White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/144727/

Version: Accepted Version

Article:
Toner, Paul, Bőhnke, Jan, Andersen, Phil et al. (2019) Alcohol screening and assessment measures for young people: a systematic review and meta-analysis of validation studies. Drug and alcohol dependence. pp. 1-44. ISSN: 1879-0046

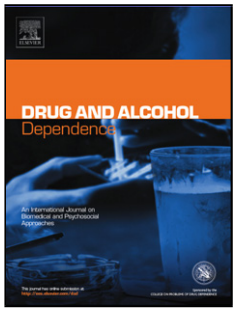
https://doi.org/10.1016/j.drugalcdep.2019.01.030

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Accepted Manuscript



Title: Alcohol screening and assessment measures for young people: A systematic review and meta-analysis of validation studies

Authors: Paul Toner, Jan R. Böhnke, Phil Andersen, Jim McCambridge

PII: S0376-8716(19)30062-6
DOI: https://doi.org/10.1016/j.drugalcdep.2019.01.030
Reference: DAD 7338

To appear in: *Drug and Alcohol Dependence*

Received date: 10 August 2018
Revised date: 10 January 2019
Accepted date: 14 January 2019

Please cite this article as: Toner P, Böhnke JR, Andersen P, McCambridge J, Alcohol screening and assessment measures for young people: A systematic review and meta-analysis of validation studies, *Drug and Alcohol Dependence* (2019), https://doi.org/10.1016/j.drugalcdep.2019.01.030

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Alcohol screening and assessment measures for young people: A systematic review and meta-analysis of validation studies*

Paul Toner[1,2#], p.toner@qub.ac.uk

Jan R. Böhnke[1,3], j.r.boehnke@dundee.ac.uk

Phil Andersen[1], phil.andersen@york.ac.uk

Jim McCambridge[1], jim.mccambridge@york.ac.uk

[1]Department of Health Sciences, University of York, York, YO10 5DD, England, UK

[2]School of Psychology, Queen's University Belfast, Belfast, BT7 1NN, Northern Ireland, UK

[3]Dundee Centre for Health and Related Research, School of Nursing and Health Sciences, University of Dundee, Dundee, DD1 4HJ, Scotland, UK

[#]Correspondence:

Paul Toner

School of Psychology, University of York

David Keir Building, 18-30 Malone Road,

Queen's University Belfast, Belfast, BT7 1NN, Northern Ireland, UK

Email: p.toner@qub.ac.uk

* Supplementary material can be found by accessing the online version of this paper at http://dx.doi.org and by entering doi:...

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Highlights

- First review of validation evidence for measures to assess young people's drinking
- Alcohol frequency or quantity is an efficient single item screener for risky drinking
- Most validation evidence for screening measures exists for the AUDIT and AUDIT-C
- Little validation evidence for assessment measures of alcohol-related problems
- New assessment instruments are needed to capture alcohol-related problems

Abstract

Background: There is a strong rationale for clinicians to identify risky drinking among young people given the harms caused by alcohol. This systematic review evaluates the quality of evidence in the validation literature on alcohol screening and assessment measures for young people under 25.

Methods: Six electronic databases (MEDLINE; EMBASE; PsycINFO; SSCI; HMIC; ADAI) were searched in May 2016 for published and grey literature. Full-text reports published in English since 1980 were included if they aimed to validate an alcohol screening or assessment measure in comparison with a previously validated alcohol measure. Risk of bias was assessed in studies surpassing *a priori* quality thresholds for predictive validity, internal and test-retest reliability using COSMIN and QUADAS-2.

Results: Thirty nine reports comprising 135 discrete validation studies were included. Summary estimates indicated that the screening instruments performed well - AUC 0.91 (95% CI: 0.88 to 0.93); sensitivity 0.98 (0.95 to 0.99); specificity 0.78 (0.74 to 0.82). Noting a paucity of validation evidence for existing assessment instruments, aggregated reliability estimates suggest

ACCEPTED MANUSCRIPT

a reliability of 0.81 (0.78 to 0.83) adjusted for 10 items. Risk of bias was high for both types of studies.

Conclusions: The volume and quality of available evidence are superior for screening measures. It is recommended that clinicians use alcohol frequency or quantity items if asking a single question. If there is an opportunity to ask more questions either the 3-item AUDIT-C or the 10-item AUDIT are recommended. There is a need to develop new instruments to assess young people's alcohol-related problems.

Keywords: Alcohol; Screening; Assessment; Young people; Systematic review; Meta-analysis

1.1 Background

Adolescent drinking is a major global health concern (Gore et al., 2011). The proportion of school children aged 11 to 15 in England who drink alcohol decreased from 61% in 2003 to 44% in 2016 (Niblett, 2017). Prevalence rises steeply within this age band, however, meaning that it remains normative to drink alcohol at age 15 (Niblett, 2017). Many drink hazardously and place themselves at risk of harm (Townshend, 2013). Systematic review evidence suggests that it is the nature of alcohol use in adolescence rather than age of first use which confers risk of adverse consequences (Maimaris and McCambridge, 2014). There is consistent evidence that higher alcohol consumption in late adolescence continues into adulthood, and is also associated with alcohol problems including dependence (McCambridge et al., 2011). This means that assessing drinking behavior among adolescents potentially offers opportunities for early interventions.

ACCEPTED MANUSCRIPT

There is a long history of efforts to better understand the nature of alcohol problems and how they develop over the life course (Edwards, 2012; Edwards et al., 1992). In line with this thinking, alcohol problems are broadly defined as adverse consequences experienced due to alcohol use. Therefore, constituting a form of substance use disorder that may vary in severity. Alcohol consumption items are prominent in screening measures, though these may also address adverse consequences directly, and be predictive of alcohol problems both contemporaneously and over time (Saunders et al., 1993). Assessment instruments are primarily concerned with alcohol problems. Alcohol problems may stem from intoxication in acute episodes or continued heavy use, and include physical (e.g., injuries), psychological (e.g., depression) and social (e.g., educational) harms (Lester et al., 2018). The contribution of alcohol to complex psychosocial problems requires careful assessment. Alcohol consumption may or may not be implicated directly in the reasons for presentation to services. Adolescents who drink heavily also risk physical health consequences later in the life course (Hagström et al., 2018).

Expert guidance in the UK has emphasized the pressing need for research to identify a 'gold standard' screening measure to assess the drinking behavior of young people under the age of 18 (NICE, 2010). Indeed the adequacy of existing measures for young people up to the age of 25 - in line with the United Nations definition of adolescence (Secretary-General, 1981) - is also unknown. As there are no systematic reviews of validation studies of alcohol screening and problems assessment instruments for this age group. This study aims to identify the best performing measures for screening and assessment respectively for young people up to the age of 25, based on their psychometric properties and the methodological quality of the underpinning validation studies.

2.1 Methods

ACCEPTED MANUSCRIPT

The Centre for Reviews and Dissemination (CRD) (CRD, 2009) and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al., 2009) guidelines were adhered to in conducting and reporting this study.

2.1.1 Electronic Searches. The following databases were searched in May 2016 for published literature: Medical Literature Analysis and Retrieval System Online (MEDLINE; Ovid 1946-), Excerpta Medica Database (EMBASE; Ovid 1974-), Psychological Information Database (PsycINFO; Ovid 1806-), and Social Sciences Citation Index (SSCI; Web of Science 1956-). Additionally, the Health Management Information Consortium Database (HMIC; Ovid 1979-) and the University of Washington Alcohol and Drug Abuse Institute (ADAI) Library Search – Substance Use Screening and Assessment Instruments Database were searched for grey literature.

The search strategies were designed using Medical Subject Headings (MeSH) and free text words adapted for each database. Three sets of search terms were combined: (1) Alcohol use and alcohol problems including substance use. (2) Young people. (3) Validation studies. See (supplementary material) SI[1] for the MEDLINE search strategy.

2.1.2 Searching Other Resources. Reverse and forward citation searching were performed using the Social Sciences Citation Index. Authors of relevant reports (n=12) were contacted to identify additional reports not identified through the database searches.

2.1.3 Selection Of Studies. Two reviewers (PT, PA) independently screened the titles and abstracts of reports retrieved by the searches using EndNote X7. Those carried forward were obtained as full text articles which were assessed for inclusion using a checklist based on pre-specified selection criteria (see section 2.2). Discrete validation studies within included reports

[1] Supplementary material can be found by accessing the online version of this paper at http://dx.doi.org and by entering doi:...

ACCEPTED MANUSCRIPT

were also assessed for inclusion using the same criteria. This was necessary because it was possible, and indeed common, for research reports to include more than one validation study. Two reviewers (PT, JB) separately screened both reports and validation studies within reports for inclusion. Where eligibility was unclear this was resolved by discussion with a third reviewer (JM).

2.2 Selection criteria

2.2.1 Types Of Studies**.** Any type of validation study published in the English language from 1980 onwards which aimed to validate an alcohol screening or assessment measure (index test) in comparison with a previously validated alcohol measure (reference test) were eligible for inclusion.

