Psychometric properties of alcohol screening tests in the emergency department in Argentina, Mexico and the United States

Mariana Cremonte, Rubén Daniel Ledesma, Cheryl J. Cherpitel, Guilherme Borges

PII: S0306-4603(10)00100-0
DOI: doi: 10.1016/j.addbeh.2010.03.021
Reference: AB 3242

To appear in: *Addictive Behaviors*

Please cite this article as: Cremonte, M., Ledesma, R.D., Cherpitel, C.J. & Borges, G., Psychometric properties of alcohol screening tests in the emergency department in Argentina, Mexico and the United States, *Addictive Behaviors* (2010), doi: 10.1016/j.addbeh.2010.03.021

PSYCHOMETRIC PROPERTIES OF ALCOHOL SCREENING TESTS IN THE EMERGENCY DEPARTMENT IN ARGENTINA, MEXICO AND THE UNITED STATES.

Mariana Cremonte [a] (corresponding author)

Rubén Daniel Ledesma [b]

Cheryl J. Cherpitel [c]

Guilherme Borges [d]

[a] Consejo Nacional de Investigaciones Científicas y Técnicas

Facultad de Psicología Universidad Nacional de Mar del Plata

Centro Interdiscipliario de Investigaciones en Psicología Matemática y Experimental

Funes 3250 Cuerpo V Nivel III

(7600) Mar del Plata

Argentina

Ph/fax (54 223) 475-9993

e-mail: mcremont@mdp.edu.ar

[b] Facultad de Psicología Universidad Nacional de Mar del Plata

Consejo Nacional de Investigaciones Científicas y Técnicas

Funes 3250 Cuerpo V Nivel III

(7600) Mar del Plata

Argentina

e-mail: rdledesma@gmail.com

[c] Public Health Institute

Alcohol Research Group

6475 Christie Avenue suite 400

Emeryville, California 94608-1010

United States of America

e-mail: ccherpitel@arg.org

[d] Universidad Autónoma Metropolitana

Instituto Nacional de Psiquiatría

Calzada Mexico Hochimilco No.101,

Tlalpan, Mexico DF

Mexico

e-mail: guibor@imp.edu.mx

ABSTRACT

The objective of this article is to report psychometric characteristics of the AUDIT, CAGE, RAPS4, and TWEAK and to compare them across three countries: Argentina, Mexico, and the United States which used a similar protocol and methodology. Probability samples of patients 18 years and older were drawn from emergency departments in Mar del Plata, Argentina (n=780), Pachuca, Mexico (n=1624) and Santa Clara, U.S. (n=1220). Concurrent validity was assessed by comparing their performance against a diagnosis of alcohol dependence (DSM-IV) obtained through the Composite International Diagnostic Interview, and for the briefer measures, also by their correlation with the AUDIT. The internal consistency of the CAGE, RAPS4, and TWEAK scores was estimated by the KR-20 formula and by Cronbach's Alpha for the

AUDIT. Corrected item-total correlation and D-values were used as item discrimination measures.

In Argentina and Mexico the AUDIT and the RAPS4 showed the highest validity. Reliability of all instruments was higher in the US than in Argentina or Mexico. In all three countries, reliability of the TWEAK was lowest, while the AUDIT was highest. With a few exceptions, all items showed good discrimination powers.

Keywords

Alcohol-screening-psychometric-Emergency Department-Argentina-Mexico

1. INTRODUCTION

A number of self-report measures to screen alcohol use disorders have been developed for use in clinical settings, and have generally been found to demonstrate better sensitivity than physiological measures or laboratory testing (Aertgeerts, Buntinx, Ansoms, & Fevery, 2001). Most of these screeners are brief and designed to be administered by a lay interviewer. Among those most commonly used are the Alcohol Use Disorders Identification Test (AUDIT), the CAGE, the Rapid Alcohol Problems Screen (RAPS4), and the TWEAK. Attention to the performance of these brief screeners for alcohol use disorders cannot be overemphasized given the importance of a valid and timely diagnosis and referral to treatment. Furthermore, their utility in clinical and hospital settings is especially important since it has been shown that admission to the Emergency Department (ED) presents a unique opportunity for a reduction in drinking and/or acceptance of referral to treatment (Gentilello, Ebel, Wickizer, Salkever, & Rivara, 2005). Additionally, having sound screening instruments is fundamental not only for actual or prospective screening and brief intervention, but for surveillance and research as well.

The main goal of this article is to present evidence of the validity, reliability and item-level statistics of the AUDIT, CAGE, RAPS4 and TWEAK (described below) in ED settings in Argentina, Mexico and the US, and to compare these measures, under similar conditions, across the three countries.

The AUDIT has been the most extensively researched of these instruments (reviewed in Reinert & Allen, 2002; Reinert & Allen, 2007). A vast body of research was conducted which includes an examination of various psychometric properties such as temporal stability, internal consistency, and construct, concurrent and predictive validity (Cremonte & Cherpitel, 2008; Medina-Mora, Carreno, & De la Fuente, 1998; Rubio Valladolid, Bermejo Vicedo, Caballero Sanchez-Serrano, & Santo-Domingo Carrasco, 1998; Rumpf, Hapke, Meyer, & John, 2002). This evidence supports the reliability and validity of the AUDIT as a screening tool. Many non-English versions also had been successfully tested and validated in a wide array of countries and cultural settings (Cherpitel, Ye, Moskalewicz, & Swiatkiewicz, 2005; Cremonte & Cherpitel, 2008; Gache et

