3-2015

Project INTEGRATE: An Integrative Study of Brief Alcohol Interventions for College Students

Eun-Young Mun
Center of Alcohol Studies, Rutgers, The State University of New Jersey

Jimmy de la Torre
Rutgers, The State University of New Jersey

David C. Atkins
University of Washington

Helene R. White
Center of Alcohol Studies, Rutgers, The State University of New Jersey

Anne E. Ray
Center of Alcohol Studies, Rutgers, The State University of New Jersey

See next page for additional authors

Authors

Eun-Young Mun, Jimmy de la Torre, David C. Atkins, Helene R. White, Anne E. Ray, Su-Young Kim, Yang Jiao, Nickeisha Clarke, Yan Huo, Mary E. Larimer, David Huh, and The Project INTEGRATE Team

Published in final edited form as:

Psychol Addict Behav. 2015 March ; 29(1): 34–48. doi:10.1037/adb0000047.

Project INTEGRATE: An Integrative Study of Brief Alcohol Interventions for College Students

Eun-Young Mun[1], **Jimmy de la Torre**[2], **David C. Atkins**[3], **Helene R. White**[1], **Anne E. Ray**[1], **Su-Young Kim**[1,4], **Yang Jiao**[1], **Nickeisha Clarke**[1], **Yan Huo**[2], **Mary E. Larimer**[3], **David Huh**[3], and **The Project INTEGRATE Team**[5]

[1] Center of Alcohol Studies, Rutgers, The State University of New Jersey

[2] Department of Educational Psychology, Rutgers, The State University of New Jersey

[3] Department of Psychiatry and Behavioral Sciences, The University of Washington

[4] Department of Psychology, Ewha Womans University

Abstract

This paper provides an overview of a study that synthesizes multiple, independently collected alcohol intervention studies for college students into a single, multisite longitudinal data set. This research embraced innovative analytic strategies (i.e., integrative data analysis or meta-analysis using individual participant-level data), with the overall goal of answering research questions that are difficult to address in individual studies such as moderation analysis, while providing a built-in replication for the reported efficacy of brief motivational interventions for college students. Data were pooled across 24 intervention studies, of which 21 included a comparison or control condition and all included one or more treatment conditions. This yielded a sample of 12,630 participants (42% men; 58% first-year or incoming students). The majority of the sample identified as White (74%), with 12% Asian, 7% Hispanic, 2% Black, and 5% other/mixed ethnic groups. Participants were assessed two or more times from baseline up to 12 months, with varying assessment schedules across studies. This paper describes how we combined individual participant-level data from multiple studies, and discusses the steps taken to develop commensurate measures across studies via harmonization and newly developed Markov chain Monte Carlo algorithms for two-parameter logistic item response theory models and a generalized partial credit model. This innovative approach has intriguing promises, but significant barriers exist. To lower the barriers, there is a need to increase overlap in measures and timing of follow-up assessments across studies, better define treatment and control groups, and improve transparency and documentation in future single, intervention studies.

Correspondence concerning this article should be addressed to Eun-Young Mun, Center of Alcohol Studies, Rutgers, The State University of New Jersey, 607 Allison Road, Piscataway, NJ 08854. eymun@rci.rutgers.edu.

[5]The Project INTEGRATE Team consists of the following contributors in alphabetical order: John S. Baer, Department of Psychology, The University of Washington, and Veterans' Affairs Puget Sound Health Care System; Nancy P. Barnett, Center for Alcohol and Addiction Studies, Brown University; M. Dolores Cimini, University Counseling Center, The University at Albany, State University of New York; William R. Corbin, Department of Psychology, Arizona State University; Kim Fromme, Department of Psychology, The University of Texas, Austin; Joseph W. LaBrie, Department of Psychology, Loyola Marymount University; Matthew P. Martens, Department of Educational, School, and Counseling Psychology, The University of Missouri; James G. Murphy, Department of Psychology, The University of Memphis; Scott T. Walters, Department of Behavioral and Community Health, The University of North Texas Health Science Center; and Mark D. Wood, Department of Psychology, The University of Rhode Island.