2.2.2 Participants. Studies of young people aged 24 or under were eligible for inclusion. Many alcohol questionnaires have been validated in (university) student samples, where study populations are not defined by age, so it is possible for some participants to be over the age of 24 in these studies. We, therefore, required at least 80% of participants were aged 24 or under in studies including older participants to allow such studies to be included. Where only mean or median age was reported, it was decided *a priori* that this was required to be not older than 21 years for the report to be included. Studies undertaken in student samples without age being defined were eligible for inclusion, unless there were specific reasons to be concerned that below 80% of the participants were aged 24 or under.

2.2.3 Index Tests. Alcohol screening or assessment measures as above.

2.2.4 Reference Tests – Comparators. The reference tests were previously validated questionnaires or diagnostic interviews assessing alcohol use or problems. Where alcohol was

ACCEPTED MANUSCRIPT

assessed alongside other drugs, the study was included only if the reference test provided an alcohol-specific result against which the index test was compared in the validation study.

The following were not considered to be valid reference tests; clinician judgment; alcohol biomarkers; alcohol diagnoses which were a composite of information contained within medical records; generic substance use measures which did not report a validated assessment of alcohol; and alcohol questions which had not been previously validated.

2.2.5 Outcomes. The direct reporting of predictive, including concurrent, validity of the index test against a comparator was required. Acceptable data were: standardized regression coefficients, odds ratios, correlation statistics, area under the curve (AUC) or % sensitivity; % specificity or % positive predictive value (PPV); % negative predictive value (NPV) or likelihood ratio.

2.3 Data Collection And Analysis

2.3.1 Data Extraction. One reviewer (PT) extracted all relevant data (see below) from included studies using a dedicated form. This was checked by a second reviewer (JB).

Many included full papers/reports contained multiple validation studies, defined for the purposes of this review as comparisons of index and reference tests. A single record for data extraction was created for each validation study. The process was as follows:

(1) The eligibility criteria used to include reports in the review were also applied to each of the validation studies within the included reports.

(2) If a validation study was included, then quality threshold data (see section 2.3.2) were extracted. If the index test failed to make any of the *a priori* quality thresholds on predictive validity, internal or test-retest reliability, this study was recorded as included in the review at step 2, with no further data extraction.

ACCEPTED MANUSCRIPT

(3) If the quality thresholds (see section 2.3.2) were met, then full data extraction and quality assessment were conducted in step 3.

For validation studies that reported only data on the subscales of a questionnaire, data were extracted as described in steps 2 and 3, thus treating the subscale as the index test. The same approach was taken with studies only reporting validation data for specific subpopulations, for example, in age categories or by gender, each subsample was treated as a separate validation study.

The following data were extracted from index tests:

(1) Predictive validity: cut-off scores (thresholds on each questionnaire), standardized regression coefficient, odds ratio, correlation coefficient, AUC, % sensitivity, % specificity, % PPV, % NPV, and likelihood ratio.

(2) Internal validity: item-to-total correlations and percentage of explained variance by proposed factor model.

(3) Reliability: Cronbach's alpha (which was adjusted for 10 items), Guttman's lambda, omega, Pearson correlation, intraclass correlation coefficient and kappa coefficient.

(4) Information on acceptability and/or feasibility.

Descriptive details on the index tests such as instrument name and acronym, whether used for screening and/or assessment were also recorded on a Microsoft Excel 2010 spreadsheet (see (Toner et al., 2017) for more details).

2.3.2 Quality Assessment**.** In line with standard practices in psychometric research, there were *a priori* quality thresholds used to determine which studies warranted full data extraction (see the published protocol (Toner et al., 2017) for further details).

The index test was required to achieve:

ACCEPTED MANUSCRIPT

(1) A predictive validity of above 0.7 (e.g., standardized regression coefficient) or 0.8 AUC/% Sensitivity OR

(2) An internal consistency above 0.8 (adjusted Cronbach's alpha for 10 items) OR

(3) A test-retest value of above 0.7 (e.g., Kappa coefficient).

Studies which compared a short version of a parent instrument as index and reference tests respectively were excluded from quantitative synthesis. This was due to the potential for overestimation of validity.

The quality of included studies above the quality thresholds was further assessed using two tools:

1) A modified consensus-based standards for the selection of health measurement instruments (COSMIN) checklist (Terwee et al., 2012) to assess the methodological quality of studies.

2) A quality assessment tool for diagnostic accuracy studies (QUADAS-2) (Whiting et al., 2011) to evaluate the risk of bias.

As recommended the COSMIN checklist was applied in modular fashion. Therefore, it was not necessary to complete the whole checklist when evaluating studies. The measurement properties evaluated in studies determine what domains are rated. For example, the internal consistency domain would not be applicable to studies evaluating single item screeners. The QUADAS-2 assesses the quality of diagnostic accuracy studies. It comprises four domains: participant selection; index test; reference standard; participant flow and timing of index test/reference standard administration (flow and timing). Each domain is assessed for risk of

ACCEPTED MANUSCRIPT

bias. Equal weight was given to both tools *a priori* for interpretation. Domain codes for included studies and overall ratings/scores are presented in Tables SII and SIII[2].

2.3.3 Data Synthesis**.** The selection criteria were primarily geared towards gathering information on scales rather than on population characteristics, apart from age. Therefore, the generalisability of meta-analytic estimates may be challenging to assess. The meta-analytic estimates and their uncertainty (i.e., confidence intervals) as demonstrated in this study are proposed as empirical benchmarks that should be met by new instruments covering related content. This is a departure from the usual meta-analytic goal of providing a single summary estimate of an effect.

For diagnostic validity studies, data on sensitivity and specificity were extracted for the cut-offs suggested by the authors of the primary studies. Where these data were not available, raw data on true and false positives and true and false negatives were extracted. As direct reporting of predictive validity was part of the inclusion criteria, primary study authors were not contacted for additional data. Data were checked and agreed by two reviewers (PT, JB) and exported from Microsoft Excel 2010 to Stata V.14 (StataCorp, 2015). Aggregate diagnostic validity statistics were calculated using the Stata extensions midas (Dwamena, 2007) and metandi (Harbord, 2008). For studies reporting reliability estimates for included instruments, the same extraction procedure was undertaken and calculations were performed on normalized reliability estimates α_i ($T_i = (1- \alpha_i)^{1/3}$; (Botella et al., 2010)). All aggregates and meta-analytic estimates took account of the nested structure of the data and were calculated using the Stata

[2] Supplementary material can be found by accessing the online version of this paper at http://dx.doi.org and by entering doi:...

ACCEPTED MANUSCRIPT

extension metaan (Kontopantelis and Reeves, 2009). Meta-analytic estimates and forest plots are presented in Figures SIV-1 and 2[3].

Heterogeneity was assessed via I^2 statistics for diagnostic and reliability studies, and Q statistics for reliability studies. In addition to the aggregates across all available studies, subgroup analyses and subgroup estimates are also reported where there was a minimum of five studies per subgroup (see study protocol (Toner et al., 2017) and Tables SIV-1 and 2)[4]. Subgroup heterogeneity was explored for year of publication, sample size, percentage female, mean age, country, ethnicity, index tests, reference tests, population (i.e., clinical, community) and setting (i.e., health, school; see protocol (Toner et al., 2017)). The potential for a meta-regression was evaluated, but due to the small number of studies compared to the large number of predictors and complex nesting within reports (especially for diagnostic studies), this was not performed.

3.1 Results

The PRISMA flowchart is presented in Figure 1. There were 39 research reports included in this review (Aertgeerts et al., 2000; Chung et al., 2002; Chung et al., 2000; Chung et al., 2012; Clark et al., 2006; Clark et al., 2016; Cook et al., 2005; DeMartini and Carey, 2012; Dick et al., 2011; Earleywine et al., 2008; Edelen et al., 2009; Ferreira et al., 2014; Fleming et al., 1991; Hurlbut and Sher, 1992; Kahler et al., 2008; Kahler et al., 2005; Kelly et al., 2009; Kelly et al., 2004; Kelly et al., 2002; Knight et al., 2003; Kokotailo et al., 2004; Levy et al., 2016; López-Núñez et al., 2012; Martens et al., 2007; McCambridge and Thomas, 2009; McGee and Kypri, 2004; Neal et al., 2006; Northrup et al., 2013; O'Hare, 2005; O'Hare et al., 1997; O'Hare and Sherrer, 1999; Read et al., 2007; Rumpf et al., 2013; Santis et al., 2009; Taylor et al., 2008;

[3-4] Supplementary material can be found by accessing the online version of this paper at http://dx.doi.org and by entering doi:...

ACCEPTED MANUSCRIPT

Thomas and McCambridge, 2008; Van den Bruel et al., 2004; Verster et al., 2009; West and Graham, 2001), including 135 discrete validation studies evaluating (coincidentally) 39 instruments. Twenty eight instruments yielded dichotomous classifications, for example, drinking at risky levels or not at risk, and were primarily utilized as screening measures. There were 11 continuous measures, mainly used for assessing alcohol problems. Table 1 provides details on the study characteristics of included studies.

