Medicine's PubMed (Medline) databases was performed to identify studies employing the eight adolescent alcohol screening identified above. Two search strategies were used for each instrument. First, the full name of the instrument was entered and results were obtained. Next, the instrument acronym along with the terms *alcohol* OR *screening* was entered and results were obtained. Duplicates were eliminated. The search was limited to articles published in peer-reviewed journals in English prior to May 2006. The final step of the literature search involved manually reviewing reference lists of the remaining articles to identify studies not captured in the original search.

After eliminating false-positive citations, meta-analyses, review papers, letters to the editor, and other nonempirical reports, a list of 120 articles employing these measures was produced. Of these 120 articles, we were unable to locate one. All articles were read and categorized by the reliability reporting practice employed.

Criterion Variable

The criterion in RG studies is generally some numeric index of reliability. In the present study we use internal consistency reliability estimates as our criterion variable. Internal consistency estimates are important for interpreting individual test scores and are vital in assessing outcomes based on group administrations (Charter, 1997; Charter & Feldt, 2002). Further, internal consistency estimates are an increasingly common type of reported estimate of reliability (Hogan, Benjamin, & Brezinski, 2000) and the most commonly reported reliability coefficient in this review of the adolescent alcohol screening literature. Because the vast majority of studies reporting internal consistency reported it as Cronbach's alpha (α; Cronbach, 1951), we will report a as the family of internal consistency estimates used in this analysis. While a detailed explanation of internal consistency and α is outside the scope of the present study, the interested reader is directed to Henson (2001) for an excellent primer.

Predictor Variables

In addition to describing the typical reliability of scores from the identified adolescent alcohol screening tools, this study sought to

explore sample characteristics associated with their respective score reliability estimates. Although a wide variety of factors could potentially influence score reliability, predictor variable selection was limited in two important ways. First, we are not aware of any other study evaluating determinants of adolescent alcohol screening tool score reliability. Therefore, we lack any theoretical or practical guide in this regard. Next, the type of variables is restricted to what is provided in the literature. With this in mind, five predictor variables were identified. These include score variability (coded as the SD of scores), sample age (coded as the average sample age), gender (coded as the proportion of males in the sample), sample ethnicity (coded as the proportion of Caucasians in the sample), and sample type (coded dichotomously as a patient [1] or non-patient [0] sample).

Data Analysis

Reporting practices were evaluated via the frequency with which researchers engaged in RI by omission, RI by report, or reported reliability information based on data gathered for their study. Next, estimates of central tendency, dispersion, and tests of homogeneity were determined to illustrate the common reliability estimates and their variance within measures, as reported in past samples. Further 95% confidence intervals (CIs) were constructed around the aggregate score reliability estimate to help future test users determine, with 95% certainty, what type of reliability to expect when they use a particular measure. Finally, the predictor variables were correlated with the reliability coefficients to test their relationship with score reliability.

RESULTS

Reliability Induction

Of the 119 adolescent screening instances observed, a total of 49 (41%) were accompanied by reliability information based on the data from that study. The remaining measurement instances ($n = 70$; 59%) used RI procedures. More specifically, 27 (23%) reported reliability information based on data other than that being analyzed (i.e., RI by report) and 43 (36%) provided no reliability information

TABLE 2. Reliability Reporting within and Across Instruments

	Na	Accurate Reliability Reported	%	Reliability Induction by Report	%	Reliability Induction by Omission	%
Instrument							
DUSI-SUS	29	10	34	9	31	10	14
AAIS	25	9	36	3	12	13	52
POSIT-ADS	21	9	43	5	24	7	33
RAPI	17	9	53	2	12	6	35
PBDS	9	3	33	4	44	2	22
PESQ-PS	7	3	43	0	0	4	57
YAAPST	7	5	72	1	14	1	14
ADI	4	1	25	3	75	0	0
Total	119	49	41	27	23	43	36

Note. [a] Total number of studies using the identified instrument (all obtained from the peer-reviewed literature).

(i.e., RI by omission). Table 2 specifies reliability reporting practices by instrument.