Keywords

Integrative Data Analysis; Meta-analysis; Brief Motivational Interventions; Alcohol Interventions; College Students

This paper provides an overview of a collaborative study entitled Project INTEGRATE. Project INTEGRATE is the first behavioral treatment research project to embrace recent advances in psychometrics and statistical methods (e.g., meta-analysis using individual participant-level data [IPD] or integrative data analysis [IDA]). The overall goals are to provide answers to evasive research questions (e.g., identification of mediational paths and subgroup differences), as well as to provide a built-in replication for the reported efficacy of brief motivational interventions for college student populations. The term IDA was coined by Curran and Hussong (2009) to highlight some of the unique promises, as well as challenges, that arise when combining studies in the psychological sciences. The term meta-analysis using IPD has been utilized more frequently in evaluating randomized control trials (RCTs) in medical research. We interchangeably use IDA and meta-analysis using IPD (or IPD meta-analysis) in the present article. This paper does not report clinical treatment outcomes. Rather, we provide an overview of this research project and discuss the challenges encountered, steps taken to overcome these challenges, and lessons learned thus far. This overview sets the stage for papers that focus on clinical outcomes and mechanisms of behavior change to follow.

Available reviews of brief motivational interventions (BMIs) for college students have documented that BMIs (e.g., the Brief Alcohol Screening and Intervention for College Students [BASICS]; Dimeff, Baer, Kivlahan, & Marlatt, 1999) are effective in reducing alcohol use and related problems at least on a short-term basis (Carey, Scott-Sheldon, Carey, & DeMartini, 2007; Cronce & Larimer, 2011). Furthermore, those delivered in-person provide more enduring effects compared to computer-delivered feedback interventions, including computer-delivered, normative feedback interventions and computer-delivered, educational alcohol interventions (Carey, Scott-Sheldon, Elliott, Garey, & Carey, 2012). However, the estimated effect sizes of these brief interventions are fairly small (e.g., Cohen's *d* ranging from 0.04 to 0.21 from random-effects models for outcome variables at short-term [4-13 weeks post intervention] follow-up of individually-delivered interventions; Carey et al., 2007), and vary from study to study across key outcome variables, such as alcohol use and alcohol-related problems. Furthermore, only a small subset of studies had a statistically significant effect when reanalyzed in a meta-analysis (Carey et al., 2007). Thus, there appears to be incongruence in the strength of the overall effect between single studies and meta-analysis studies.

Emerging evidence suggests that single studies may be more susceptible to biased statistical inference than previously thought. For example, recent meta-analytic studies examining the efficacy of anti-depressant medication aptly demonstrate the potential pitfalls of relying on evidence only from single studies. Turner, Matthews, Linardatos, Tell, and Rosenthal (2008) meta-analyzed aggregated data (AD; e.g., effect size estimates) on anti-depressant medication submitted to the Food and Drug Administration (FDA) and in published articles

from 74 trials (12 drugs and 12,564 patients) that were registered with the FDA between 1987 and 2004. Their analyses indicated that effect sizes had been substantially overestimated in published articles. For example, whereas 94% of the 37 published studies reported a significant positive result, only 51% had a positive outcome according to the meta-analysis of the FDA data. On average, Turner et al. found a 32% difference in effect sizes between the FDA data and the published data. Moreno et al. (2009) further showed that this false positive outcome bias was associated with publications, and found that deviations from study protocol, such as switching from an intent-to-treat analysis to a per-protocol analysis (i.e., excluding dropouts and/or those who did not adhere to treatment protocol), accounted for some of the discrepancies between the FDA and published data. Subsequent meta-analyses examined this controversy further. Fournier et al. (2010) obtained raw, individual participant-level data (i.e., IPD) from six of the 23 short-term RCTs of anti-depressant medication (a total of 718 patients). Using IPD, these authors found that anti-depressant drugs were minimally effective for patients with mild or moderate depressive symptoms (Cohen's $d = 0.11$), but their effects were better for those with severe ($d = 0.17$) or very severe ($d = 0.47$) depression. The controversy regarding the efficacy of anti-depressant medication illustrates that quantitative synthesis, especially utilizing IPD, can play a unique role in drawing unbiased and robust inference in treatment research.

Unfortunately, controversies like this are not limited to pharmaceutical clinical trials. A recent review of meta-analytic studies published in psychological journals also reveals a clear publication bias (Bakker, van Dijk, & Wicherts, 2012). Bakker et al. demonstrated in a simulation study that it is easier to find inflated and statistically significant effects in underpowered samples than larger and more powerful samples, especially when the true effect size is small. This may be because smaller samples capitalize on chance variations in effect sizes (Tversky & Kahneman, 1971) and also because questionable research practices (e.g., failing to report data on all outcomes) make it more likely to discover statistically significant effects. This may explain the paradox where typical psychological studies are underpowered; yet 96% of all papers in the psychological literature report statistically significant outcomes (Bakker et al., 2012). Overall, there is evidence of generally larger effects in smaller, compared to larger, studies (Borenstein, Hedges, Higgins, & Rothstein, 2009, p. 291; see also Kraemer, Mintz, Noda, Tinklenberg, & Yesavage, 2006).