For instruments capturing alcohol consumption, a single frequency item had most validation studies supporting its use (n=18) – *On how many days have you had an alcoholic drink in the past 12 months?* A single quantity item was the next best supported (n=10) - *On the days that you drank during the past 30 days, how many drinks did you usually have each day?*

However, one report (Chung et al., 2012) provided 78% (14 of 18 studies) of the available validation data for the frequency item and 70% (7 of 10 studies) of the available data for the quantity item. The AUDIT-C (Bush et al., 1998), the first three items of the full AUDIT (Alcohol Use Disorders Identification Test (Saunders et al., 1993)) also had 10 validation studies supporting its use, assessed against a wider variety of reference tests. There was also support for a modified version of the AUDIT-C from an additional validation study (see Table 2).

For instruments examining both consumption and consequences, the full 10-item AUDIT had the most validation studies supporting its use (n=10). In addition, there were two validation studies supporting a modified version of the full AUDIT. The AUDIT also had seven validation studies supporting combinations of the alcohol problem items only (see Table 2).

There were fewer studies of other multi-item consumption and/or problems screening measures, though some performed satisfactorily in validation studies. For example, the CRAFFT (Car, Relax, Alone, Forget, Friends, Trouble (Knight et al., 1999)) modified to screen for alcohol

ACCEPTED MANUSCRIPT

had three validation studies over the quality threshold. Other instruments produced mixed findings in validation studies. This was in line with findings for instruments assessing alcohol problems only. There was quite limited validation evidence available compared to screening measures, and more mixed evidence in study findings (see Table 3). The YAACQ (Young Adult Alcohol Consequences Questionnaire (Read et al., 2006)) was supported by four validation studies, whilst a brief version yielded mixed findings.

Table 4 presents aggregate statistics for diagnostic criteria for screening instruments where at least five studies in independent samples were available (Toner et al., 2017). For single item screeners of alcohol frequency and quantity, only three (Chung et al., 2012; Clark et al., 2006; Clark et al., 2016) and two (Chung et al., 2012; Clark et al., 2016) reports were available. Reporting on 18 and 10 independent samples respectively. Regarding average sensitivity and specificity, both single items tend to perform better than the AUDIT. Although the majority of the estimates emanate from a single large scale epidemiological study (Chung et al., 2012).

The largest number of independently undertaken studies was available for the full AUDIT - nine reports containing 10 studies. Across those studies, an average sensitivity above the *a priori* threshold can be reported. Also, the full AUDIT and AUDIT-C are the only screening instruments with enough validation data to report an average reliability. For AUDIT-C, five reports were available presenting data on 10 studies, with virtually the same average estimates for all diagnostic criteria as the full AUDIT. The only exception is higher reported reliability, potentially due to the three consumption items being more homogenous than the full set of 10 AUDIT items.

3.2 Quality Assessment

ACCEPTED MANUSCRIPT

All studies were found to be at risk of bias using the QUADAS-2. Across studies, the greatest risk of bias was flow and timing, specifically that all participants were not included in the analyses reported. There were many unclear codes both for the index test and reference standard domains. These reveal problems in reporting where it is not possible to assess the rigor of study conduct. Since all studies were rated as being at risk of bias, QUADAS-2 ratings were not used in further heterogeneity analysis (see Tables SIV-1 and 2)[4].

Using the COSMIN checklist, although criterion validity was assessed in all included studies, none of the studies scored excellent for methodological quality. Only three validation studies from two reports (Aertgeerts et al., 2000; Edelen et al., 2009) were rated as good. The vast majority of studies were rated fair (n = 93) (Chung et al., 2002; Chung et al., 2000; Chung et al., 2012; Clark et al., 2006; Clark et al., 2016; Cook et al., 2005; DeMartini and Carey, 2012; Fleming et al., 1991; Hurlbut and Sher, 1992; Kahler et al., 2008; Kelly et al., 2009; Kelly et al., 2004; Knight et al., 2003; Kokotailo et al., 2004; Levy et al., 2016; López-Núñez et al., 2012; McGee and Kypri, 2004; Northrup et al., 2013; O'Hare, 2005; O'Hare et al., 1997; O'Hare and Sherrer, 1999; Read et al., 2007; Rumpf et al., 2013; Santis et al., 2009; Thomas and McCambridge, 2008; Van den Bruel et al., 2004; Verster et al., 2009) and the remaining six studies from four reports (Ferreira et al., 2014; Kahler et al., 2005; Kelly et al., 2002; West and Graham, 2001) were rated as poor. Similar to QUADAS-2, lower scores were given mainly because only participants with complete data were included in the analyses and/or no details were provided on missing data.

Internal consistency was the second most assessed domain across studies, especially for assessment measures. Only one study was rated excellent (Edelen et al., 2009) for methodological quality. Thirty three studies from 12 reports were rated as fair (Chung et al.,

ACCEPTED MANUSCRIPT

2002; Fleming et al., 1991; Hurlbut and Sher, 1992; Kelly et al., 2009; Kelly et al., 2004; Kelly et al., 2002; Kokotailo et al., 2004; López-Núñez et al., 2012; O'Hare and Sherrer, 1999; Read et al., 2007; Rumpf et al., 2013; Thomas and McCambridge, 2008) and nine studies from eight reports were rated as poor (Ferreira et al., 2014; Kahler et al., 2008; Kahler et al., 2005; Kelly et al., 2004; McGee and Kypri, 2004; Santis et al., 2009; Verster et al., 2009; West and Graham, 2001). Studies scored lower mostly due to lack of reporting relating to missing data and/or factor analysis not being performed for new instruments or were not referenced for index tests that were not new. As a consequence of only including validation studies in this review, content validity and hypotheses testing were not assessed in any of the included studies. However, scores for all the domains assessed in each study and average study scores for methodological quality are presented in Table SIII[5].

Aggregated quality assessments (average of all rated domains) were used to explore heterogeneity in diagnostic accuracy and reliability. Table 4 reports the averages for the five measures with sufficient studies available. The aggregated average is two (approximately so for the full AUDIT), which is equivalent to fair methodological quality.

3.3 Meta-Analytic Estimates Of Diagnostic Accuracy And Reliability

A total of $n = 53$ independent samples in screening measure validation studies were drawn from 18 reports (Aertgeerts et al., 2000; Chung et al., 2002; Chung et al., 2000; Chung et al., 2012; Clark et al., 2006; Clark et al., 2016; Cook et al., 2005; DeMartini and Carey, 2012; Fleming et al., 1991; Kelly et al., 2009; Kelly et al., 2004; Knight et al., 2003; Kokotailo et al., 2004; López-Núñez et al., 2012; O'Hare, 2005; O'Hare et al., 1997; Rumpf et al., 2013; Van den Bruel et al., 2004) and included in an examination of diagnostic accuracy. In these studies,

[5] Supplementary material can be found by accessing the online version of this paper at http://dx.doi.org and by entering doi:...

ACCEPTED MANUSCRIPT

26,806 (14%) participants were reference-test positive and 159,803 were reference-test negative. The forest plot of all study estimates is presented in Figure SIV-1. The aggregated area under the curve was 0.91 (95% CI: 0.88 to 0.93); the estimate for sensitivity was 0.98 (0.95 to 0.99); and the estimate for specificity was 0.78 (0.74 to 0.82).

Figure 2 presents the estimated ROC curve in Panel A. Panel B shows how assumptions about the prevalence of hazardous drinking in the youth population influence the posterior probability associated with a test result. With a positive test result (long dashed), there is always a tangible probability that the young person is using alcohol in a hazardous way, regardless of the prevalence. For a negative test result (short dashed) this probability is close to 0 and only rises when the prevalence in the population is above 80 percent. Assuming for illustrative purposes that the true prevalence of hazardous drinking among all young people is between 1 and 14 percent which was the prevalence estimate from studies included in the meta-analysis. The negative predictive value is 1 and the positive predictive value is 0.26.

The heterogeneity observed in the sample was extremely high. The overall inconsistency was $I^2 = 100$, primarily due to the selection of studies over a minimal sensitivity threshold which reduces the sample's total variance. Detailed findings are presented in Table SIV-1. For several of these criteria the I^2 was reduced notably when the data was split, particularly for more recent studies, and in samples with a low percentage of females.

Since one report (Chung et al., 2012) provided a large amount of data (26 discrete validation studies), the diagnostic meta-analysis was also conducted without this report. The aggregated area under the curve was 0.90 (95% CI: 0.88 to 0.93; virtually unchanged); the estimate for sensitivity was 0.88 (0.86 to 0.90; reduced by 0.10); the estimate for specificity was 0.77 (0.71 to 0.81; slightly lower). Heterogeneity was still very high - $I^2 = 98$.

ACCEPTED MANUSCRIPT

A total of *n* = 20 reports (Chung et al., 2002; Edelen et al., 2009; Ferreira et al., 2014; Fleming et al., 1991; Hurlbut and Sher, 1992; Kahler et al., 2008; Kahler et al., 2005; Kelly et al., 2009; Kelly et al., 2004; Kelly et al., 2002; Kokotailo et al., 2004; López-Núñez et al., 2012; McGee and Kypri, 2004; O'Hare and Sherrer, 1999; Read et al., 2007; Rumpf et al., 2013; Santis et al., 2009; Thomas and McCambridge, 2008; Verster et al., 2009; West and Graham, 2001) provided reliability estimates for 26 instruments. In these studies, data were gathered on 12,760 participants. The forest plot of all study estimates is presented in Figure SIV-2. The aggregated reliability estimate for an expected scale length of 10 items was 0.81 (95% CI: 0.78 to 0.83; transformed estimates: 0.58; 95% CI: 0.55 to 0.61). The heterogeneity in reliability estimates was investigated taking into account the same criteria for diagnostic studies and was also high (I^2 = 96.8). The heterogeneity was not substantially reduced by any of the criteria apart from the index test. Five different reports tested the AUDIT and amongst those the heterogeneity was substantially lower than within all other groups - I^2 = 22.7 (see Table SIV-2). The average estimates for reliability were all very similar. The largest difference was observed for the median split of aggregated COSMIN scores, with an estimated reliability of 0.77 in studies with lower quality ratings versus 0.84 in studies with higher quality ratings.