al., 2005; Kim, Gulick, Nam, & Kim, 2008; Lima et al., 2005; Rumpf et al., 2002). Despite this abundant literature, in a recent review Reinert and Allen (2007) concluded the need for additional research to further study the psychometric properties of non-English versions of the AUDIT. This suggests an even greater need for similar additional research on other brief screening instruments which have received far less attention The lack of motivation to collect data on the psychometric performance of these other instruments comes from their relative simplicity, in comparison with the more complex, multidimensional, structure of the AUDIT. While the AUDIT has three subscales containing ten items on a five point answer scale for the first eight items and a three point scale for items nine and ten, the CAGE and RAPS4 have four dichotomous items each and the TWEAK has an additional non-dichotomous item. Most reports evaluating the performance of the CAGE, TWEAK and RAPS4 have been limited to their concurrent validity, primarily through sensitivity and specificity based on standard diagnostic criteria. Studies evaluating other psychometric properties, such as reliability, or other forms of validity, are scarce. In a systematic review of studies on the CAGE (Dhalla & Kopec, 2007) only three provided evidence of its reliability. Moreover, as with other behavioral and psychological measures, reliability and validity can vary with age, gender, and ethnic background (Cherpitel & Borges, 2000; Knight, Goodman, Pulerwitz, & DuRant, 2000). Additionally, studies reporting item level statistics, such as item discrimination power, are extremely rare. Given that the CAGE is the oldest and most widely used of these simpler instruments, even fewer studies report psychometric properties (other than concurrent validity) of the RAPS4 and TWEAK. The TWEAK has been most extensively tested among pregnant women and found, with a few exceptions (Bush et al., 2003), to perform reasonably well (Moraes, Viellas, & Reichenheim, 2005). The RAPS4 on the other hand, has been tested primarily in large samples of ED patients in several countries (Cherpitel & Borges, 2000; Cherpitel, Ye, Moskalewicz et al., 2005; Cremonte & Cherpitel, 2008) against ICD-10 (World Health Organization, 1992) and DSM-IV (American Psychiatric Association, 1994) criteria for alcohol dependence and harmful drinking/abuse, as well as against tolerance across 13 countries (Cherpitel, Ye, Bond et al., 2005). No studies, however, have addressed other psychometric properties other than sensitivity and specificity of the RAPS4; nor are studies reported on the

reliability of the TWEAK in ED patients. To the best of our knowledge the present study is the first that examines a number of psychometric properties of four of the commonly used screeners, in three diverse countries and under similar conditions. Thus, the main objective of this article is to present new evidence of validity and reliability of the AUDIT, CAGE, RAPS4, and TWEAK, and to compare the performance of these instruments across three countries which exhibit diverse drinking styles: Argentina, Mexico, and the U.S. Argentina has an European integrated style of drinking, with low abstention rates and high per capita consumption, while Mexico typifies the fiesta pattern of drinking, with a higher abstention rate and a higher rate of infrequent but heavy drinking, while the U.S. demonstrates a drinking style somewhere between these two (World Health Organization, 2004).

2. METHODS

2.1. Participants

Probability samples of patients were drawn from EDs in each country: Argentina, Mexico and the U.S. At each site a sample of patients 18 years and older was obtained from ED admissions reflecting consecutive arrival to the ED. Each sample reflected an equal representation of each shift for each day of the week during the study period. Patients who arrived at the ED too severely ill or injured to be interviewed were followed into the hospital and interviewed once their condition had stabilized. In Argentina the sample was collected from the largest ED of the city of Mar del Plata in the state of Buenos Aires (n=780); in Mexico from three EDs in Pachuca in the state of Hidalgo (n=1624); and in the US from an ED in Santa Clara, California (n=1220). Completion rates were 92 % in Argentina, 93% in Mexico and 73% in the US. These samples are part of the Emergency Room Collaborative Alcohol Analysis Project (ERCAAP) and additional information about methods and data collection procedures can be found elsewhere (Cherpitel, Ye, & Bond, 2004). The data analyzed here include only those patients who reported having consumed at least one drink during the last twelve months (current drinkers): 85% in Argentina (n= 662), 34% in Mexico (n= 559), and 72 % in the U.S. (n=884). Data on socio-demographic and drinking characteristics of the sample in

each country is presented in Table 1.

[Table 1 about here]

2.2. Instruments

Eligible patients were asked to provide informed consent as soon possible after arriving in the ED, and were subsequently interviewed by a cadre of trained field workers who administered a similar questionnaire developed by Cherpitel (1989) in all three countries. The questionnaire included, among other items, the Alcohol Section of the Composite International Diagnostic Interview (CIDI) Core (World Health Organization, 1993) to obtain a DSM-IV diagnosis for alcohol dependence for the last 12 months, and items comprising the AUDIT, CAGE, RAPS4, and TWEAK. The AUDIT (Saunders, Aasland, Amundsen, & Grant, 1993) was developed by the World Health Organization (WHO) with the primary purpose of identifying harmful and hazardous drinking in primary care settings. It is a ten-item measure comprised of three subscales evaluating recent alcohol use, alcohol dependence symptoms, and alcohol-related problems (See Appendix 1). The CAGE (Ewing, 1984), the oldest and most widely used of the brief screening instruments, is a four item instrument, with the advantage of its brevity and simplicity. Its name is an acronym based on the following four questions: 1) Have you ever felt you should cut down on your drinking? 2) Have people annoyed you about your drinking? 3) Have you ever felt bad or guilty about your drinking? 4) Have you ever had a drink first thing in the morning to steady your nerves or get rid of a hangover (eye-opener)? The TWEAK is a short instrument developed to screen pregnant women in clinical settings (Chan, Pristach, Welte, & Russell, 1993). Its name is also an acronym based on five items: 1) How many drinks does it take to make you feel high? 2) Have close friends or relatives worried or complained about your drinking in the past year? 3) Do you sometimes take a drink in the morning when you first get up? 4) Has a friend or family member ever told you about things you said or did while you were drinking

that you could not remember? 5) Do you sometimes feel the need to cut down on your drinking? The first two items have a weighted score. The Rapid Alcohol Problems Screen (RAPS4) is the more recently developed of these instruments (Cherpitel, 1995), based on the five best-performing items from the AUDIT, CAGE , Brief MAST (Pokorny, Miller, & Kaplan, 1972), and TWEAK and subsequently refined into a four-item instrument (Cherpitel, 2000): 1) During the last year have you had a feeling of guilt or remorse after drinking? (**R**emorse) 2) During the last year has a friend or family member ever told you about things you said or did while you were drinking that you could not remember? (**A**mnesia, also called blackouts) 3) During the last year have you failed to do what was normally expected from you because of drinking? (**P**erform) 4) Do you sometimes take a drink in the morning when you first get up? (**S**tarter, also called eye-opener).