In sum, findings from single studies may not provide sufficient, unbiased evidence as to the true magnitude of the effect of an intervention and the extent to which the effect can be applied (Ioannidis, 2005). In addition, published findings in the biomedical, as well as psychological, research fields have poor replicability (Begley & Ellis, 2012; Ioannidis, 2005; Nosek, Spies, & Motyl, 2012). Given that serious negative implications are associated with such poor reproducibility, calls have been made to raise standards for clinical trials (Begley & Ellis, 2012) and psychological research in general (Simmons, Nelson, & Simonsohn, 2011), as well as to improve transparency in reporting methodology and findings (Schulz, Altman, Moher, & CONSORT Group, 2010; Tse, Williams, & Zarin, 2009). Accordingly, integrative studies synthesizing IPD may be one promising alternative to a large-scale, multisite RCT.

Project INTEGRATE: Data and Design

Project INTEGRATE was motivated to overcome limitations of single studies and AD meta-analyses via pooling IPD from multiple, college alcohol intervention trials. More specifically, Project INTEGRATE was developed to examine (1) whether BMIs are efficacious for bringing about changes in theory-based behavior targets, such as normative perceptions about peer alcohol use and the use of protective behavioral strategies while drinking; (2) whether positive changes in behavior targets predict greater reductions in alcohol use and negative consequences; (3) whether subsets of interventions are more promising; and (4) whether subgroups exist for whom different interventions are more efficacious.

The present paper (1) provides a summary of the Project INTEGRATE data and its unique design characteristics; (2) describes how we established commensurate measures across studies; and (3) discusses lessons learned and offers practical recommendations for single intervention studies. Once commensurate measures across studies are established, the stated project goals can be examined using a number of appropriate analytical methods. Thus, this paper does not delve into any specific analytical models, as they would depend on the research questions being examined.

A group of investigators who had published studies assessing the efficacy of BMIs for college students were contacted in the spring of 2009, asking for their willingness to contribute their deidentified data. All but one agreed, resulting in a total of 24 studies (Studies 1 through 7, 8a through 8c, and 9 through 22; see Table 1 and the Online Supplement). Note that Studies 7 and 10 are single studies each with two distinct subsamples. In addition to examining BMIs, all 24 studies sampled college or university students in the United States, and assessed alcohol use outcome measures. Existing review studies provide some perspective about our sample of 24 studies as it relates to the body of work on college alcohol BMIs as a whole. Larimer and Cronce (2002, 2007) and Cronce and Larimer (2011) systematically searched the literature covering the combined period from 1984 to early 2010 on individual-focused preventative intervention studies, and summarized results from a combined total of 110 studies, of which approximately a third came from the last three years (2007-2010). Similarly, Carey et al. (2007) meta-analyzed data from 62 studies that focused on individual-level interventions published between 1985 and early 2007. Thus, the sample of 24 studies included in Project INTEGRATE represents a good proportion of the existing BMIs conducted between 1990 and 2009 (published between 1998 and 2010). These studies are diverse in terms of original investigators, college campuses from which participants were recruited, demographic characteristics, and intervention study designs. Our combined data set also includes data from unpublished studies (Studies 8b, 8c, and 9) and unreported data from published studies (e.g., additional cohorts; Study 20). Investigators who contributed data provided clarifications about study design and data, documentation, and intervention content for their studies.

Combined Sample

Data pooled from all 24 studies consisted of 12,630 participants. All studies included one or more BMI conditions, with the majority (21 studies) including either a control condition or