Reliability Generalization

Data Reduction

Of the 49 measurement opportunities that included score reliability information, some had to be eliminated. Internal consistency reliability, specifically a, is the selected criterion in the present study. Therefore, other types of reliability estimates (e.g., interrater or test–retest reliability) were removed. Reliability coefficients reported in previous research on the same sample, in an unusable form (e.g., average reliability across scales or subsamples) or for scales other than the identified alcohol screening subscale (e.g., Santisteban, Tejeda, Donimicis, & Szapocznik, 1999, reported reliability only for the Family domain of the POSIT), were also eliminated from the present analysis. Other reasons for exclusion included using a significantly shortened version of the original scale (e.g., Pedersen & Skrondal, 1999) or significantly modified version of the original scale (e.g., Mayer & Ligman, 1989). Note that while Borsari and Carey (2000) did not include reliability data in their original work, it was included in this meta-analysis as that information was provided

in Neal and Carey (2004). Data reduction procedures left a total of 40 samples evaluated in the present study.

Characterizing Score Reliability Across Samples

The second set of analyses generated estimates of central tendency and dispersion for the criterion and predictor variables. Table 3 provides these descriptive statistics, both within and across measures, and highlights the typical reliability and range of reliability coefficients reported in past samples. Results from the ADI and PBDS are not presented in Table 3 due to insufficient reliability information found in the literature (i.e., less than two studies with reliability information).

To ensure that aggregate reliability statistics were unbiased with respect to sample size, both weighted and unweighted estimates are presented in Table 3. The unweighted mean reliability estimate across instruments was .85 and this was very similar to the weighted average

TABLE 3. Descriptive Statistics for Internal Consistency Reliability Coefficients Across Instruments

Variable	*K*/*N*	*M*	*Mdn*	*SD*	Low	High	Range
Unweighted score reliability (α)							
PESQ-PS	5	0.93	0.93	0.02	0.90	0.96	0.06
RAPI	8	0.89	0.88	0.05	0.81	0.95	0.14
DUSI-SUS	6	0.90	0.89	0.04	0.87	0.99	0.12
POSIT-ADS	5	0.86	0.86	0.06	0.77	0.92	0.15
YAAPST	4	0.83	0.84	0.08	0.72	0.91	0.19
AAIS	12	0.76[†]	0.79	0.14	0.51	0.94	0.43
Statistics across all measures	40	0.85	0.88	0.11	0.51	0.99	0.49
Weighted score reliability (α)							
PESQ-PS	2433	0.92	0.92	0.02	0.90	0.96	0.06
RAPI	3085	0.91	0.92	0.04	0.81	0.95	0.14
DUSI-SUS	1816	0.90	0.89	0.03	0.87	0.99	0.12
POSIT-ADS	4495	0.86	0.85	0.03	0.77	0.92	0.15
YAAPST	1271	0.85	0.84	0.05	0.72	0.91	0.19
AAIS	2501	0.77[a]	0.79	0.12	0.51	0.94	0.43
Statistics across all measures	15,601	0.87	0.88	0.07	0.51	0.99	0.49

Note. K = the total number of independent samples; N = the number of participants within those samples. The AAIS total N of 11 samples yielded 2,501 respondents (while 12 samples provided reliability coefficients, only 11 specified the sample size).

[a]Heterogeneity tests demonstrated that the unweighted weighted mean reliability coefficients for the AAIS were statistically lower then the other five measures as a set.

[†]Heterogeneity tests demonstrated that the unweighted weighted mean reliability coefficients for the AAIs were statistically lower than the other five measures as a set.

(0.87). Unweighted reliability estimates across all six instruments ranged from a high of 0.93 for the PESQ-PS scores to a low of 0.76 for AAIS scores. This trend held for weighted reliability estimates. In addition to having the highest mean reliability, the PESQ-PS showed the least amount of variability in reliability estimates in terms of SD (0.02) and range (0.06). The AAIS showed the most variability in reliability estimates in terms of SD (0.14) and range (0.43). The 95% CIs for the evaluated instruments include PESQ-PS (0.90–0.96), DUSI-SUS (0.86–0.95), RAPI (0.85–0.93), POSIT-ADS (0.79–0.94), YAPPST (0.70–0.85), and AAIS (0.67–0.85). Though each instrument has variability in reliability when examined alone, when combined, the 95% confidence interval is quite tight (0.81–0.88).