All screening items were also framed to inquire about the last twelve months (although the CAGE is typically used on a life-time basis). In Argentina and Mexico the instruments were independently translated into Spanish and back translated into English in order to obtain locally adapted versions. Sensitivity and specificity of the Argentinean version of the AUDIT, CAGE, RAPS4 and TWEAK has been reported elsewhere (Cremonte & Cherpitel, 2008) as has been the sensitivity and specificity of the Mexican version (Cherpitel & Borges, 2000).

2.3. Data Analysis

The cut-point at which a screen was considered positive was: AUDIT (a weighted score of 8), CAGE (1), RAPS4 (1), and TWEAK (a weighted score of 2). Sensitivity and specificity for the AUDIT, CAGE, RAPS4, and TWEAK were estimated for each country against a standard diagnosis of alcohol dependence according to DSM-IV criteria obtained from the CIDI core (World Health Organization, 1993). Sensitivity and specificity are psychometric properties of instruments that constitute measures of concurrent validity. Sensitivity refers to the capacity of a given instrument to correctly identify those positive on the criterion

from among all those who are positive, while specificity refers to the capacity of an instrument to correctly identify those negative on the criterion among all those who are negative. Concurrent validity was additionally assessed for the briefer screeners (CAGE, RAPS4, and TWEAK) by estimating their Pearson correlation (one-tailed) with the AUDIT total scores (as the longest scale). Descriptive statistics (mean and S.D.) of the AUDIT, CAGE, RAPS4, and TWEAK's total scores were computed for each country. As a measure of reliability the internal consistency of the AUDIT was estimated by Cronbach's Alpha and of the CAGE, RAPS4, and TWEAK by the KR-20 formula (Kuder, & Richardson, 1937). The prevalence of each item for each screener in each country is also reported.

The corrected item-total correlation was used as an item discrimination measure for the AUDIT, and corrected point-biserial correlation and D-value were used for the shorter instruments. The D-value was calculated by subtracting the proportion of positive answers to an item among those negative on the criterion (alcohol dependence) from the proportion of positive answers among those positive on the criterion. While the corrected-item total correlation indicates the capacity of an item to discriminate high from low scorers (considering the measured trait as a continuous variable), the D-value measures the capacity of an item to discriminate between discrete states (those positive on the criterion from those negative).

The Statistical Package for the Social Sciences (SPSS) for Windows version 11.5 was used for data processing and analysis. Complementary psychometric analyses were performed using ViSta-CITA, a software for classic item and test analysis (Ledesma & Molina, 2009). Due to missing data the effective sample sizes vary depending on the instrument analyzed.

3. RESULTS

Sensitivity and specificity values, correlation coefficients between the AUDIT and the brief measures, descriptive statistics for the total scores, internal consistency, the proportion of indicator presence, and item discrimination measures for the AUDIT, CAGE, RAPS4 and TWEAK are presented in Tables 2, 3, 4 and 5,

respectively.

[Table 2 about here]

[Table 3 about here]

[Table 4 about here]

[Table 5 about here]

3.1. Psychometric properties of AUDIT in 3 countries

The AUDIT showed high sensitivity and specificity in all the three countries. Internal consistency for the AUDIT was higher in the U.S., although good (above .80) in all three countries. Subscale 1 had a somewhat lower (below .70) reliability in Argentina and Mexico, as did subscale 3 in Mexico. In Argentina, scale reliability was increased by eliminating item 1. Except for the items belonging to subscale 1 (drinking patterns), means for all other items (4 through 10) were higher in the U.S. Item 1 (*frequency of drinking*) had the highest mean across the three countries and it was highest in Argentina, followed by the U.S. and Mexico. In Argentina this item had the lowest discrimination power, although discrimination power of this item was good in Mexico and the U.S and discrimination power of all other items was good across all three countries, with the most discriminating item in all countries being item 4 (*inability to stop drinking*). Overall, all items appeared to have higher discriminating power in the U.S. than in Mexico or Argentina.

3.2. Psychometric properties of CAGE in 3 countries

Sensitivity of the CAGE appeared (no formal statistical testing was done) higher in the U.S. and Mexico, and somewhat lower in Argentina, while specificity seemed higher in Argentina and lower in the other two countries (not unexpected since sensitivity and specificity are inversely correlated). Correlation of the CAGE with the AUDIT was higher in the U.S. Internal consistency was good in all three countries, especially considering that the reliability coefficient depends on the number of items in an instrument, and

the CAGE has only four. As true of the AUDIT, internal consistency was higher in the U.S. than in Argentina and Mexico. Although item 4 (*eye opener*) was the least prevalent item in the three countries it had a good (above .40) discrimination index. The best discriminating items were 1 (*Cut*) in Argentina and 3 (*Guilt*) in Mexico and the U.S. (corrected item-test correlation).

3.3. Psychometric properties of RAPS4 in 3 countries

Sensitivity of the RAPS4 was high in the three countries; specificity appeared somewhat lower in the U.S. In the three countries, the RAPS4 was the brief screener that showed the highest correlation with the AUDIT. As with the AUDIT and CAGE, the RAPS4 had good internal consistency in the three countries, but seemed higher in the U.S. Similar to the CAGE, item 4 (*starter/eye opener*) was the least prevalent, although it had an acceptable discrimination index (above .30) in all three countries. Items that showed a higher corrected correlation with total scores were 1 (*remorse)* in Argentina and 3 (*performance*) in Mexico and the U.S. Item 1 (*remorse*) had the highest D-value in all three countries.