other comparison condition (i.e., alcohol education). Because condition labels varied across studies, we relabeled them based on shared intervention characteristics to one of the following five categories for Project INTEGRATE (Ray et al., in press): motivational interview plus personalized feedback (MI + PF, $n = 10$), stand-alone personalized feedback (PF, $n = 11$), group motivational interview (GMI, $n = 11$), alcohol education (AE, $n = 6$), and control ($n = 19$). There were three unique conditions that did not fit these categories: an MI + PF combined with an AE intervention, an MI without PF, and an MI + PF combined with a parent-based intervention (see Table S1 for all 60 intervention groups and their new labels included in Project INTEGRATE). Participant recruitment and selection also varied across studies, ranging from volunteer students recruited with flyers to students who were required to complete an alcohol program because they violated university rules about alcohol. Although some studies (i.e., Studies 8a, 8b, 8c, 10, 20, and 22) had assessments beyond 12 months post baseline, we decided to focus only on follow-up data up to a year, as there was a considerable lack of overlap in timing of assessments beyond this point. Each study assessed participants at least twice from baseline up to 12 months. More details on participant characteristics, assessment schedules, and classification of study conditions can be found in Table 1.

More than half of the combined sample is comprised of women (58%) and first-year or incoming students (58%). The majority of the sample is White (74%), with 12% Asian, 7% Hispanic, 2% Black, and 5% belonging to other or mixed ethnic groups. Approximately 15% are college students mandated to complete a university program as a result of alcohol-related infractions; 27% are members (or pledged to be a member) of fraternities and sororities; and 13% are varsity athletes or members of club sports. Two studies of mandated students (Studies 2 and 7.1) utilized a waitlist control within the 12-month follow-up period. To preserve the original randomized group assignment at baseline, we excluded data from those control cases who were waitlisted initially at baseline and received an intervention at a time that the follow-up assessment took place for other treatment groups (i.e., 119 from Study 2 at the 6-month follow-up; 24 from Study 7.1 at the 6-month follow-up; see Table 1). The majority of the individual studies included in Project INTEGRATE have been previously described in the published literature. Additional study details that were not described previously are provided in the Online Supplement.

In addition to this combined intervention data set, there were additional participants who were not part of the original intervention studies. Adding these screening or nonrandomized participants resulted in a total of 24,336 participants (60% women; 48% first-year or incoming students). This larger data set was used for item response theory (IRT) analysis and sensitivity analysis, as well as for research questions that did not involve intervention efficacy (e.g., racial and gender differences in alcohol-related problems; Clarke, Kim, White, Jiao, & Mun, 2013).

Study Design Characteristics and Analytic Considerations

IDA studies can be developed for specific research questions and there are a number of appropriate analytical approaches that can be utilized, depending on the nature of those questions as well as characteristics of the pooled data itself. Nonetheless, a discussion of

some of the challenges and our counter measures to overcome them may be helpful for other IDA studies. The Project INTEGRATE data have a three-level data structure where multiple repeated assessments are nested within individuals who are nested within studies. If no adjustment is made, any resulting standard error from the nested data tends to be underestimated and the power to detect any effects tends to be overestimated. This nested, correlated data structure can be measured using an intraclass correlation coefficient (ICC). Although the study ICC may be relatively small in our pooled data, the average cluster size (i.e., study sample size) is large, and the design effect, which is estimated as 1 + ICC * (average cluster size –1), can be substantial. In one analysis of a subsample, ICCs were small, ranging from 0.05 to 0.26 but the design effects were huge, ranging from 34.6 to 166.0, due to the large average cluster size (N = 648). To address this issue, we can use a sandwich variance estimator (see Hardin, 2003 for a review) suited for complex survey data. In conjunction with complex survey analysis, we can weight or scale data at the individual level (e.g., by using a weight of 1 over the square root of the sample size of each study) to prevent large studies from exerting overly dominating influences on overall estimates (see Table 1 for discrepant sample sizes across studies). In principle, large studies contribute more information and should count more toward estimates. However, a weighting strategy like this places slightly higher value on individuals' information from smaller studies relative to individuals' information from larger studies. An alternative approach is to utilize the multilevel modeling framework using either fixed-effects or random-effects models, which weight data differently when combining effects across studies (see DerSimonian & Laird, 1986) due to different assumptions involved in each modeling approach. This multilevel modeling approach can also accommodate weights although the best practice may differ for each research project. Both complex survey analysis and multilevel analysis can readily be implemented by using commercially available software programs.

Project INTEGRATE: Measures

One of the most important challenges in conducting IDA or meta-analysis using IPD is to ensure that measures are comparable across studies (Cooper & Patall, 2009; Curran & Hussong, 2009; Hussong, Curran, & Bauer, 2013). To address this issue, we utilized harmonization and developed innovative item response theory (IRT) models. Table S2 in the Online Supplement provides a list of our key constructs and overlap across studies, as well as the approach taken for each construct. For IRT analysis, some harmonization steps were taken to find common response options or to derive items that could be collapsed and linked across studies. Note that the overlap in measures across studies was excellent at the level of construct, but not at the item level. Within each study, most of the conceptual mediator variables were assessed at the same time as outcome measures.