Homogeneity Tests

Homogeneity tests were performed to determine if the reliability coefficients across instruments were statistically similar. When examining all 40 reliability coefficients we find them to be heterogeneous ($\chi^2_{(39)} = 91.36$; $p = .000002$). This suggests that the reliability coefficients across the six instruments are not consistent. Due to the low average reliability of the AAIS, a homogeneity test excluding the 16 reliability estimates from the AAIS was performed. When these reliability coefficients were removed from the homogeneity test, the remaining reliability coefficients are statistically homogenous ($\chi^2_{(23)} = 19.21$; $p = .68$). This indicates that the reliability coefficients are consistent across the PESQ-PS, RAPI, DUSI-SUS, POSIT-ADS, and the YAAPST.

The chi-square test of homogeneity is a fixed effects test, which lacks generalizability to current and future studies utilizing these screening measures. To determine whether the AAIS will tend to generate less reliable scores in future studies, we analyzed the mean difference between the AAIS compared to the mean of the five other screening measures combined. The unweighted mean of the five homogenous screening measures was higher ($M = 0.90$) than the unweighted mean of the AAIS ($M = 0.78$) and this difference was statistically significant ($F[1, 38] = 17.38$; $p = .0002$; $\eta^2 = 0.31$). Moreover, this difference was consistent when controlling for the percentage of the sample that is Caucasian, the percentage of the sample that is male, if the sample was from a clinical population, the average age of the sample, and the variance within the sample. This suggests that

the actual test administered is a moderator of reliability, with the AAIS, on average, generating statistically less reliable scores than the other five instruments.

Correlates of Reliability

With the administered test seen as contributor in the variance of obtained reliability coefficients, the next analytic step was to explore other factors that may predict variability in the observed reliability coefficients. This was done by computing the Pearson correlation coefficients between the identified predictor variable and score reliabilities, both within and across the screening measures (Table 3). While sample size restricts absolute interpretability of these estimates (*Ns* range from 30 to 38), a marginally statistically significant ($p < .10$) relationship provides some evidence for the tested moderator. Using this criterion, we see that all six instruments generate more reliable scores in clinical settings ($M = 0.89$) as opposed to nonclinical settings ($M = 0.82$; r effect size (38) $= 0.29$, $p < .10$). Further, there are several within instrument correlations that are statistically significant.

DISCUSSION

To contextualize these results, it is helpful to evaluate them relative to proposed minimum levels of acceptable reliability. While only "rules-of-thumb," conventionally used standards suggest a minimum score reliability cutoff value of 0.70 for the early stage of measure development, 0.80 for basic research purposes, and 0.90 when important clinical decisions are being made (Nunnally & Bernstein, 1994). The PESQ-PS was the only instrument to generate both weighted and unweighted mean and median score reliability estimates exceeding 0.90. Additionally, PESQ-PS score reliability estimates generated the smallest range (0.06) with low values at 0.90. The 95% CI indicates that 95 out of the next hypothetical studies using the PESQ-PS should observe internal consistency reliability estimates between 0.90 and 0.96. Indeed, the PESQ-PS was the only scale in the present study that obtained a lower bound reliability estimate at or above 0.90. Taken together, these results suggest that the PESQ-PS tends to generate adequately reliable scores for both research and clinical purposes even after accounting for variation in these estimates.

Both the RAPI and the substance use scale of the DUSI achieved comparable weighted and unweighted mean and median values that approached those of the PESQ. However, the ranges of reliability estimates for each of the former instruments indicate more variability in those estimates. Nevertheless, with low values still above the suggested 0.80 value for research purposes, and respectable 95% CIs, both the RAPI (95% CI = 0.85–0.93) and DUSI-SUS (95% CI = 0.86–0.95) tend to produce scores with adequate reliability for most research purposes and show promise for clinical use in this regard as well. The wide sample age range from which RAPI reliability estimates were drawn suggests the instrument's usefulness across adolescent groups in different developmental stages (e.g., high school and college).