3.4. Psychometric properties of TWEAK in 3 countries

Sensitivity of the TWEAK was high in the three countries. Specificity appeared to be lowest in Argentina (below 70%). Correlation with the AUDIT was higher in the U.S. The scale`s internal consistency was below .70 in Argentina and Mexico and .71 in the US, and increased in all three countries when item 1 (*tolerance*) was eliminated (not shown) The resulting internal consistency appeared to be higher in the U.S. and Mexico, than in Argentina). This same item also showed a poorer performance (item-test corrected correlation below .30) in México. Similar to the other instruments, all items were more likely to be endorsed in the U.S. than in Mexico or Argentina. The best discriminating items were 5 (*Cut*) in Argentina and Mexico and 2 (*Worried*) in Mexico and the U.S.

4. DISCUSION

4.1. Concurrent validity

Concurrent validity of all instruments was assessed through their sensitivity and specificity against a standard diagnosis of alcohol dependence based on the CIDI (World Health Organization, 1993) and, for the brief screeners, also by their correlation with the AUDIT. However, performance based on a standard diagnostic tool is believed to be a better indicator of validity, given that the performance of each screener is evaluated against a standard of adequacy, while inter-correlations of screening instruments might be altered given that some scales posses the same items, and depend on the reliability and validity of each measure.

In Argentina, the instruments with the highest sensitivity were the TWEAK and AUDIT, followed closely by the RAPS4. However, TWEAK's specificity was low (67%) making the AUDIT and the RAPS4 the best performing instruments, with both adequate sensitivity and specificity. In Mexico, the RAPS4 and the AUDIT performed equally well, while the TWEAK had a slightly lower sensitivity. In the U.S. the CAGE, RAPS4 and the AUDIT had high sensitivity; although the RAPS4 had a somewhat lower specificity (75%) and the CAGE even lower. The TWEAK also had high levels of sensitivity and specificity. Overall, all instruments showed similar validity in the U.S.; however, considering that for screening purposes given reasonable specificity, sensitivity is preferred over specificity, the best performing screener appeared to be the RAPS4.

Comparing the instruments' concurrent validity among the three countries, the CAGE had the poorest performance, although somewhat better in the U.S. than in Argentina and Mexico. The CAGE has typically been scored positive at a cut point of 2, rather than 1 as used in this study; however, this lower cutpoint only serves to increase sensitivity (at the sake of specificity). In both Argentina and Mexico the RAPS4 and AUDIT showed a higher validity. Higher validity of the AUDIT might be related to the fact that it was developed on international samples. Furthermore, the RAPS4 showed the highest correlation with the AUDIT in all three countries, rendering additional support to its higher concurrent validity. The observed correlations among the brief screeners and the AUDIT in all three countries are higher than those reported by

Kelly et al. (2002) in a U.S. sample of adolescent ED patients, possibly due to better functioning of these instruments among adults, on whom these brief screeners were originally developed.

Findings regarding the performance of the brief screeners against a standard diagnosis of current alcohol dependence are within the expected range, as is the higher validity found for the AUDIT and RAPS4 among the Non-English speaking countries (Cherpitel & Bazargan, 2003; Dhalla & Kopec, 2007; Fiellin, Reid, & O'Connor, 2000; Gache et al., 2005; Rumpf et al., 2002).

4.2. Reliability

Results presented here indicate that internal consistencies of all screeners, except the TWEAK, were above the proposed criterion of .70 as an acceptable value (Nunnally & Bernstein, 1995).Internal consistency improved for the TWEAK to at least .70 when item 1 (*tolerance*) was eliminated in Mexico and the U.S., although the resulting coefficient was still somewhat low in Argentina. Taking into account that reliability coefficients depend upon the number of items included, and the CAGE, RAPS4 and TWEAK are very short, reliabilities, overall, were satisfactory.

Internal consistencies found for the TWEAK and CAGE were similar to those previously reported (Bell, Williams, Senier, Strowman, & Amoroso, 2003; Kelly et al., 2002; Shields & Caruso, 2004), although higher than that reported for a Brazilian sample of pregnant women (Moraes et al., 2005), and possibly due to better functioning of the instruments among males and mixed samples for the CAGE (the TWEAK was originally developed for use among pregnant women). The internal consistency coefficients for the RAPS4 were good in all three countries, and to our knowledge, these are the first estimates of reliability reported for this measure.

Reliabilities of all instruments were higher in the U.S. Among the instruments, the TWEAK had the lowest estimates in all three countries, while the AUDIT had the highest. Higher reliability of the AUDIT relative to the other instruments was an expected result given that the AUDIT is the longest scale. However,

reliability of the 3 and 4 item subscales was also good. Overall, the AUDIT presented good reliability in the three countries, and this finding is consistent with the majority of literature reporting reliability in English and Non-English speaking countries. For example a Cronbach's Alpha of .82 was found for U.S. residents of Korean origin (Kim, 2008), .87 in a French speaking clinic sample (Gache et al, 2005), and .81 in a general Brazilian population (Lima et al., 2005). Moreover, our finding of a lower reliability for subscale 1 in Argentina and Mexico has been reported in a general population sample in Germany (Rumpf et al., 2002). What appeared in our findings to be small differences in the reliabilities of the AUDIT subscales and also among instruments in each country might also be partially explained by cultural differences in drinking patterns and related problems. For example, in Argentina subscale 1 which measured *drinking habits* which includes items related to *frequency of drinking* (item 1), *quantity per occasion* (item 2) and *frequency of heavy drinking* (item 3) had a somewhat low reliability. Because Argentina is primarily a wine drinking culture where low quantities of alcohol are typically consumed (Munné, 2005); these three questions would be expected to have a low inter-correlation since those with higher scores on item 1 may score low on items 2 and 3, and vice versa. Furthermore, although all means tended to be higher in the U.S. (likely due to a higher prevalence of dependence), the highest mean for item 1 was in Argentina, followed by the U.S. and Mexico, reflecting each country's patterns of drinking as noted in the introduction. Additionally, in Argentina this item had the lowest discrimination power, possibly measuring a different construct since reliability increased when this item was eliminated.