Hierarchical Item Response Theory Approaches

When a construct was assessed using multiple items or scales that are well established in the literature, and when there was a subset of construct items that could be linked across studies, we conducted IRT analysis to obtain commensurate measures across studies. IRT or latent trait theory (Lazarsfeld & Henry, 1968; Lord & Novick, 1968) has been used extensively in the area of educational testing and measurement, and with increasing frequency in

psychological research (e.g., Gibbons et al., 2012). Unlike classical test theory, in the IRT framework, item parameters are independent of parameters describing individuals (or studies), which is a critical advantage for the current project, for which item subsets vary by individual and by study. Given the unique qualities of the Project INTEGRATE data, existing IRT methods were extended to handle sparse data, take into account study-level information (e.g., different trait means across studies), and borrow information, when possible, from related or higher-order dimensions. More specifically, we developed several IRT models adapted from hierarchical, multi-unidimensional, as well as unidimensional, two-parameter logistic IRT (2-PL IRT) models, and developed Markov chain Monte Carlo (MCMC) algorithms to fit these IRT models within a hierarchical Bayesian perspective. Huo et al. (2014) provides the theoretical and technical details of the 2-PL IRT models and MCMC algorithms, as well as the findings of two simulation studies and real data analysis. The MCMC codes were written in Ox (Doornik, 2009), a matrix-based, object-oriented programming language, and are available upon request.

Scoring of latent trait scores across time—For each construct, item parameters were calibrated using baseline data, and these calibrated item parameters were then used to estimate latent trait scores for baseline and subsequent follow-up data. Prior to longitudinal scoring, we checked whether different items were assessed at different time points, and whether different sets of items used at different time points could have introduced bias in our estimation of latent trait scores. Furthermore, not all individuals assessed at baseline were followed up, either by study design or due to attrition. Therefore, we conducted sensitivity analyses by recalibrating data using different sets of items and different subsets of participants across time. We compared the descriptive statistics (e.g., means and standard deviations) of the estimated item parameters, structural parameters, and trait scores by using different sets of items calibrated and checked their correlations (r = 0.99), which led us to conclude that the differences in items and participants over time did not exert any meaningful influence on our estimates. Below, we give examples of how latent trait scores – often called ‘theta (θ) scores’ in IRT – were established for two key constructs.

Alcohol-related problems—A total of 71 individual items were assessed in all 24 studies. Of the 71 items, three pairs of very similarly worded items were combined (e.g., I have become very rude, obnoxious, or insulting after drinking; Have you become very rude, obnoxious, or insulting after drinking?) and 68 unique items were subsequently analyzed. Items came from the Rutgers Alcohol Problem Index (RAPI; White & Labouvie, 1989), the Young Adult Alcohol Problems Screening Test (YAAPST; Hurlbut & Sher, 1992), the Brief Young Adult Alcohol Consequences Questionnaire (BYAACQ; Kahler, Strong, & Read, 2005), the Alcohol Use Disorders Identification Test (AUDIT; Saunders, Aasland, Babor, De La Fuente, & Grant, 1993), the Positive and Negative Consequences Experienced questionnaire (PNCE; D'Amico & Fromme, 1997), and the Alcohol Dependence Scale (ADS; Skinner & Allen, 1982; Skinner & Horn, 1984). For each item, responses were dichotomized to indicate 1 = *Yes*; 0 = *No*, because this response format was the common denominator across studies. When someone did not drink during the time frame referenced, their score was recoded as zero.

Several existing psychometric studies on alcohol-related problems have used a single-factor structure (RAPI - Neal, Corbin, & Fromme, 2006; YAAPST - Kahler, Strong, Read, Palfai, & Wood, 2004; BYAACQ - Kahler et al., 2005; Diagnostic and Statistical Manual of Mental Disorders, 4th edition [DSM-IV; American Psychiatric Association, 1994]; alcohol use disorder symptoms - Martin, Chung, Kirisci, & Langenbucher, 2006). Thus, we derived latent trait scores using a unidimensional 2-PL IRT model, which assumes that a single overall, severity latent trait gives rise to item responses. We also estimated a four-dimensional 2-PL IRT model (the four related, but distinct dimensions were Neglecting responsibilities, Interpersonal difficulties, Dependence-like symptoms, and Acute heavy-drinking). The estimated correlations among the four dimensions exceeded 0.8. For two small studies (Studies 13 and 14; combined $N = 138$), only sum scores of the RAPI, but not individual item scores, were available. We matched latent trait scores for these participants using their RAPI sum scores with those from studies that had both latent trait scores and RAPI sum scores.