Previous studies using the POSIT-ADS and YAAPST also created a pattern of statistics that allow their respective reliability estimates to be described similarly. Across samples, these instruments tend to produce scores with reliability in the mid .80 s with low range values falling below 0.80. These, along with their 95% CIs (POSIT-ADS, 0.79–0.94; YAPPST, 0.70–0.85), suggest that while these tools tend to generate score with adequate reliability for most research purposes, they cannot always be counted on to do so. This may be particularly true for the YAAPST.

The descriptive and inferential statistics indicate that the AAIS may produce lower estimates of reliability. In terms of mean and median score reliability estimates, the AAIS achieved the lowest values due to the fact that almost 60% ($n = 7$) of reported reliability estimates fell below the suggested value of 0.80 for research purposes. The low range value of 0.51 along with the 95% CI (0.67–0.85) suggests that researchers should take precautions to minimize measurement error when using the AAIS and clinicians may wish to consider alternate tools when screening for adolescent alcohol problems. Both fixed and random model homogeneity tests indicate that the AAIS generates less reliable scores than the other five screening measures and this mean difference in reliability was not due to differing sample characteristics between the AAIS and the other five measures.

Moderators of Reliability: A Guide for Future Research

Insufficient score reliability and sample characteristics reporting tempers our enthusiasm for the results regarding the moderators of score reliability within the adolescent alcohol screening literature.

Nevertheless, as this study is the first to evaluate potential predictors of score reliability estimates within this literature, results will inform future research. The remainder of the discussion will provide interpretations of the correlations between sample-specific factors and the reliability coefficients within the six screening measures. These results cannot be taken literally but instead be used to guide researchers evaluating the future performance of the instruments under evaluation.

When is Sample Type a Moderator?

The only significant moderator across all six measures was sample type; studies administering the screening test in clinical settings (e.g., emergency room, outpatient mental health clinic) have higher estimates of reliability when compared to studies administering them in nonclinical settings. This marginally significant relation is consistent with RG studies of adult alcohol screening tools (Shields & Caruso, 2004; Shields, Howell, Potter, & Weiss, 2007). This result supports the use of this variable as a candidate for future reliability studies among adolescent alcohol screening tools. This is particularly true for the RAPI, as it was developed for use in both clinical and nonclinical settings (White & Labouvie, 1989), yet reliability is only available for a single clinical sample (Runchkin, Koposov, Eisemann, & Hagglof, 2002).

When is Ethnicity a Moderator?

Across the 40 studies, 30 (75%) reported ethnicity information and no statistically significant relationship was found between percentage of sample that was Caucasian and the reliability coefficients (r (28) = 0.18, $p = .35$). However, within two of the measures (the RAPI and the AAIS), the relationships between score reliability and sample ethnicity were marginally significant (the YAAPST was marginally significant, but the number of samples was too low, $N = 3$). For AAIS the direction of the relationship was positive and this suggests that score reliability tends to increase as the proportion of the sample that is Caucasian increases. Indeed, among AAIS-administered samples made up of at least 80% Caucasians ($n = 7$), the average reliability estimate was .86, and the single study that administered the AAIS to an all–African American sample yielded a reliability estimate of 0.62 (Rodney, Mupier, & Crafter,

1997). The relationship between RAPI score reliability and sample ethnicity indicates that RAPI score reliability tends to increase as the proportion of Caucasians in the sample decreases. This is attributed primarily to the highest observed RAPI score reliability estimate ($a = 0.91$) coming from the most ethnically heterogeneous sample (Levy & Earleywine, 2003). Noteworthy is the overall lack of ethnic variation observed in the samples contributing to our aggregate RAPI reliability estimate. Aside from the DUSI literature, this finding generalizes across measures and is indicative of a relative lack of reliability data among non-Caucasian samples in the adolescent alcohol screening literature. Because previous findings show ethnicity as an important determinant of the performance characteristics, including score reliability, of adult alcohol screening tools (Cherpitel, 1998; Cherpitel & Clark, 1995; Shields & Caruso, 2004), it would be beneficial for those using these instruments among minority populations if more reliability information among ethnic minorities were available.

When is Age a Moderator?

Though the correlation across all measures was not statistically significant, there was a statistically significant and negative relationship among POSIT-ADS score reliability estimates and sample age. While the magnitude of effects produced by age on score reliability is difficult to interpret (and inconsistent), this statistically significant relationship is consistent with an RG study of an adult alcohol screening tool, the Alcohol Use Disorders Identification Test (Shields & Caruso, 2003).