4.3. Item analysis

All items were more likely to be endorsed in the U.S. than in either Mexico or Argentina, likely due to the higher prevalence of dependence in that sample. In all three countries the item that was least likely to be endorsed was *starter/eye opener*, despite which evidenced a good discrimination power.

All items, except the first from the TWEAK in all three countries (*how many drinks can you hold*)

and the first from the AUDIT in Argentina, showed good discrimination power. The poor performance found here for the first AUDIT item in Argentina has been previously reported in a general population study in Germany (Rumpf et al., 2002). The poorer performance of item 9 found here for the Mexico and U.S. samples, has also been reported by these same authors (Rumpf et al., 2002) and by Kelly et al. (2002). This item evaluates alcohol-related injuries, and its performance is likely affected by the particular alcohol-injury relationship and perception of such a link in a culture, which has been found to be related to drinking patterns of a culture (Cherpitel, et al., 2004).

Despite what appeared to be small variations in the performance of screeners and their items, some patterns are found across the three countries. For example, item 4 (*inability to stop drinking*) was the AUDIT's most discriminating item in all three countries. This item likely targets the core of alcohol dependence (Moss, Chen, & Yi, 2008). Likewise items related to *starter/eye opener* tended to be the least reported, although showing good discrimination power in all three countries. It could be hypothesized that this item may be endorsed only by those subjects who experience physiological dependence and need to drink in the morning in order to avoid withdrawal symptoms. However, because those subjects are at the more severe end of the alcohol use disorders spectrum, they may also tend to give a positive response to other items on the scale, resulting in the good discrimination power observed for this item.

Another item that had an uniform performance across the three countries was item 1 from the TWEAK (*tolerance: how many drinks can you hold?*), and poor performance of this item has also been reported elsewhere (Kelly et al., 2002), where it was hypothesized that this item might not be well understood in their adolescent sample due to lack of experience with tolerance. However, the fact that a similarly poor performance was observed here with samples that had a relatively high prevalence of alcohol use disorders might point in a different direction. Our findings indicate this item measured a rather different construct, since internal consistency of the TWEAK increased when this item was eliminated. The multidimensionality this item introduces might be related to the complex link between alcohol use

disorders and specific patterns of drinking since it has been previously shown that quantity of drinking, per se, may not be directly related to the degree of severity of alcohol problems (Russell, Light, & Gruenewald, 2004). Noticeably, another consumption item (3 from the AUDIT) measuring *frequency of drinking 5 or more drinks per occasion* performed well.

4.4. Limitations

One limitation of findings reported here is that the order in which the instruments and the CIDI (World Health Organization, 1993) were presented might have affected their psychometric performance (Rumpf, Hapke, Meyer, & Ulrich, 2005; Steinweg & Worth, 1993), and this possible source of bias was not controlled. However, since the same questionnaire was used in all three countries, any such effect likely resulted in a similar bias across the three countries. While the administration of multiple measures along with the criterion measure under the same conditions is, by itself, an advantage (Fiellin et al., 2000),whether the ordering of instruments has an effect over their performance and the magnitude of such an effect is an area requiring further research.

Lastly, since regional variations affecting samples and psychometric results within EDs in the same country have been reported (Cherpitel, Ye, Moskalewicz et al., 2005), present findings should not be generalized to other populations, or to other regions within the same country.

4.5. Practical implications for screening in ED settings

Findings presented here indicate that in Argentina and Mexico the RAPS4 and the AUDIT had higher validity and reliability, suggesting that in these countries these may be the instruments of choice. Overall, psychometric performance of all instruments seemed more similar in the U.S. However, considering that the TWEAK demonstrated lower reliability, other screeners might be preferable, and of these, the best choices appear to be the RAPS4 and the AUDIT. Also noteworthy is that in all three countries, the RAPS4 had the highest correlation with the AUDIT, adding further support for use of the RAPS4 when a shorter, simpler

instrument is needed.

Noteworthy, the first three items of the AUDIT (comprising the consumption subscale) have been proposed as a stand-alone screener named AUDIT-C (Bush et al., 1998). Although this new screener has been found to perform relatively well (Frank et al., 2008) further studies should be conducted before attempting its use in Argentina and Mexico, given findings reported here of the poor performance of the first item in Argentina, and a somewhat low reliability of the subscale in Mexico.

Despite study limitations, findings presented here on psychometric characteristics of the most widely used screening instruments, in ED settings in Argentina, Mexico and the U.S., using a similar methodology under similar conditions suggest distinct cultural differences in instrument performance. Possible factors accounting for variability in findings are drinking practices (affecting cultures and socio-demographic groups), differences in the manifestation of related problems and disorders, prevalence of such problems, and their degree of severity (spectrum range). However, the relative impact of these factors and the mechanisms by which they account for variations and consistencies in findings across different studies and sites is not known, and this is an area in need of new empirical and, may be more importantly, theoretical developments.

References

Aertgeerts, B., Buntinx, F., Ansoms, S., & Fevery, J. (2001). Screening properties of questionnaires and laboratory tests for the detection of alcohol abuse or dependence in a general practice population. *Br J Gen Pract. , 51*(464), 206-217.

American Psychiatric Association (Ed.). (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association.

Bell, N. S., Williams, J. O., Senier, L., Strowman, S. R., & Amoroso, P. J. (2003). The reliability and validity of the self-reported drinking measures in the army's health risk appraisal survey. *Alcohol Clin Exp Res, 27*(5), 826-834.