In the factor analysis environment, items are evaluated in terms of their factor loadings and thresholds (intercepts for continuous indicators), whereas in IRT analysis items are typically evaluated by their discrimination and difficulty (or severity) item parameters. The item discrimination parameter is the slope of the item characteristic curve that indicates an item's ability to discriminate among respondents, and how strongly an item is correlated to the underlying latent trait. Items with steeper slopes indicate better discrimination. For example, Item C (*The quality of work suffered because of drinking*) in Figure 1 discriminates respondents better than Item E (Getting into trouble because of drinking at work or school). The item difficulty parameter indicates the location of the item along the latent trait continuum where the probability of endorsement of the item is 0.5, and indicates how easy or difficult the item is to endorse. Items with higher difficulty are less frequently endorsed. We examined the total information curve (see Figure S1 in the Online Supplement), which provides the overall performance of the measure at each level of an underlying latent trait (Markon, 2013). Overall, the items for alcohol-related problems provided less reliable or precise information about individuals whose underlying latent traits were at the lower end of the spectrum, but more reliable information for individuals whose traits were at the higher end of the spectrum (e.g., θ scores between 1 and 3). This also reflects that few alcohol-related problems items are easy to endorse in the present study, and that the majority of these items are more sensitive and informative for those who report high levels of alcohol-related problems, which is similar to findings from a previous analysis (Neal et al., 2006).

It is worth mentioning that in deriving latent trait scores, there was a need to reconcile different referent time frames across studies. Most of the studies used a short referent time frame (3 months or shorter) for alcohol-related problems at baseline. More specifically, 20 studies out of 24 used a 1- to 3-month time frame, and three studies (Studies 4, 10, and 12) used a 6-month time frame. Only one study (Study 3) measured past-year alcohol-related problems using the YAAPST items and also included the AUDIT items, which ask about the last year (see Table S3 in the Online Supplement for measure overlap and referent time frames at baseline across studies). A few studies asked about problems that occurred in two or three different referent time frames and we examined their responses. Study 20, in

particular, had 1-month, 6-month, and 1-year referent time frames for each RAPI item. Since there is a part-whole relationship between answers for the 1-month time frame and answers for the 6-month time frame, item endorsement rates should be higher for items assessed over the longer time frame. However, the differences across the three time frames were relatively small in magnitude, and also depended on item characteristics. For example, for a relatively easy item to endorse, such as "Got into fights, acted bad, or did mean things," endorsement rates went up progressively across time frames (i.e., 15%, 23%, and 28%, respectively). For a relatively more difficult or severe item, such as "Felt that you had a problem with alcohol," endorsement rates tended to be stable regardless of the referent time frame (i.e., 8%, 10%, and 11%, respectively). Correlations between 1-month and 6-month responses were also high (0.78 for the easy item; and 0.90 for the more difficult item). Most of the studies had a 1-6 month referent time frame at baseline, and follow-up assessments utilized a 1- to 3-month time frame in all studies. Note also that through IRT analysis, the measurement perspective was changed from the number of alcohol-related problems that occurred within a given time frame (i.e., a count variable) to the severity of alcohol-related problems (i.e., a trait score in a normal distribution).

The correlations between the original scale sum scores (e.g., the RAPI or YAAPST sum scores) and latent trait scores within studies were, on average, 0.83, suggesting that the rank orders of individuals within studies were similar across the two approaches (i.e., summed scale scores and theta scores from the IRT analysis). However, these two approaches are based on different measurement models and items, and are not directly comparable.