Other Moderators?

A variety of factors could influence score reliability in addition to those explored in the present study. Most notable is score variability, which can demonstrate tremendous predictive power regarding score reliability (Caruso, Witkiewitz, Belcourt-Dittloff, & Gottlieb, 2001). This is consistent with classical test theory (cf. Henson, 2001), and without its inclusion as a predictor variable, fluctuations in score reliability that are really due to changes in observed score variability may be erroneously attributed to other predictor variables. In the present analysis, only 16 (14%) studies simultaneously reported score reliability and variability (e.g., SD) information and this precluded the exploration of this relationship in all but two of the identified

TABLE 4. Descriptive Statistics for Predictor Variables Across Instruments and an Index of their Relationship with Score Reliability

Variable	*K*/*N*	*M*	*Mdn*	*SD*	Low	High	Range	*r**	*p*
Predictor variables									
Sample type									
PESQ-PS	5	0.60	1.0	–	0.00	1.0	1.0	.84	.07
RAPI	8	0.13	0.00	–	0.00	1.0	1.0	− .08	.86
DUSI-SUS	4	0.50	0.50	–	0.00	1.0	1.0	− .54	.27
POSIT-ADS	5	1.0	1.0	–	1.0	1.0	.00	–	–
YAAPST	4	0.00	0.00	–	0.00	0.00	.00	–	–
AAIS	12	0.17	0.00	–	0.00	1.0	1.0	0.34	.28
Statistics across all Measures	**38**	**0.34**	**0.00**	**–**	**0.00**	**1.0**	**1.0**	**0.29**	**.08**
Proportion male									
PESQ-PS	5	0.80	1.0	0.45	0.00	1.0	1.0	− 0.42	.48
RAPI	6	0.51	0.49	0.26	0.24	1.0	0.76	− 0.26	.63
DUSI-SUS	6	0.58	0.74	0.49	0.00	1.0	1.0	− 0.11	.84
POSIT-ADS	5	0.58	0.61	0.15	0.31	0.68		0.72	.17
YAAPST	3	0.35	0.35	0.01	0.34	0.35	0.01	− 0.78	.43
AAIS	10	0.59	0.54	0.20	0.33	1.0	0.67	0.22	.55
Statistics across all Measures	**35**	**0.58**	**0.54**	**0.30**	**0.00**	**1.0**	**1.0**	**0.12**	**.48**
Average age (years)									
PESQ-PS	5	15.9	16.3	2.1	13.9	18.8	4.9	− 0.38	.53
RAPI	4	18.1	18.1	2.5	15.6	20.7	5.1	0.68	.32
DUSI-SUS	6	15.6	15.6	0.32	15.2	16.0	0.81	− 0.15	.78
POSIT-ADS	4	15.5	15.4	0.36	15.2	15.9	0.75	− 0.97	.03
YAAPST	3	18.6	18.8	0.41	18.2	19.0	0.79	− 0.37	.76
AAIS	10	14.9	15.6	1.4	12.0	16.4	4.4	0.21	.56
Statistics across all Measures	**33**	**16.00**	**15.61**	**1.88**	**12.00**	**20.70**	**8.70**	**0.16**	**.40**
Proportion white									
PESQ-PS	5	0.76	0.80	0.15	0.60	0.94	0.34	− 0.03	.97
RAPI	6	0.91	0.94	0.11	0.70	1.0	0.30	− 0.76	.08
DUSI-SUS	5	0.33	0.00	0.44	0.00	0.81	.81	− .54	.35
POSIT-ADS	4	0.53	0.59	0.26	0.16	0.77	.61	.71	.29
YAAPST	3	0.83	0.85	0.06	0.77	0.88	.11	.99	.07
AAIS	7	0.81	0.97	0.36	0.00	1.0	1.0	.80	.03
Statistics across all Measures	**30**	**0.71**	**0.82**	**0.33**	**0.00**	**1.0**	**1.0**	**.18**	**.35**
SD of scores									
POSIT-ADS	4	2.7	2.8	0.87	1.80	3.48	1.68	.82	.19
AAIS	7	9.76	9.33	2.66	6.60	14.25	7.65	.33	.47

Note. K = the total number of independent samples.