4.1 Discussion

This systematic review evaluates existing evidence with contrasting findings for screening and assessment instruments for alcohol use and problems respectively in young people. The volume of evidence is larger for screening instruments with the alcohol use frequency single item screener having somewhat more validation evidence supporting it than alcohol use quantity. It should be borne in mind, however, that many studies of both measures originate from the same report. For the multi-item screeners, there were 10 studies supporting

ACCEPTED MANUSCRIPT

both full AUDIT and AUDIT-C, and in general, they performed similarly to each other. Although this may suggest the briefer 3-item AUDIT-C is preferable, it is important to note that the full AUDIT was validated in a larger number of separate reports (nine versus five). The predictive validity values are identical, with the AUDIT-C marginally outperforming the AUDIT on AUC. Overall, the psychometric evidence including heterogeneity data for the full AUDIT is stronger at this point in time. However, the promising performance of the AUDIT-C is striking and better quality studies are needed (see below) to confirm this finding, which should be regarded as preliminary. In contrast, there was limited and weak evidence for alcohol assessment measures, with none of the identified instruments having enough validation studies to support meta-analysis. The YAACQ appears most promising among the existing measures (see Table 3). The risk of bias afflicting both studies of screening and assessment measures is a major finding of this review and calls for stronger designs and analyses in future research.

In terms of the quality of the primary studies included in the review, with the exception of the study by Edelen and colleagues (Edelen et al., 2009) which validated alcohol problem items and had an overall COSMIN ranking of good to excellent, the other 38 reports were rated fair or poor. Additionally, many studies did not perform all relevant tests for instrument development or validation (see Table SIII). Raising awareness of available frameworks to plan and evaluate psychometric work such as COSMIN (Terwee et al., 2012) and the importance of setting clear goals for a particular study (especially, measurement versus identification; Smits et al., 2018), could help improve the quality of psychometric assessments for alcohol problems. Also, all studies validating both screening and assessment measures included were scored at risk of bias on the QUADAS-2. Therefore, a review-level finding is that the conduct and reporting of the primary studies need to be improved. Potentially good quality diagnostic reports, for example

ACCEPTED MANUSCRIPT

(Aertgeerts et al., 2000), also suffered from reporting problems relating to whether the results of either index or reference tests were interpreted without knowledge of the other. Other potential biases were not examined in the quality criteria. Attention is warranted, for example, to instruments validated by their developers (see for instance (Manea et al., 2017) on allegiance effects) particularly for alcohol problems assessment measures where limited validation evidence was available.

The common use of single item screeners in the included studies is not an indication of superior validity. Their popularity likely points to the field's preference for short screening instruments, and relatedly time pressures in practice. Single item measures have implied challenges such as a potentially higher propensity for response errors and a limited range of construct content being represented. Whether or not a single item is appropriate depends ultimately on its intended use. The usefulness of an instrument for screening purposes largely depends on its ability to correctly classify individuals (i.e., criterion or predictive validity) for which the use of a multi-item instrument may potentially even be detrimental (Smits et al., 2018). Alternatively, if assessment of the severity of alcohol problems is the main goal, increasing the precision of measurement (reliability) is important, which is potentially improved by using multiple items. Planning instrument development and validation studies with these competing goals in mind (Smits et al., 2018; Costa, 2016) and informing study design with existing psychometric evaluation frameworks such as COSMIN (Terwee et al., 2012) is important for future advances in the field.

There are a number of limitations to the current systematic review. English language reports only were included, although evaluated instruments could be developed or exist in other languages (e.g., Ferreira et al., 2014; B-YAACQ). This review provides an appraisal of the

ACCEPTED MANUSCRIPT

validation literature as it existed at the time the searches were completed. It is possible that more recent studies may address, at least in part, some of the limitations of the literature identified. Further studies are unlikely, however, to undermine the main findings of this review.

The evaluated instruments were largely drawn from studies that were not independent of each other. There were 39 reports, each containing an average of 3.28 validation studies. Only one report with multiple data points entered the diagnostic accuracy meta-analysis (Chung et al., 2012), the impact of which was addressed by a sensitivity analysis (see Results). To assess the potential impact due to clustering of studies within reports for the reliability meta-analysis, intraclass correlations (ICCs) were estimated for the transformed reliabilities. An ICC of < .01 indicated a very small cluster effect.

The most in-depth quantitative analyses focused on the best performing screening and assessment measures to provide a benchmark for further research on instrument development. Therefore, the presentation of results may be skewed in favor of instruments which surpassed the *a priori* quality thresholds. Study findings should be interpreted with this important caveat in mind.

Despite study limitations, this is the first review to synthesize the available validation evidence on alcohol measures for young people. Additionally, it combines and appraises both the alcohol screening and assessment literature. For screening, the ROC summary plot brings together the best performing instruments to identify for the first time, benchmarks against which future studies validating and/or developing measures in the population can compare their performance. New instruments which do not fulfill diagnostic criteria within or above the range of values identified as benchmarks, cannot be regarded as equivalent in performance to existing measures.

ACCEPTED MANUSCRIPT

For assessment measures, there is a lack of validation evidence to support similar recommendations. However, aggregated reliability estimates for multi-item instruments suggest an adjusted (for 10 items) alpha value of 0.81 (0.78 to 0.83). There is clearly considerable scope for improvement in the assessment of alcohol problems for young people and the development and testing of new measures. As the majority of studies included within this review had important methodological weaknesses, future validation studies should apply best practices in relation to appropriate quality checklists and reporting guidance (e.g., Standards for Reporting Diagnostic Accuracy Studies; Cohen et al., 2016).

4.2 Conclusions

This review provides a rigorous analysis of available evidence on the psychometric performance of instruments for alcohol screening and assessment in young people. On the basis of this evidence, we can make recommendations in relation to both practice and research for screening. In relation to practice implications from the review, both alcohol frequency or quantity single item screeners performed well. The AUDIT-C is a promising screening tool for alcohol consumption in young people, and further studies are needed to determine whether it may in time replace the full AUDIT in this population. The YAACQ is the best existing instrument for assessing alcohol problems, though it contains 48 items which may not be practical to implement in many settings. Also, the small body of evidence supporting it may be biased. Therefore, we suggest that research efforts to develop new assessment instruments draw on existing item content found to perform well in psychometric studies.

It is appropriate to situate this study within the wider context of developments in thinking about the nature of alcohol and other addiction problems, and how they may be conceptualized and assessed to inform interventions. DSM-V and International Classification of Diseases (ICD)-

ACCEPTED MANUSCRIPT

11 offer contrasting conceptualizations of the nature of substance use disorders, including whether or not dependence may be meaningfully separated from non-dependence problems (Edwards, 2012; Edwards et al., 1992). In such circumstances, unresolved construct validity issues are highly likely to limit the progress possible in measurement studies. This is without engaging with issues such as whether the nature of alcohol problems are importantly different in young people compared to older adults or across populations characterized in other ways. The findings of this review on assessment instruments are perhaps unsurprising in this light. What is required are agreed ways to conceptualize, as well as to measure, alcohol-related problems in populations of different ages. This review demonstrates that there are existing instruments able to capture relatively well whether young people's drinking is risky, but we are not well placed to identify how far their drinking is problematic.

Author Disclosures

Role of Funding Source
This study was funded by the Society for the Study of Addiction (SSA) who supported PT with a Griffith Edwards Academic Fellowship award to conduct the review. The SSA had no involvement in the design of the study, data collection, analysis, interpretation or in the decision to publish the study.

Contributors
PT conceived the study, is guarantor and co-designed the review, and led on data collection, analysis, interpretation and preparing the manuscript. JB co-designed the review and provided advanced psychometric expertise on data collection, analysis and interpretation. He was involved in drafting the manuscript for important intellectual content. PA assisted with data collection, analysis and interpretation. He revised the manuscript for important intellectual content. JM co-

ACCEPTED MANUSCRIPT

designed the review and provided expert guidance and support on data collection, analysis and interpretation. He critically revised the manuscript for important intellectual content.

All authors read and approved the final manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflict of Interest
No conflict declared.

Data Sharing
Data sets generated are available from the corresponding author on reasonable request.

Acknowledgements
The authors would like to thank Nick Meader (CRD, University of York) and David Brown (Department of Health Sciences Academic Liaison Librarian, University of York) for their advice on review methods and search strategy respectively.

Ethics Statement
Ethical approval for this study was not required.