Protective behavioral strategies—Protective behavioral strategies refer to specific cognitive-behavioral strategies that can be employed to reduce risky drinking prior to and during drinking, and to limit harm from drinking (Martens et al., 2005). A total of 58 protective behavioral strategy items assessed by 13 studies were analyzed. Items came from the 10-item Protective Behavioral Strategies (PBS) measure taken from the National College Health Assessment survey (American College Health Association, 2001), the 15-item Protective Behavioral Strategies Scale (PBSS; Martens et al., 2005), the 37-item Self Control Strategies Questionnaire (SCSQ; Werch & Gorman, 1988), a seven-item Drinking Restraint Strategies (DRS) scale used in Wood, Capone, Laforge, Erickson, and Brand (2007), and a nine-item Drinking Strategies (DS) scale reported in Wood et al. (2010). We removed items that indicated either abstinence (i.e., *Chose not to drink alcohol*) or risky (as opposed to protective) drinking behaviors (i.e., *Drink shots of liquor*), as well as items that were used in only one study, as they could not be linked to measures of other studies for our IRT analysis. Of the remaining items, 20 were combined into five individual items because they were very similarly worded (e.g., use a designated driver; used a designated driver; use a safe ride or taxi service when you have been drinking; make arrangements not to drive when drinking). Forty-three remaining items were analyzed via hierarchical IRT, specifying a single, underlying dimension of protective behavioral strategies. Although the literature varies as to the dimensionality of these behaviors (e.g., three dimensions for the PBSS in Martens et al. [2005]; four dimensions for the PBSS in Walters, Roudsari, Vader, & Harris [2007]; seven factors for the external SCSQ in Werch & Gorman [1988]; one summed score for the DS in Wood et al., 2010), we used a unidimensional IRT model because of lack of

overlap in items across studies and also because protective behavioral strategies are often used as a single overall score (e.g., Benton, Downey, Glider, & Benton, 2008; Martens, Ferrier, & Cimini, 2007). Furthermore, the three dimension scores of the PBSS are similarly related to alcohol use, alcohol-related problems, and depressive symptom scores (Martens et al., 2005; Martens et al., 2008). Only data for individuals who reported recent drinking (i.e., past 1 to 3 months) were included.

Items were recoded to indicate 0 = *never*; 1 = *rarely, seldom, occasionally, or sometimes*; 2 = *usually or often*; and 3 = *always*. Although some of the past studies dichotomized item responses (0 and 1 vs. 2 and 3; e.g., Walters, Roudsari, et al., 2007), we deemed it important that a protective behavior that is *often* used be differentiated from one that is *always* used, and that this difference be reflected in estimating latent trait scores. Thus, we used a generalized partial credit model (GPCM; Muraki, 1992) to assign partial credit for polytomous items. Unlike the previous IRT model, the single difficulty parameter of an item is replaced by three step difficulty parameters, each of which can be interpreted as the intersection point of two adjacent item response curves (0 and 1, 1 and 2, 2 and 3; see Figure 2). These intersection points are the points on the latent trait scale axis (*x*-axis) where one response (e.g., 2 = *usually or often*) becomes relatively more likely than the preceding response (e.g., 1 = *rarely*).

Figure 2 shows category response curves for two protective behaviors under the partial credit model. It is relatively easy for participants to endorse "*rarely*" or "*usually*" as opposed to "*never*" for Item B (*Eat before and/or during drinking*), compared to Item A (*Stop drinking at a predetermined time*). Most of the responses to Item A were either "*never*" or "*rarely.*" In contrast, most of the responses to Item B occurred between "*rarely*" and "*always.*" Item step difficulty parameter estimates reflect this relative difficulty. Item step difficulty parameter estimates for Item A were higher than those for Item B at intersection points (e.g., 1.62 vs. –0.48 for Item A vs. Item B for the intersection between "*rarely*" and "*usually*"). In sum, it is relatively more difficult to stop drinking at a predetermined time than to eat food during or before drinking. Polytomous items, therefore, can meaningfully be interpreted in terms of how difficult one item is to endorse compared to other items. The correlations between the original scale sum scores (e.g., the PBS, PBSS) and latent trait scores within studies exceeded 0.96.

Differential item functioning and latent traits—Differential item functioning (DIF) tests examine whether participants with the same level of a construct but different backgrounds respond similarly to the same items, and are often conducted in IDA research (Curran et al., 2008; Hussong et al., 2007). Likewise, important covariates, which can be different for different IDA studies, can be included in measurement models when estimating latent traits (e.g., moderated nonlinear factor analysis [MNLFA]; Bauer & Hussong, 2009; Curran et al., 2014). Each IDA study may also make certain assumptions about data and item performance.