Note. Sample type was coded as clinical (1) and non-clinical (0).

*Pearson *r* refers to the relationship between the unweighted reliability estimates and identified predictor variables.

**Note: insufficient reporting of SD precluded the individual evaluation of the PESQ-PS, RAPI, DUSI-SUS, and YAAPST.

instruments. In addition to score variability, adolescent alcohol screening researchers are encouraged to evaluate those variables that will have theoretical and practical usefulness for others as they make instrument selection decisions. For example, it would be of considerable importance to know the extent that mode of administration (e.g., face-to-face interview vs. group administration) and/or the time frame of the assessment period (e.g., past 30 days vs. past year) affect the reliability of scores gathered by a particular tool.

CONCLUSIONS

There are clear recommendations to consider reliability information when selecting and using adolescent alcohol screening measures (SAMHSA, 1999). Yet, clinicians and researchers wishing to adhere to this sound assessment advice may have difficulty because available reliability data is thinly spread over a large body of peer-reviewed literature. Results from this meta-analysis address this issue by providing a central repository of reliability information that both care providers and behavioral scientists can use when facing important measurement selection decisions. Additionally, results can inform future psychometric and reliability studies across adolescent alcohol screening tools. Lastly, results suggest that the reliability reporting practices within the adolescent alcohol screening literature could be improved. The recommendation of Wilkinson and the American Psychological Association/Task Force on Statistical Inference (1999) that authors "provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric" (p. 596) is particularly relevant in this regard.

REFERENCES

Articles marked with an * were included in the meta-analysis.

American Academy of Pediatrics, Committee on Substance Abuse. (1993). Role of the pediatrician in prevention and management of substance abuse. *Pediatrics*, *91*, 1010–1013.

* Aytaclar, S., Erkiran, M., Kirisci, L., & Tarter, R. (2003). Substance abuse and associated psychosocial risk factors among Turkish male adolescents. *Addictive Behaviors*, *28*, 1419–1429.

Borsari, B. E., & Carey, K. B. (2000). Effects of brief motivational interventions for college student drinkers. *Journal of Consulting and Clinical Psychology*, *68*, 728–733.

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 & Psychological Measurement*, *61*, 675–689.

Charter, R. A. (1996). Revisiting the standard errors of measurement, estimate, and prediction and their application to test scores. *Perceptual and Motor Skills*, *82*, 1139–1144.

Charter, R. A. (1997). Effect of measurement error on tests of statistical significance. *Journal of Clinical & Experimental Neuropsychology*, *19*, 458–462.

Charter, R. A., & Feldt, L. S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical & Experimental Neuropsychology*, *23*, 530–537.

Charter, R. A., & Feldt, L. S. (2002). The importance of reliability as it relates to true score confidence intervals. *Measurement & Evaluation in Counseling & Development*, *35*, 104–112.

Chassin, L., Tetzloff, C., & Hershey, N. (1985) Self-image and social-image factors in * adolescent alcohol use. *Journal of Studies on Alcohol*, *49*, 39–47.

Cherpitel, C. J. (1998). Differences in performance of screening instruments for problem drinking among Blacks, Whites, harmful drinking and alcohol dependence in the emergency room. *Alcoholism: Clinical & Experimental Research*, *19*, 628–634.

Cherpitel, C. J., & Clark, W. B. (1995). Ethnic differences in performance of screening instruments for identifying problem behaviors among adolescents at risk for substance use. *Journal of Child & Adolescent Substance Abuse*, *11*, 17–36.

Clark, D. B. (2004). The natural history of adolescent alcohol use disorders. *Addiction*, *99*, 5–22.

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

Danseco, E. R., & Marques, P. R. (2002). Development and validation of a POSIT-Short Form: Screening for and Hispanics in an emergency room population. *Jour- * nal of Studies on Alcohol*, *59*, 420–426.

Deas, D., Riggs, P., Langenbucher, J., Goldman, M., & Brown, S. (2000). Adolescents are not adults: Developmental considerations in alcohol users. *Alcoholism: Clinical & Experimental Research*, *24*, 232–237.