References

Aertgeerts, B., Buntinx, F., Bande-Knops, J., Vanderrneulen, C., Roelants, M., Ansoms, S., Fevery, J., 2000. The value of CAGE, CUGE, and AUDIT in screening for alcohol abuse and dependence among college freshmen. Alcohol. Clin. Exp. Res. 24, 53-57. https://doi.org/10.1111/j.1530-0277.2000.tb04553.x

Botella, J., Suero, M., Gambara, H., 2010. Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. Psychol. Methods 15, 386-397. https://doi.org/10.1037/a0019626

ACCEPTED MANUSCRIPT

Bush, K., Kivlahan, D.R., McDonell, M.B., Fihn, S.D., Bradley, K.A., 1998. The AUDIT alcohol consumption questions (AUDIT-C): An effective brief screening test for problem drinking. Arch. Intern. Med. 158, 1789-1795. https://doi.org/10.1001/archinte.158.16.1789

Chung, T., Colby, S.M., Barnett, N.P., Monti, P.M., 2002. Alcohol use disorders identification test: Factor structure in an adolescent emergency department sample. Alcohol. Clin. Exp. Res. 26, 223-231. https://doi.org/10.1111/j.1530-0277.2002.tb02528.x

Chung, T., Colby, S.M., Barnett, N.P., Rohsenow, D.J., Spirito, A., Monti, P.M., 2000. Screening adolescents for problem drinking: performance of brief screens against DSM-IV alcohol diagnoses. J. Stud. Alcohol Drugs 61, 579-587. https://doi.org/10.15288/jsa.2000.61.579

Chung, T., Smith, G.T., Donovan, J.E., Windle, M., Faden, V.B., Chen, C.M., Martin, C.S., 2012. Drinking Frequency as a Brief Screen for Adolescent Alcohol Problems. Pediatrics 129, 205-212. https://doi.org/10.1542/peds.2011-1828

Clark, D.B., Chung, T., Martin, C., 2006. Alcohol use frequency as a screen for alcohol use disorders in adolescents. Int. J. Adolesc. Med. Health 18, 181-187. https://doi.org/10.1515/IJAMH.2006.18.1.181

Clark, D.B., Martin, C.S., Chung, T., Gordon, A.J., Fiorentino, L., Tootell, M., Rubio, D.M., 2016. Screening for underage drinking and DSM-5 alcohol use disorder in rural primary care practice. J. Pediatr. 173, 214-220. https://doi.org/10.1016/j.jpeds.2016.02.047

Cohen, J.F., Korevaar, D.A., Altman, D.G., Bruns, D.E., Gatsonis, C.A., Hooft, L., Irwig, L., Levine, D., Reitsma, J.B., de Vet, H.C.W., Bossuyt, P.M.M., 2016. STARD 2015

ACCEPTED MANUSCRIPT

guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 6. https://doi.org/10.1136/bmjopen-2016-012799

Cook, R.L., Chung, T., Kelly, T.M., Clark, D.B., 2005. Alcohol Screening in Young Persons Attending a Sexually Transmitted Disease Clinic: Comparison of AUDIT, CRAFFT, and CAGE Instruments. J. Gen. Intern. Med. 20, 1-6. https://doi.org/10.1111/j.1525-1497.2005.40052.x

Costa, D.S.J., 2015. Reflective, causal, and composite indicators of quality of life: A conceptual or an empirical distinction? Qual. Life Res. 24, 2057-2065. https://doi.org/10.1007/s11136-015-0954-2

CRD, 2009. Systematic reviews: CRD's guidance for undertaking reviews in health care. CRD, University of York, York. https://www.york.ac.uk/crd/guidance/

DeMartini, K.S., Carey, K.B., 2012. Optimizing the Use of the AUDIT for alcohol screening in college students. Psychol. Assess. 24, 954-963. https://doi.org/10.1037/a0028519

Dick, D.M., Aliev, F., Viken, R., Kaprio, J., Rose, R.J., 2011. Rutgers alcohol problem index scores at age 18 predict alcohol dependence diagnoses seven years later. Alcohol. Clin. Exp. Res. 35, 1011-1014. https://doi.org/10.1111/j.1530-0277.2010.01432.x

Dwamena, B.A., 2007. MIDAS: Stata module for meta-analytical integration of diagnostic accuracy studies. https://econpapers.repec.org/software/bocbocode/s456880.htm

Earleywine, M., LaBrie, J.W., Pedersen, E.R., 2008. A brief Rutgers alcohol problem index with less potential for bias. Addict. Behav. 33, 1249-1253. https://doi.org/10.1016/j.addbeh.2008.05.006

Edelen, M.O., McCaffrey, D.F., Ellickson, P.L., Tucker, J.S., Klein, D.J., 2009. Creating a developmentally sensitive measure of adolescent alcohol misuse: An application of item

ACCEPTED MANUSCRIPT

response theory. Subst. Use Misuse 44, 835-847. https://doi.org/10.1080/10826080802484686

Edwards, G., 2012. "The evil genius of the habit": DSM-5 seen in historical context. J. Stud. Alcohol Drugs 73, 699-701. https://doi.org/10.15288/jsad.2012.73.699

Edwards, G., Lader, M., Drummond, D.C., 1992. Problems and Dependence: The history of two dimensions, The nature of alcohol and drug related problems. Oxford Medical Publications, Oxford.

Ferreira, J.A., Martins, J.S., Coelho, M.S., Kahler, C.W., 2014. Validation of Brief Young Adult Alcohol Consequences Questionnaire (B-YAACQ): Portuguese Version. Span. J. Psychol. 17, E71. https://doi.org/10.1017/sjp.2014.74

Fleming, M.F., Barry, K.L., Macdonald, R., 1991. The Alcohol Use Disorders Identification Test (AUDIT) in a college sample. Int. J. Addict. 26, 1173-1185. https://doi.org/10.3109/10826089109062153

Gore, F.M., Bloem, P.J., Patton, G.C., Ferguson, J., Joseph, V., Coffey, C., Sawyer, S.M., Mathers, C.D., 2011. Global burden of disease in young people aged 10-24 years: A systematic analysis. Lancet 377, 2093-2102. https://doi.org/10.1016/S0140-6736(11)60512-6

Hagström, H., Hemmingsson, T., Discacciati, A., Andreasson, A., 2018. Alcohol consumption in late adolescence is associated with an increased risk of severe liver disease later in life. J. Hepatol. 68, 505-510. https://doi.org/10.1016/j.jhep.2017.11.019

Harbord, R., 2008. METANDI: Stata module to perform meta-analysis of diagnostic accuracy. https://EconPapers.repec.org/RePEc:boc:bocode:s456932

ACCEPTED MANUSCRIPT

Hurlbut, S.C., Sher, K.J., 1992. Assessing alcohol problems in college students. J. Am. Coll. Health 41, 49-58. https://doi.org/10.1080/07448481.1992.10392818

Kahler, C.W., Hustad, J., Barnett, N.P., Strong, D.R., Borsari, B., 2008. Validation of the 30-Day Version of the Brief Young Adult Alcohol Consequences Questionnaire for use in longitudinal studies. J. Stud. Alcohol Drugs 69, 611-615. https://doi.org/10.15288/jsad.2008.69.611

Kahler, C.W., Strong, D.R., Read, J.P., 2005. Toward efficient and comprehensive measurement of the alcohol problems continuum in college students: The Brief Young Adult Alcohol Consequences Questionnaire. Alcohol. Clin. Exp. Res. 29, 1180-1189. https://doi.org/10.1097/01.ALC.0000171940.95813.A5

Kelly, T.M., Donovan, J.E., Chung, T., Bukstein, O.G., Cornelius, J.R., 2009. Brief screens for detecting alcohol use disorder among 18–20 year old young adults in emergency departments: Comparing AUDIT-C, CRAFFT, RAPS4-QF, FAST, RUFT-Cut, and DSM-IV 2-Item Scale. Addict. Behav. 34, 668-674. https://doi.org/10.1016/j.addbeh.2009.03.038

Kelly, T.M., Donovan, J.E., Chung, T., Cook, R.L., Delbridge, T.R., 2004. Alcohol use disorders among emergency department–treated older adolescents: A new brief screen (RUFT-Cut) using the AUDIT, CAGE, CRAFFT, and RAPS-QF. Alcohol. Clin. Exp. Res. 28, 746-753. https://doi.org/10.1097/01.ALC.0000125346.37075.85

Kelly, T.M., Donovan, J.E., Kinnane, J.M., Taylor, D.M., 2002. A comparison of alcohol screening instruments among under-aged drinkers treated in emergency departments. Alcohol Alcohol. 37, 444-450. https://doi.org/10.1093/alcalc/37.5.444

ACCEPTED MANUSCRIPT

Knight, J.R., Sherritt, L., Harris, S.K., Gates, E.C., Chang, G., 2003. Validity of brief alcohol screening tests among adolescents: A comparison of the AUDIT, POSIT, CAGE, and CRAFFT. Alcohol. Clin. Exp. Res. 27, 67-73. https://doi.org/10.1111/j.1530-0277.2003.tb02723.x

Knight, J.R., Shrier, L.A., Bravender, T.D., Farrell, M., Vander Bilt, J., Shaffer, H.J., 1999. A new brief screen for adolescent substance abuse. Arch. Pediatr. Adolesc. Med. 153, 591-596. https://doi.org/10.1001/archpedi.153.6.591

Kokotailo, P.K., Egan, J., Gangnon, R., Brown, D., Mundt, M., Fleming, M., 2004. Validity of the alcohol use disorders identification test in college students. Alcohol. Clin. Exp. 28, 914-920. https://doi.org/10.1097/01.ALC.0000128239.87611.F5

Kontopantelis, E., Reeves, D., 2009. METAAN: Stata module to perform fixed- or random-effects meta-analyses. Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s457071.html

Lester, L., Baker, R., Coupland, C., Orton, E., 2018. Alcohol misuse and injury outcomes in young people aged 10–24. J. Adolesc. Health 62, 450-456. https://doi.org/10.1016/j.jadohealth.2017.10.003

Levy, S., Dedeoglu, F., Gaffin, J.M., Garvey, K.C., Harstad, E., MacGinnitie, A., Rufo, P.A., Huang, Q., Ziemnik, R.E., Wisk, L.E., Weitzman, E.R., 2016. A screening tool for assessing alcohol use risk among medically vulnerable youth. PLoS ONE 11, e0156240. https://doi.org/10.1371/journal.pone.0156240

López-Núñez, C., Fernández-Artamendi, S., Fernández-Hermida, J.R., Álvarez, Á.C., Secades-Villa, R., 2012. Spanish adaptation and validation of the Rutgers Alcohol Problem Index (RAPI). Int. J. Clin. Health Psychol. 12, 251-264.