Bush, K., Kivlahan, D. R., McDonell, M. B., Fihn, S. D., Bradley, K. A., & for the Ambulatory Care Quality Improvement, P. (1998). The Audit alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. *Arch Intern Med, 158*(16), 1789-1795.

Bush, K. R., Kivlahan, D. R., Davis, T. M., Dobie, D. J., Sporleder, J. L., Epler, A. J., et al. (2003). The TWEAK is weak for alcohol screening among female veterans affairs outpatients. *Alcohol Clin Exp Res, 27*(12), 1971-1978.

Cremonte, M., & Cherpitel, C. J. (2008). Performance of screening instruments for alcohol use disorders in emergency department patients in Argentina. *Subst Use Misuse, 43*(1), 125-138.

Chan, A. W. K., Pristach, E. A., Welte, J., & Russell, M. (1993). Use of the TWEAK test in screening for alcoholism/ heavy drinking in three populations. *Alcohol Clin Exp Res, 17*(6), 1188-1192.

Cherpitel, C. (1989). A study of alcohol use and injuries among emergency room patients. In N. Giesbrecht, R. Gonzales, M. Grant, E. Osterberg, R. Room, I. Rootman & L. Towle (Eds.), *Drinking and casualties: accidents, poisonings and violence in an international perspective* (pp. 288-299). London, New York: Tavistock/Routledge.

Cherpitel, C. (1995). Screening for alcohol problems in the emergency room: a rapid alcohol problems

screen. *Drug and Alcohol Depend, 40*(2), 133-137.

Cherpitel, C. (2000). A brief screening instrument for problem drinking in the emergency room: the RAPS4. *J Stud Alcohol, 61*(3), 447-449.

Cherpitel, C., & Bazargan, S. (2003). Screening for alcohol problems: comparison of the AUDIT, RAPS4 and RAPS4-QF among African American and Hispanic patients in an inner city emergency department. *Drug Alcohol Depend, 71*(3), 275-280.

Cherpitel, C., & Borges, G. (2000). Performance of screening instruments for alcohol problems in the ER: a comparison of Mexican-Americans and Mexicans in Mexico. *Am J Drug Alcohol Abuse, 26*(4), 683-702.

Cherpitel, C. J., Ye, Y., & Bond, J. (2004). Alcohol and injury: multi-level analysis from the emergency room collaborative alcohol analysis project (ERCAAP). *Alcohol Alcohol, 39*(6), 552-558.

Cherpitel, C. J., Ye, Y., Bond, J., Borges, G., Cremonte, M., Marais, S., et al. (2005). Cross-national performance of the RAPS4/RAPS4-QF for tolerance and heavy drinking: data from 13 countries. *J Stud Alcohol, 66*(3), 428-432.

Cherpitel, C. J., Ye, Y., Moskalewicz, J., & Swiatkiewicz, G. (2005). Screening for alcohol problems in two emergency service samples in Poland: comparison of the RAPS4, CAGE and AUDIT. *Drug Alcohol Depend, 80*(2), 201-207.

Dhalla, S., & Kopec, J. A. (2007). The CAGE questionnaire for alcohol misuse: a review of reliability and validity studies. *Clin Invest Med, 30*(1), 33-41.

Ewing, J. (1984). Detecting alcoholism. The CAGE questionnaire. *JAMA, 252*(14), 1905-1907.

Fiellin, D. A., Reid, M. C., & O'Connor, P. G. (2000). Screening for alcohol problems in primary care: A Systematic Review. *Arch Intern Med, 160*(13), 1977-1989.

Frank, D., DeBenedetti, A. F., Volk, R. J., Williams, E. C., Kivlahan, D. R., & Bradley, K. A. (2008). Effectiveness of the AUDIT-C as a screening test for alcohol misuse in three race/ethnic groups J *Gen Intern Med, 23*(6), 781-787.

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. *Alcohol Clin Exp Res, 29*(11), 2001-2007.

Gentilello, L. M., Ebel, B. E., Wickizer, T. M., Salkever, D. S., & Rivara, F. P. (2005). Alcohol interventions for trauma patients treated in emergency departments and hospitals: a cost benefit analysis. *Ann Surg, 241*(4), 541-550.

Kelly, T. M., Donovan, J. E., Kinnane, J. M., & Taylor, D. M. C. D. (2002). A comparison of alcohol screening instruments among under-aged drinkers treated in emergency departments. *Alcohol Alcohol, 37*(5), 444-450.

Kim, S. S., Gulick, E. E., Nam, K. A., & Kim, S. H. (2008). Psychometric properties of the alcohol use disorders identification test: a Korean version. *Arch Psychiatr Nurs, 22*(4), 190-199.

Knight, J. R., Goodman, E., Pulerwitz, T., & DuRant, R. H. (2000). Reliabilities of short substance abuse screening tests among adolescent medical patients. *Pediatrics, 105*(4 Pt 2), 948-953.

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

Ledesma, R. D., & Molina, J. G. (2009). Classical Item and Test Analysis with Graphics: the ViSta-CITA program. *Behav Res Methods, in press*.

Lima, C. T., Freire, A. C. C., Silva, A. P. B., Teixeira, R. M., Farrell, M., & Prince, M. (2005). Concurrent and construct validity of the AUDIT in an urban Brazilian sample. *Alcohol Alcohol., 40*(6), 584-589.

Medina-Mora, E., Carreno, S., & De la Fuente, J. R. (1998). Experience with the Alcohol use disorders identification test (AUDIT) in Mexico. *Recent Dev Alcohol, 14*, 383-396.

Moraes, C. L., Viellas, E. F., & Reichenheim, M. E. (2005). Assessing alcohol misuse during pregnancy: evaluating psychometric properties of the CAGE, T-ACE and TWEAK in a Brazilian setting. *J Stud Alcohol, 66*(2), 165-173.