De Micheli, D., & Formigoni, M. L. (2002). Psychometric properties of the Brazilian version of the Drug Use Screening Inventory. *Alcoholism: Clinical & Experimental * Research*, *26*, 1523–1528.

Downs, W. R. (1987). A panel study of normative structure, adolescent alcohol use * and peer alcohol use. *Journal of Studies on Alcohol*, *48*, 167–175.

Elster, A., & Kuznets, N. (1994). *American Medical Association's guidelines for adolescent preventive services (GAPS)*. Baltimore, MD: Williams & Wilkins.

Gilksman, L., & Smythe, P. C. (1982). Adolescent involvement with alcohol: A cross- * sectional study. *Journal of Studies on Alcohol*, *43*, 370–379.

Harrell, A. V., & Wirtz, P. W. (1989). Screening for adolescent problem drinking: Validation of a multidimensional instrument for case identification. *Psychological Assessment*, *1*, 61–63.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement & Evaluation in Counseling & Development*, *34*, 177–189.

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement & Evaluation in Counseling & Development*, *35*, 113–127.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational & Psychological Measurement*, *60*, 523–531.

Hurlbut, S. C., & Sher, K. J. (1992). Assessing alcohol problems in college students. *Journal of American College Health*, *41*, 49–58.

Johnston, L. D., O'Malley, P. M., & Bachman, J. G. (2003). *Monitoring the future national survey results on drug use, 1975–2002: Vol. 1. Secondary school students* (NIH Publication No. 03–5375). Bethesda, MD: National Institute on Drug Abuse.

Knight, J. R., Goodman, E., Pulerwitz, T., & DuRant, R. H. (2001). Reliability of the Problem Oriented Screening Instrument for Teenagers (POSIT) in adolescent
* medical practice. *Journal of Adolescent Health*, *29*, 125–130.

Koposov, R. A., Ruchkin, V. V., Eisemann, M., & Sidorov, P. (2002). Alcohol use in adolescents from northern Russia: The role of the social context. *Alcohol Abuse &*
* *Alcoholism*, *37*, 297–303.

Latimer, W. W., Winters, K. C., & Stinchfield, R. D. (1997). Screening for drug abuse among adolescents in clinical and correctional settings using the Problem-Oriented Screening Instrument for Teenagers. *American Journal of Drug & Alcohol*
* *Abuse*, *23*, 79–98.

* Levy, B., & Earleywine, M. Reinforcement expectancies for studying predict drinking problems among college students: Approaching drinking from an expectancies choice perspective. *Addictive Behaviors*, *28*, 551–559.

Lewis, M. A., & Neighbors, C. (2004). Gender-specific misperceptions of college
* student drinking norms. *Psychology of Addictive Behaviors*, *18*, 334–339.

Martin, C. S., & Winters, K. C. (1998). Diagnosis and assessment of alcohol use disorders among adolescents. *Alcohol Health & Research World*, *22*, 95–105.

Mayer, J., & Filstead, W. J. (1979). The Adolescent Alcohol Involvement Scale: An instrument for measuring adolescents' use and misuse of alcohol. *Journal of Studies on Alcohol*, *40*, 291–300.

Mayer, J. E., & Ligman, J. D. (1989). Personality characteristics of adolescent marijuana users. *Adolescence*, *24*, 965–976.

McLaney, M. A., & Boca, F. D. (1994). A validation study of the Problem-Oriented Screening Instrument for Teenagers (POSIT). *Journal of Mental Health*, *3*,
* 363–376.

McNally, A. M., & Palfai, T. P. (2003). Brief group alcohol interventions with college students: Examining motivational components. *Journal of Drug Education*,
* *33*, 159–176.

Moberg, D. P. (1983). Identifying adolescents with alcohol problems: A field test of the Adolescent Alcohol Involvement Scale. *Journal of Studies on Alcohol*, *44*, 701–721.

Moss, H. B., Boniacatto, S., Kirisci, L., Giradelli, A. M., & Murrelle, L. (1998). Substance abuse and associated psychosocial problems among Argentine adolescents: Sex, heterogeneity and familial transmission. *Drug & Alcohol Dependence*, *52*, 221–230.

National Institute on Alcohol Abuse and Alcoholism. (1995). *Assessing alcohol problems: A guide for clinicians and researchers (NIH Publication No. 95–3745)*. Bethesda, MD: Author.