ACCEPTED MANUSCRIPT

MacArthur, G.J., Smith, M.C., Melotti, R., Heron, J., Macleod, J., Hickman, M., Kipping, R.R., Campbell, R., Lewis, G., 2012. Patterns of alcohol use and multiple risk behaviour by gender during early and late adolescence: the ALSPAC cohort. J. Public Health 34, i20-i30. https://doi.org/10.1093/pubmed/fds006

Maimaris, W., McCambridge, J., 2014. Age of first drinking and adult alcohol problems: systematic review of prospective cohort studies. J. Epidemiol. Community Health 68, 268-274. https://doi.org/10.1136/jech-2013-203402

Manea, L., Boehnke, J.R., Gilbody, S., Moriarty, A.S., McMillan, D., 2017. Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis. BMJ Open 7. https://doi.org/10.1136/bmjopen-2016-015247

Martens, M.P., Neighbors, C., Dams-O'Connor, K., Lee, C.M., Larimer, M.E., 2007. The factor structure of a dichotomously scored Rutgers alcohol problem index. J. Stud. Alcohol Drugs 68, 597-606. https://doi.org/10.15288/jsad.2007.68.597

McCambridge, J., McAlaney, J., Rowe, R., 2011. Adult consequences of late adolescent alcohol consumption: A systematic review of cohort studies. PLoS Med. 8, e1000413. https://doi.org/10.1371/journal.pmed.1000413

McCambridge, J., Thomas, B.A., 2009. Short forms of the AUDIT in a Web-based study of young drinkers. Drug Alcohol Rev. 28, 18-24. https://doi.org/10.1111/j.1465-3362.2008.00010.x

McGee, R., Kypri, K., 2004. Alcohol-related problems experienced by university students in New Zealand. Aust. N. Z. J. Public Health 28, 321-323. https://doi.org/10.1111/j.1467-842X.2004.tb00437.x

ACCEPTED MANUSCRIPT

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. PLoS Med. 6, e1000097. https://doi.org/10.1371/journal.pmed.1000097

Neal, D.J., Corbin, W.R., Fromme, K., 2006. Measurement of alcohol-related consequences among high school and college students: Application of item response models to the Rutgers Alcohol Problem Index. Psychol. Assess. 18, 402-414. https://doi.org/10.1037/1040-3590.18.4.402

Niblett, P., 2017. Smoking, drinking and drug use among young people. England: 2016. NHS Digital, England.

NICE, 2010. Alcohol-use disorders: preventing the development of hazardous and harmful drinking. NICE public health guidance 24, England.

Northrup, T.F., Malone, P.S., Follingstad, D., Stotts, A.L., 2013. Using Item Response Theory to Improve Alcohol Dependence Screening for African American and White Male and Female College Students. Addict. Disord. Their Treat. 12, 99-109. https://doi.org/10.1097/ADT.0b013e3182627431

O'Hare, T., 2005. Comparing the AUDIT and 3 drinking indices as predictors of personal and social drinking problems in freshman first offenders. J. Alcohol Drug Educ. 49, 37-61.

O'Hare, T., Cohen, F.C., Sherrer, M.V., 1997. Comparing the QFI, the retrospective diary, and binge drinking in college first offenders. J. Alcohol Drug Educ. 42, 40-53.

O'Hare, T., Sherrer, M.V., 1999. Validating the Alcohol Use Disorder Identification Test With College First-Offenders. J. Subst. Abuse Treat. 17, 113-119. https://doi.org/10.1016/S0740-5472(98)00063-4

ACCEPTED MANUSCRIPT

Read, J.P., Kahler, C.W., Strong, D.R., Colder, C.R., 2006. Development and preliminary validation of the young adult alcohol consequences questionnaire. J. Stud. Alcohol Drugs 67, 169-177. https://doi.org/10.15288/jsa.2006.67.169

Read, J.P., Merrill, J.E., Kahler, C.W., Strong, D.R., 2007. Predicting functional outcomes among college drinkers: Reliability and predictive validity of the Young Adult Alcohol Consequences Questionnaire. Addict. Behav. 32, 2597-2610. https://doi.org/10.1016/j.addbeh.2007.06.021

Rumpf, H.J., Wohlert, T., Freyer-Adam, J., Grothues, J., Bischof, G., 2013. Screening Questionnaires for Problem Drinking in Adolescents: Performance of AUDIT, AUDIT-C, CRAFFT and POSIT. Eur. Addict. Res. 19, 121-127. https://doi.org/10.1159/000342331

Santis, R., Garmendia, M.L., Acuña, G., Alvarado, M.E., Arteaga, O., 2009. The Alcohol Use Disorders Identification Test (AUDIT) as a screening instrument for adolescents. Drug Alcohol Depend. 103, 155-158. https://doi.org/10.1016/j.drugalcdep.2009.01.017

Saunders, J.B., Aasland, O.G., Babor, T.F., de la Fuente, J.R., Grant, M., 1993. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption - II. Addiction 88, 791-804. https://doi.org/10.1111/j.1360-0443.1993.tb02093.x

Secretary-General, 1981. Secretary-General's Report to the General Assembly, A/36/215. United Nations, New York.

Smits, N., van der Ark, L.A., Conijn, J.M., 2018. Measurement versus prediction in the construction of patient-reported outcome questionnaires: can we have our cake and eat it? Qual. Life Res. 27, 1673-1682. https://doi.org/10.1007/s11136-017-1720-4.

ACCEPTED MANUSCRIPT

StataCorp, 2015. Stata Statistical Software: Release 14. StataCorp LP, College Station,TX.

Taylor, J., James, L.M., Bobadilla, L., Reeves, M.D., 2008. Screening for disinhibited disorder cases in a college population: Performance of the SMAST, DAST, SCID-II-Q, and PDQ-4. Psychol. Assess. 20, 351-360. https://doi.org/10.1037/a0013452

Terwee, C.B., Mokkink, L.B., Knol, D.L., Ostelo, R.W., Bouter, L.M., de Vet, H.C., 2012. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual. Life Res. 21, 651-657. https://doi.org/10.1007/s11136-011-9960-1

Thomas, B.A., McCambridge, J., 2008. Comparative psychometric study of a range of hazardous drinking measures administered online in a youth population. Drug Alcohol Depend. 96, 121-127. https://doi.org/10.1016/j.drugalcdep.2008.02.010

Toner, P., Böhnke, J.R., McCambridge, J., 2017. A systematic review of alcohol screening and assessment measures for young people: a study protocol. BMJ Open 7. https://doi.org/10.1136/bmjopen-2017-016406

Townshend, T.G., 2013. Youth, alcohol and place-based leisure behaviours: A study of two locations in England. Soc. Sci. Med. 91, 153-161. https://doi.org/10.1016/j.socscimed.2013.02.017

Van den Bruel, A., Aertgeerts, B., Hoppenbrouwers, K., Roelants, M., Buntinx, F., 2004. CUGE: a screening instrument for alcohol abuse and dependence in students. Alcohol Alcohol. 39, 439-444. https://doi.org/10.1093/alcalc/agh077

Verster, J.C., Van Herwijnen, J., Olivier, B., Kahler, C.W., 2009. Validation of the Dutch version of the brief young adult alcohol consequences questionnaire (B-YAACQ). Addict. Behav. 34, 411-414. https://doi.org/10.1016/j.addbeh.2007.09.013

ACCEPTED MANUSCRIPT

West, S.L., Graham, C.W., 2001. Assessing the validity of college alcohol problem scale with African American undergraduates: Results of preliminary investigation. J. Alcohol Drug Educ. 46, 4-17.