Moss, H. B., Chen, C. M., & Yi, H.-Y. (2008). DSM-IV Criteria Endorsement Patterns in Alcohol

Dependence: Relationship to Severity. *Alcohol Clin Exp Res, 32*(2), 306-313.

Munné, M. (2005). Social consequences of alcohol consumption in Argentina. In I. S. Obot & R. Room (Eds.), *Alcohol, Gender and Drinking Problems: Perspectives from Low and Middle Income Countries*. Geneve: World Health Organization.

Nunnally, J. C., & Bernstein, I. J. (1995). *Teoría de la Psicometría.* New York: MacGraw Hill.

Pokorny, A. D., Miller, B. A., & Kaplan, H. B. (1972). The Brief MAST: A Shortened Version of the Michigan Alcoholism Screening Test. *Am J Psychiatry, 129*, 342-345.

Reinert, D. F., & Allen, J. P. (2002). The Alcohol Use Disorders Identification Test (AUDIT): A Review of Recent Research. *Alcohol Clin Exp Res, 26*(2), 272-279.

Reinert, D. F., & Allen, J. P. (2007). The alcohol use disorders identification test: an update of research findings. *Alcohol Clin Exp Res, 31*(2), 185-199.

Rubio Valladolid, G., Bermejo Vicedo, J., Caballero Sanchez-Serrano, M. C., & Santo-Domingo Carrasco, J. (1998). Validation of the Alcohol Use Disorders Identification Test (AUDIT) in primary care. *Rev Clin Esp, 198*(1), 11-14.

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 Alcohol, 37*(3), 261-268.

Rumpf, H. J., Hapke, U., Meyer, C., & Ulrich, J. (2005). Effects of item sequence on the performance of the AUDIT in general practices. *Drug Alcohol Depend, 79*(3), 373-377.

Russell, M., Light, J. M., & Gruenewald, P. J. (2004). Alcohol Consumption and Problems: The Relevance of Drinking Patterns. *Alcohol Clin Exp Res, 28*(6), 921-930.

Saunders, J. B., Aasland, O. G., Amundsen, A., & Grant, M. (1993). Alcohol consumption and related problems among primary health care patients: WHO collaborative project on early detection of persons with harmful alcohol consumption--I. *Addiction, 88*(3), 349-362.

Shields, A. L., & Caruso, J. C. (2004). A reliability induction and reliability generalization study of the

CAGE questionnaire. *Educ Psychol Meas, 64*(2), 254-270.

Steinweg, D., & Worth, H. (1993). Alcoholism: the keys to the CAGE. *Am J Med, 94*(5), 520-523.

World Health Organization. (1992). *International statistical classification of diseases and related health problems (10th rev.)*. Geneva: World Health Organization.

World Health Organization. (1993). *The composite international diagnostic interview (CIDI)* Geneve: World Health Organization.

World Health Organization. (2004). Comparative quantification of health risks In *Global Status Report on Alcohol 2004* (pp. 94). Geneva, Switzerland: World Health Organization.

Role of Funding Sources

Partial funding for this study was provided by the Universidad Nacional de Mar del Plata (UNMdP) and by the Consejo Nacional de Investigaciones científicas y Técnicas (CONICET), Argentina. Partial funding for this study was also provided by a grant from the U.S. National Institute on Alcohol Abuse and Alcoholism (NIAAA) (RO1 AA013750-04). Neither the UNMDP, CONICET, nor NIAAA had any role in the study design, collection, analysis or interpretation of the data, writing the manuscript, or the decision to submit the paper for publication.

Contributors

All authors have materially participated in the research and/or the manuscript preparation. Roles for each author were as follows:

Dr. Cherpitel designed the study and wrote the protocol. Drs. Borges and Cremonte conducted literature searches and provided summaries of previous research studies. Drs. Borges, Cherpitel, and Cremonte collected and provided data from the Mexico, US and Argentina sites, respectively. Dr. Ledesma conducted the statistical analysis. Dr. Cremonte wrote the first draft of the manuscript. All authors contributed to and have approved the final manuscript.

Conflict of Interest

All authors declare that they have no conflicts of interest.

Appendix 1

The Alcohol Use Disorders Identification Test: Interview Version

1. How often do you have a drink containing alcohol?
(0) Never [Skip to Qs 9-10]
(1) Monthly or less
(2) 2 to 4 times a month
(3) 2 to 3 times a week
(4) 4 or more times a week

2. How many drinks containing alcohol do you have on a typical day when you are drinking?
(0) 1 or 2
(1) 3 or 4
(2) 5 or 6
(3) 7, 8, or 9
(4) 10 or more

3. How often do you have six or more drinks on one occasion?
(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily

4. How often during the last year have you found that you were not able to stop drinking once you had started?
(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily

5. How often during the last year have you failed to
do what was normally expected from you
because of drinking?
(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily

6. How often during the last year have you needed
a first drink in the morning to get yourself going
after a heavy drinking session?
(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily

7. How often during the last year have you had a
feeling of guilt or remorse after drinking?
(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily

8. How often during the last year have you been
unable to remember what happened the night
before because you had been drinking?
(0) Never
(1) Less than monthly
(2) Monthly
(3) Weekly
(4) Daily or almost daily

9. Have you or someone else been injured as a result of your drinking?
(0) No
(2) Yes, but not in the last year
(4) Yes, during the last year

10. Has a relative or friend or a doctor or another health worker been concerned about your drinking or suggested you cut down?
(0) No
(2) Yes, but not in the last year
(4) Yes, during the last year

Table 1. Socio-demographic and alcohol drinking characteristics of the sample in each country (current drinkers)