Neal, D. J., & Carey, K. B. (2004). Developing discrepancy within self-regulation theory: Use of personalized normative feedback and personal strivings with heavy-drinking college students. *Addictive Behaviors*, *29*, 281–297.

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

Palfai, T., & Wood, M. D. (2001). Positive alcohol expectancies and drinking behavior: The influence of expectancy strength and memory accessibility. *Psychology of Addictive Behaviors*, *15*, 60–67.

Pedersen, W., & Skrondal, A. (1999). Ecstasy and new patterns of drug use: A normal population study. *Addiction*, *94*, 1695–1706.

Petchers, M. K., & Singer, M. I. (1987). Perceived-Benefit-of-Drinking Scale: Approach to screening for adolescent alcohol abuse. *Journal of Pediatrics*, *110*, 977–981.

Putnins, A. L. (1992). The adolescent alcohol involvement scale: Some findings with young offenders. *Drug & Alcohol Review*, *11*, 253–258.

Rahdert, E. R. (1991). *The adolescent assessment/referral system manual*. Washington, DC: U.S. Department of Health and Human Services Alcohol, Drug Abuse, and Mental Health Administration.

Rodney, H. E., Mupier, R., & Crafter, B. (1997). Predictors of alcohol drinking among African-American adolescents. *Journal of Negro Education*, *65*, 434–444.

Ruchkin, V. V., Koposov, R. A., Eisemann, M., & Hagglof, B. (2002). Alcohol use in delinquent adolescents from northern Russia: The role of personality, parental rearing and family history of alcohol abuse. *Personality & Individual Differences*, *32*, 1139–1148.

Santisteban, D. A., Tejeda, M., Donimicis, C., & Szapocznik, J. (1999). An efficient tool for screening maladaptive family functioning in adolescent drug abusers: The problem oriented screening instrument for teenagers. *American Journal of Drug & Alcohol Abuse*, *23*, 197–206.

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

Shields, A. L., & Caruso, J. C. (2004). A reliability induction and reliability generalization study of the CAGE questionnaire. *Educational & Psychological Measurement*, *64*, 254–270.

* 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 & Misuse*, *42*, 1783–1800.

Substance Abuse and Mental Health Services Administration. (1999). *Screening and assessing adolescents for substance use disorders, treatment improvement protocol (TIP) series 31*. Rockville, MD: U.S. Department of Health and Human Resources, Public Health Service, SAMHSA, Center for Substance Abuse Treatment.

Tarter, R. E. (1990). Evaluation and treatment of adolescent substance abuse: A decision tree method. *American Journal of Drug & Alcohol Abuse*, *16*, 1–46.

Tarter, R. E., Mezzich, A. C., Kirisci, L., & Kaczynksi, N. (1994). Reliability of the drug use screening inventory among adolescent alcoholics. *Journal of Child & Adolescent Substance Abuse*, *3*, 25–36.

* Thombs, D. L., & Beck, K. H. (1994). The social context of four adolescent drinking patterns. *Health Education Research*, *9*, 13–22.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational & Psychological Measurement*, *60*, 174–195.

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

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational & Psychological Measurement*, *60*, 509–522.

White, H. R., & Labouvie, E. W. (1989). Towards the assessment of adolescent problem drinking. *Journal of Studies on Alcohol*, *50*, 30–37.

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

Winters, K. C. (1992). Development of an adolescent alcohol and other drug abuse screening scale: Personal experience screening questionnaire. *Addictive Behaviors*, *17*, 479–490.

* Winters, K. C., Remafedi, G., & Chan, B. Y. (1996). Assessing drug abuse among gay-bisexual young men. *Psychology of Addictive Behaviors*, *10*, 228–236.

* Winters, K. C., Weller, C. L., & Meland, J. A. (1993). Extent of drug abuse among juvenile offenders. *Journal of Drug Issues*, *23*, 515–524.

* Wood, M. D., Read, J. P., Palfai, T. P., & Stevenson, J. F. (2001). Social influence processes and college student drinking: The mediational role of alcohol outcome expectancies. *Journal of Studies on Alcohol*, *62*, 32–43.