Whiting, P.F., Rutjes, A.S., Westwood, M.E., Mallett, S., Deeks, J.J., Reitsma, J.B., Leeflang, M.M., Sterne, J.A., Bossuyt, P.M.; QUADAS-2 Group, 2011. Quadas-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann. Intern. Med. 155, 529-536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Figure Legend

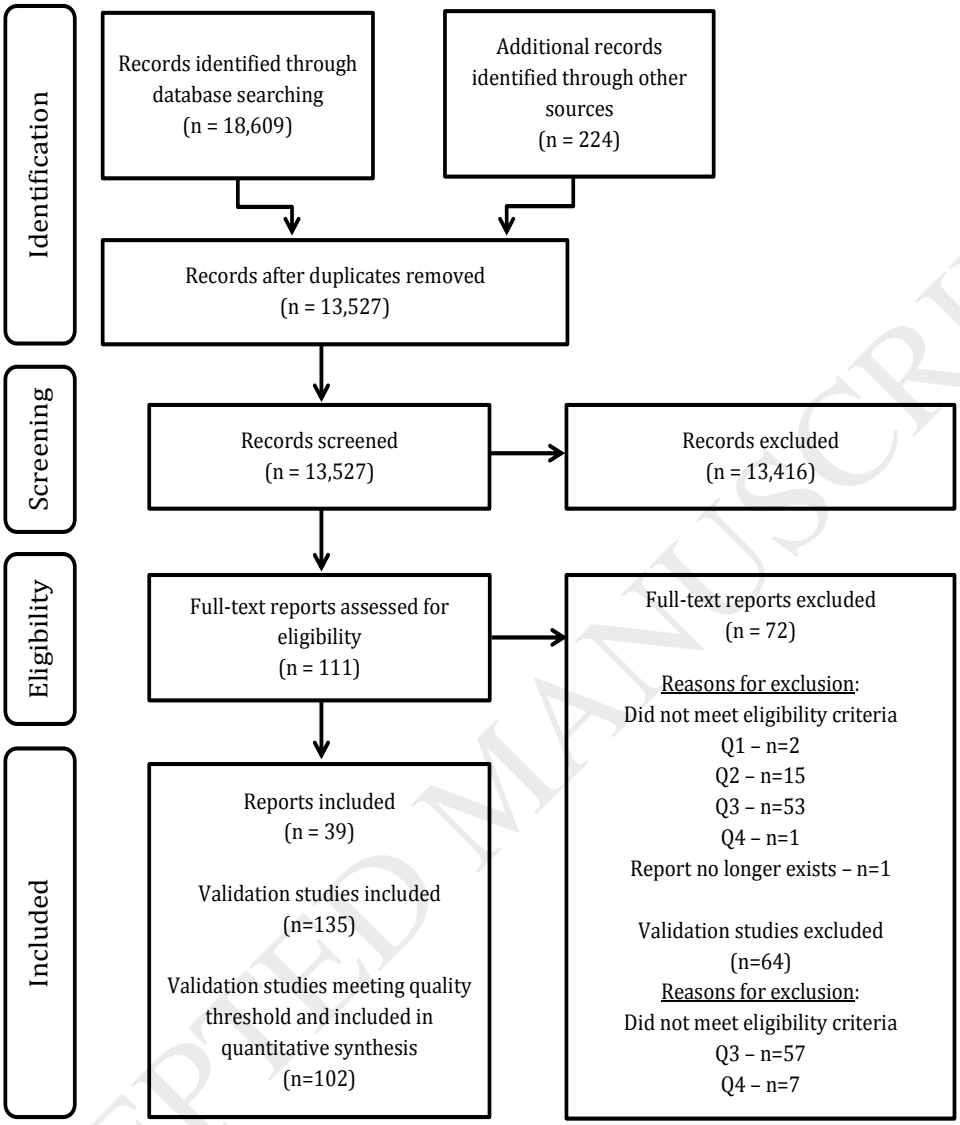
Figure 1. Search results and study selection flowchart.

Figure 2. Panel A: Aggregated ROC curve for $n = 53$ studies (grey circles) of screening instruments reporting diagnostic accuracy data; dashed horizontal line presenting the sensitivity inclusion criterion ($\geq$ .80). Panel B: Probability of using alcohol in a hazardous way depending on prevalence for positive (long dashed) and negative (short dashed) test results.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Figure 1.

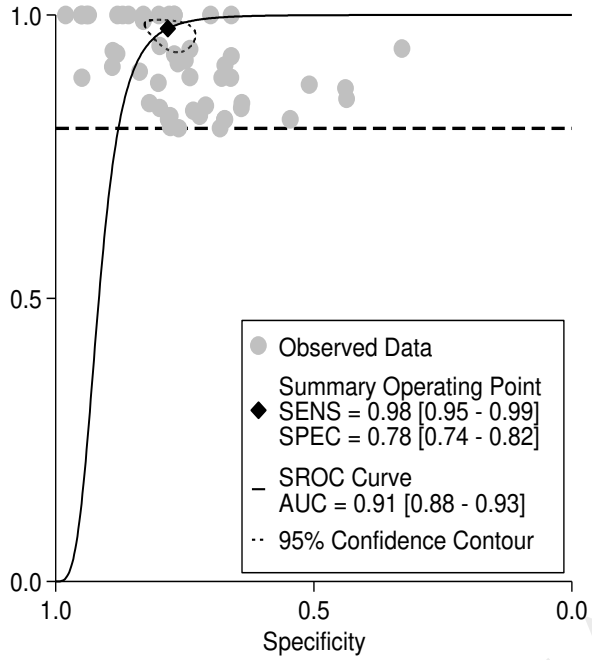


ACCEPTED MANUSCRIPT

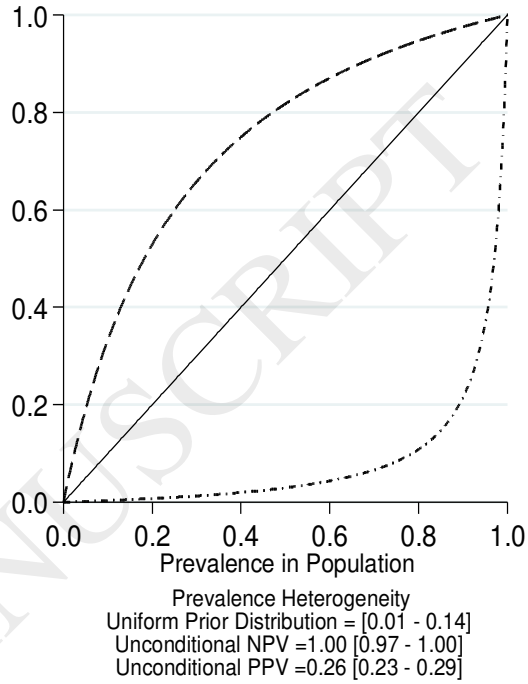
ACCEPTED MANUSCRIPT

Figure 2.

Panel A



Panel B



ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 1. Characteristics of included studies above quality thresholds.

Report	Country	Setting	Sample size	Mean Age	% Female	Ethnicity[a]	Index Tests	Reference Tests
Aertgeerts et al. (2000)	Belgium	University	3,564	18	54.4	Nationality data only	CUGE AUDIT	CIDI
Chung et al. (2002)	USA	Emergency Department	173	16.4	43	72% White	Modified AUDIT	DISC
Chung et al. (2000)	USA	Emergency Department	261	16.1	42	71% White	Modified TWEAK Modified AUDIT	DISC
Chung et al. (2012)	USA	Population Survey	166,165 (stratified by age 12-18)	N/A	48.6	62.3% White	Frequency Item Quantity Item HED Frequency Item	11 questions DSM-IV symptoms
Clark et al. (2006)	USA	Community Sample	219	16	52	81% White	Frequency Item	Modified SCID
Clark et al. (2016)	USA	Primary Care	1,193 (stratified by age 12-20)	N/A	57	93.4% White	Frequency Item Quantity Item Quantity x Frequency (QxF) Items	The National Survey on Drug Use and Health
Cook et al. (2005)	USA	STD Clinic	358	20.6	45	49% Black	AUDIT CRAFFT	Modified SCID
DeMartini & Carey (2012)	USA	University	401	19	54	64% White	AUDIT AUDIT-C	Daily Drinking Questionnaire
Edelen et al. (2009)	USA	School	5,828	12.4	48	86% White	Alcohol Misuse Items	AUDIT
Ferreira et al. (2014)	Portugal	University	560	20.6	68.8	Not reported	B-YAACQ	AUDIT
Fleming et al. (1991)	USA	University	989	20.5	69.5	93% White	AUDIT	DIS

ACCEPTED MANUSCRIPT

Report	Country	Setting	Sample size	Mean Age	% Female	Ethnicity[a]	Index Tests	Reference Tests
Hurlbut & Sher (1992)	USA	University	490	18.2	50	Not reported	YAAPST	DIS
Kahler et al. (2008)	USA	University	291	19	35	96% White	B-YAACQ	AUDIT
Kahler et al. (2005)	USA	University	126	Not reported – 19 total sample	Not reported – 51.8 total sample	Not reported – 84% White total sample	YAACQ B-YAACQ	RAPI
Kelly et al. (2009)	USA	Emergency Department	181	Not reported –18-20	43	Not reported	AUDIT-C AUDIT (FAST) RAPS4-QF RUFT-Cut	Modified SCID
Kelly et al. (2004)	USA	Emergency Department	93	19	45	81% White	AUDIT CRAFFT RAPS-QF RUFT-Cut	Modified SCID
Kelly et al. (2002)	USA	Emergency Department	103	17.5	46.6	79.5% White	Modified TWEAK CAGE	AUDIT
Knight et al. (2003)	USA	Young Adult Hospital	538	Not reported – 14-18	68	51% Black	AUDIT CAGE	ADI
Kokotailo et al. (2004)	USA	University Health Clinic	302	20.3	61.3	90% White	AUDIT	CIDI-SAM
Levy et al. (2016)	USA	Children's Hospital	118	Not reported	52.5	78% White	NIAAA Youth Alcohol Screen	DISC