		Argentina (*n*=621)	**México (*n*=559)**	**USA (*n*=827)**
Age		Mean=33 S.D.=15	Mean=35 S.D.=12	Mean=32 S.D.=12
Gender	% Females	37	32	44
Educational level	Elementary	50	45	9
	High school-secondary	35	37	52
	College and above	15	18	39
Drinking habits	% DSM-IV Abuse	9	9	11
	% DSM-IV Dependence	9	12	19
	% Daily of near daily drinking	29	7	15

Table 2. Psychometric characteristics of the AUDIT in each country (current drinkers)

	Sensitivity -S-and Specificity -Sp- for alcohol dependence		**AUDIT Mean (S.D)**	**AUDIT Reliability for all items and Subscales –S-**				**Item Mean (S.D)**								**Item discrimination index [b]**									
	S%	Sp%		All items	S1	S2	S3	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	item10	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	item10
Arg (*n*=599)	93	80	6.59 (6.88)	.88	.69	.84	.80	2.84 (1.01)	.87 (1.24)	.92 (1.30)	.31 (.91)	.22 (.76)	.17 (.71)	.29 (.86)	.20 (.69)	.31 (.99)	.45 (1.22)	.31	.60	.68	.76	.69	.60	.72	.67	.55	.66
Mex (*n*=526)	92	98	7.19 (6.99)	.86	.65	.90	.70	1.43 (.84)	2.21 (1.49)	1.02 (1.17)	.40 (.97)	.24 (.79)	.33 (.95)	.38 (.87)	26 (.76)	24 (.89)	.65 (1.43)	.54	.45	.67	.76	.73	.72	.66	.70	.40	.44
US (*n*=801)	94	81	7.69 (8.90)	.92	.84	.88	.81	1.88 (1.01)	1.08 (1.35)	1.11 (1.33)	.52 (1.13)	.38 (.94)	.38 (1.01)	.56 (1.12)	..39 (.90)	.45 (1.12)	.91 (1.28)	.63	.68	.77	.82	.72	.76	.76	.76	.51	.66

[a] Test reliability estimate by Cronbach's Alpha.

[b] Corrected Item-Test Correlation.

Table 3. Psychometric characteristics of the CAGE in each country (current drinkers)

	Sensitivity - S-and Specificity - Sp- for alcohol dependence		**Correlation with the AUDIT** [a]	**CAGE Mean and (S.D.)**	**CAGE Reliability** [b]	**Item p [Proportion of indicator presence]**				**Item discrimination index** [c]				**Item discrimination index** [d]			
	S %	Sp %				Item1	Item2	Item3	Item4	Item1	Item2	Item3	Item4	Item1	Item2	Item3	Item4
Arg (*n*=528)	75	87	.71**	.37 (.90)	.78	.14	.10	.09	.04	.66	.62	.63	.49	.60	.56	.51	.41
Mex (*n*=541)	92	64	.62**	.82 (1.15)	.72	.37	.15	.20	.10	.51	.50	.60	.44	.46	.42	.46	.50
US (*n*=860)	96	68	.75**	1.05 (1.40)	.82	.38	.20	.29	.17	.67	.66	.72	.56	.66	.55	.61	.59

[a] Pearson correlation coefficient ** p<.001 (one-tailed).

[b] Test reliability estimate by Kuder-Richardson-20 index.

[c] Corrected Item-Test Point-Biserial Correlation.

[d] D-Value.

Table 4. Psychometric characteristics of the RAPS4 in each country (current drinkers)

	Sensitivity -S- and Specificity -Sp- for alcohol dependence		Correlation with the AUDIT[a]	RAPS Mean and (S.D.)	RAPS Reliability[b]	Item p [Proportion of indicator presence]				Item discrimination index[c]				Item discrimination index[d]			
	S %	Sp %				Item1	Item2	Item3	Item4	Item1	Item2	Item3	Item4	Item1	Item2	Item3	Item4
Arg (*n*=594)	89	87	.84**	.25(.71)	.70	.10	.04	.08	.03	.62	.46	.59	.34	.63	.36	.60	.41
Mex (*n*=540)	92	98	.85**	.53(1.01)	.73	.20	.11	.11	.11	.44	.53	.60	.53	.59	.45	.55	.48
US (*n*=844)	95	75	.91**	.82(1.28)	.80	.29	.17	.19	.16	.62	.63	.66	.60	.68	.55	.64	.54

[a] Pearson correlation coefficients ** Significative at 0.01 (one-tailed).

[b] Test reliability estimate by Kuder-Richardson-20 index.

[c] Corrected Item-Test Point-Biserial Correlation.

[d] D-Value.

Table 5. Psychometric characteristics of the TWEAK in each country (current drinkers)

	Sensitivity - S-and Specificity - Sp- for alcohol dependence		**Correlation with the AUDIT[a]**	**TWEAK Mean and (S.D.)**	**TWEAK Reliability[b]**	**Item p [Proportion of indicator presence]**					**Item discrimination index[c]**					**Item discrimination index[d]**				
	S %	Sp %				Item 1	Item 2	Item 3	Item 4	Item 5	Item 1	Item 2	Item 3	Item 4	Item 5	Item 1	Item 2	Item 3	Item 4	Item 5
Arg (*n*=528)	98	67	.77**	1.07 (1.55)	.53	.70	.16	.04	.04	.14	.35	.42	.39	.36	.48	.63	.51	.41	.36	.73
Mex (*n*=539)	90	98	.64**	1.38 (1.32)	.68	.65	.18	.11	.13	.33	.27	.53	.46	.51	.50	.35	.50	.48	.45	.54
US (*n*=801)	91	81	.86**	2.07(2.24)	.71	.94	.40	.16	.17	.40	.40	.61	.56	.57	.56	.41	.66	.54	.55	.63

[a] Pearson correlation coefficients ** Significative at 0.01 (one-tailed).

[b] Test reliability estimate by Kuder-Richardson-20 index.

[c] Corrected Item-Test Point-Biserial Correlation.

[d] D-Value.