ACCEPTED MANUSCRIPT

Report	Country	Setting	Sample size	Mean Age	% Female	Ethnicity[a]	Index Tests	Reference Tests
López-Núñez et al. (2012)	Spain	School	569	16.8	45.9	Nationality data only	RAPI	Items based on the DSM-IV-TR criteria
McGee & Kypri (2004)	New Zealand	University	1,464	Not reported – 20.5 total sample	57.6	Not reported	AREAS	AUDIT-C
Northrup et al. (2013)	USA	University	1,500	19.4	68	81% White	AUDIT-C Modified AUDIT-P	SSAGAII
O'Hare (2005)	USA	University	389	18.2	36.8	95.6% White	Binge Drinking	AUDIT
O'Hare et al. (1997)	USA	University	197	18.7	41.6	89.8% White	Binge Drinking	MmMAST
O'Hare & Sherrer (1999)	USA	University	312	18.6	35.9	91.3% White	AUDIT-C AUDIT-P	CAPS
Read et al. (2007)	USA	University	92	19.1	52	72% White	YAACQ	AUDIT YAAPST
Rumpf et al. (2013)	Germany	School	225	15.5	50.7	Not reported	AUDIT AUDIT-C CRAFFT POSIT	M-CIDI
Santis et al. (2009)	Chile	School	95	15.9	44.2	Not reported	AUDIT	CIDI-SAM
Thomas & McCambridge (2008)	UK	Online Survey	167	20.3	70	86% White	APS AREAS RAPI LDQ SDS	AUDIT

ACCEPTED MANUSCRIPT

Report	Country	Setting	Sample size	Mean Age	% Female	Ethnicity[a]	Index Tests	Reference Tests
Van den Bruel et al. (2004)	Belgium	University	2,699	18.7	59.6	Nationality data only	CUGE	CIDI
Verster et al. (2009)	Netherlands	University	667	20.5	72.4	Not reported	B-YAACQ	AUDIT-PC
West & Graham (2001)	USA	University	33	19	70	100% Black	CAPS	DSM-IV Criteria

[a]Most prevalent ethnicity in study samples reported.

Instrument names: APS – The Alcohol Problems Scale; AREAS – The Academic Role Expectations and Alcohol Scale; AUDIT – Alcohol Use Disorder Identification Test; AUDIT-C – AUDIT Consumption subscale; AUDIT-P – AUDIT Problems subscale; CAGE – Concern/Cut-down, Anger, Guilt, and Eye-opener; CAPS – College Alcohol Problems Scale; CUGE – Concern/Cut-down, Under Influence, Guilt, and Eye-opener; CRAFFT – Car, Relax, Alone, Forget, Friends, Trouble; DSM – Diagnostic and Statistical Manual of Mental Disorders; FAST – Fast Alcohol Screening Test ; LDQ – Leeds Dependence Questionnaire; NIAAA – National Institute on Alcohol Abuse and Alcoholism; POSIT – Problem Oriented Screening Instrument for Teenagers; RAPI – Rutgers Alcohol Problem Index; RAPS4-QF – Rapid Alcohol Problems Screen – Quantity Frequency; RUFT-Cut – Riding with intoxicated driver, Unable to stop, Family/Friends, Trouble, Cut down; SDS – The Severity of Dependence Scale; TWEAK – Tolerance, Worried, Eye-opener, Amnesia, K/Cut down; YAACQ – Young Adult Alcohol Consequences Questionnaire; YAAPST – Young Adult Alcohol Problems Screening Test.

ACCEPTED MANUSCRIPT

Table 2. Validation studies for screening items/instruments.

Screening instruments	Validation studies (under thresholds)	Reference tests (for studies under thresholds)	Validation studies (over thresholds)	Reference tests (for studies over thresholds)
Alcohol Frequency	0		18	18 DSM (14 same report and 3 same report)
Alcohol Quantity	0		10	10 DSM (7 same report and 3 same report)
HED Frequency	2	2 DSM	5	5 DSM (5 same report)
AUDIT: Items 1-10	2	1 CAPS:SE; 1 CAPS:CS	10	8 DSM; 1 DDQ; 1 TLFB
AUDIT (Modified): Items 1-10	0		2	2 DSM
AUDIT-C: Items 1-3	0		10	6 DSM (4 same report); 2 DDQ (2 same report); 1 CAPS:SE, 1 CAPS:CS
AUDIT-C (Modified): Items 1-3	0		1	1 DSM
AUDIT: Items 4,5,6,7,8,10	0		4	4 DSM (same report)
AUDIT: Items 4-10	0		2	1 CAPS:SE, 1 CAPS:CS
AUDIT (Modified): Items 4-10	0		1	1 DSM
AUDIT (FAST) Items 3,5,8,10	0		1	1 DSM
CRAFFT	0		3	3 DSM
Quantity-Frequency (QF)	0		3	3 DSM (3 same report)
CUGE	0		2	2 DSM
RAPS4-QF	0		2	2 DSM
RUFT-Cut	0		2	2 DSM
Screening instruments	**Validation studies (under thresholds)**	**Reference tests (for studies under thresholds)**	**Validation studies (over thresholds)**	**Reference tests (for studies over thresholds)**

[a]Short version validated by parent instrument.

ACCEPTED MANUSCRIPT

CAGE	3	3 DSM	2	1 DSM; 1 AUDIT
CAGE (Modified)	1	1 DSM	0	
Binge Drinking	2	1 CAPS:SE; 1 CAPS:CS	2	1 AUDIT; 1 MmMAST
TWEAK (Modified)	1	1 CAGE	2	1 DSM; 1 AUDIT
POSIT - Substance Use/Abuse Scale	0		1	1 DSM
Heavy Drinking (QFI)	4	1 AUDIT; 1 MmMAST;	0	
Alcohol Change Index (ACI)	3	1 CAPS:SE; 1 CAPS:CS; 1 AUDIT;	0	
Peak Drinking (RD)	1	1 CAPS:SE; 1 CAPS:CS 1 MmMAST	0	
SMAST	1	1 DSM	0	
Timeline Follow-Back (TLFB)	1	1 AUDIT	0	
[a]AUDIT (Brief): Items 3,5,8	1	[a]1 AUDIT	0	
[a]DSM-IV-2	1	[a]1 DSM	0	

ACCEPTED MANUSCRIPT

Table 3. Validation studies for assessment instruments.

Assessment instruments	Validation studies (under thresholds)	Reference tests (for studies under thresholds)	Validation studies (over thresholds)	Reference tests (for studies over thresholds)
YAACQ	0		4	1 RAPI; (1 AUDIT; 1 YAAPST; 1 YAAPST-D same report)
[a]B-YAACQ	3	1 DDQ-R; 1 TLFB; [a]1 YAACQ	4	2 AUDIT; 1 AUDIT-PC; 1 RAPI
RAPI	5	1 DSM; 3 DDQ; 1 DDQ-R	2	1 DSM; 1 AUDIT
AREAS	0		2	1 AUDIT; 1 AUDIT-C
APS	1	1 AUDIT-C	1	1 AUDIT
Alcohol Misuse Items	0		1	1 AUDIT
CAPS	0		1	1 DSM
LDQ	0		1	1 AUDIT
SDS	0		1	1 AUDIT
YAAPST	0		1	1 DSM
[a]S-RAPI	1	[a]1 RAPI	0	

[a]Short version validated by parent instrument.

ACCEPTED MANUSCRIPT

Table 4. Aggregate statistics for diagnostic criteria of screening items/instruments.

Screening instruments	Reports[a] / Studies[b]	Average Sensitivity[c]	Average Specificity[c]	Average COSMIN	Average Reliability[d,e]
Alcohol Frequency	3 / 18	1.00	.84	2.00 (SD = 0)	*n* = 0[f]
Alcohol Quantity	2 / 10	.96	.91	2.00 (SD = 0)	*n* = 0[f]
AUDIT: Items 1-10	9 / 10	.83	.70	1.94 (SD = .19)	.80 *n* = 6
AUDIT-C: Items 1-3	5 / 10	.83	.70	2.00 (SD = 0)	.92 *n* = 4

Note: The total number of available reports/studies in the review is provided in column 2 for each row; in the subsequent columns *n* indicates the studies reporting each statistic if deviating from the total. [a]Full-text papers. [b]Discrete validation studies contained within reports. [c]Determined with Stata V.14, midas (Dwamena, 2007). [d]Determined with Stata V.14, metaan (Kontopantelis and Reeves, 2009) [e]Calculated for reliability adjusted for 10 items. [f]Single item instruments, no reliability estimate was reported. Although there were five validation studies for HED frequency, all estimates were from a single report (Chung et al., 2012) and it was only possible to report an average COSMIN rating of 2.