In deriving item parameters and latent trait scores in the current study, we initially made an assumption that the same items administered in different studies had the same item parameters as specified in the item response function, after taking into account different

average trait levels across studies. We reasoned that it is a sensible assumption because all participants were college students who were assessed within a narrow window of assessment (i.e., 12 months). In addition, we had a high proportion of overall missing data at the item level, which can be attributed to the large number of both studies and items that were pooled. Note also that we treated many similarly worded items as different items, which increased the number of items and, consequently, the amount of missing data. The high proportion of missing data for this large scale IDA data set made it very difficult, if not impossible, to examine DIF for many items (see Huo et al., 2014 for an example of item overlap across studies [in their Table 5] and findings from a simulation study on missing data). In addition, the amount of missingness prevented us from using existing software programs, such as Mplus (Muthén & Muthén, 1998-2014), to compute the tetrachoric correlation from which further analyses (e.g., factor analysis, structural equation modeling analysis) can be conducted.

We should note that there exists an indeterminacy between DIF and group (study) differences in latent traits, which has been well known among psychometricians for some time (e.g., Thissen, Steinberg, & Gerrard, 1986), and that DIF depends on the items or a set of items that serve as a point of reference (i.e., an anchor) because the choice of invariant items within a pool of items affects how remaining items behave (Bechger, Gunter, & Verstralen, 2010; see also Byrne, Shavelson, & Muthén, 1989 for the nonindependence of these tests in the context of confirmative factor analysis). That is the reason why DIF items within a set of items can change depending on search strategies and measurement models (Kim & Yoon, 2011; Yoon & Millsap, 2007). In other words, DIF is only relative to the reference point, which can be set in several ways in a large pool of items. Consequently, latent trait scores can also shift up and down along the theta scale depending on the invariant items or DIF items, although relative positions of individuals on this scale may remain the same across groups. Even when DIF items exist across groups within a pool of items, as long as there are invariant anchor items that provide linkage across groups, latent trait scores can reasonably be estimated. Some in the literature have stated that only one invariant (i.e., non-DIF) item is necessary to establish partial invariance across different groups (studies) for a single unidimensional construct (e.g., Steenkamp & Baumgartner, 1998; also briefly noted in Bauer & Hussong, 2009). Due to this nature of DIF, it may not be best to focus on which items show DIF.

What is central for IDA is whether trait scores are unbiased in relation to a key design variable (i.e., study). Thus, we conducted an additional IRT analyses for alcohol-related problems to examine this question. We compared the latent trait scores from the original IRT model (no-DIF model) with those from alternative models that subset a portion of items to take different item parameters (i.e., DIF items) across studies. If latent trait scores resulting from these different approaches (i.e., no-DIF model vs. DIF models) are equivalent, we can be assured that our trait scores, as a whole, are invariant to study. Our strategy is essentially equivalent to the IRT strategy adopted by Hussong et al. (2007), with the exception that in Hussong et al., DIF items were specifically specified, whereas we allowed some items to have DIF across studies. Results indicated that no meaningful differences existed in latent trait scores between these two IRT approaches (see Figures S2 and S3 in the Online Supplement). The rank orders of individuals within and across studies

were preserved across the two IRT models (rs ≥ 0.95). In addition, the rank orders of the studies in terms of their observed theta means were also largely the same. However, our original no-DIF model was the simpler, more parsimonious model, and had the lower deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) than the alternative DIF model. The DIC is appropriate for comparing models that are estimated from MCMC analysis. It can be considered as a generalized version of the Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978) for Bayesian models. Thus, through the use of novel IRT methods, we were able to combine different items from different scales across the studies included in Project INTEGRATE. The result of these IRT analyses is that all participants could be placed on the same underlying trait scale, although these traits were assessed with different scales, items, and/or response options in the original studies.

Given that we have intervention and control groups, we also checked latent trait scores obtained from our IRT analysis to make sure no systematic bias exists in separating these groups within studies. With the exception of the three studies that did not have a control group, all individual studies utilized random assignment. Thus, the treatment and control groups should be, and were, mostly equivalent at baseline when comparing either original scale scores or latent trait scores. Table S4 in the Online Supplement provides a list of important considerations and actions that we have made to estimate latent trait scores.

Harmonization

Harmonization can be described as the recoding of variables so that values from different variables assessing the same construct can be made comparable. More broadly, harmonization refers to a general approach where measures are retrospectively made comparable to synthesize large data sets, and it is increasingly utilized in biomedical epidemiological research (e.g., Fortier et al., 2010). Harmonization can be straightforward if standard measures, such as the Daily Drinking Questionnaire (DDQ; Collins, Parks, & Marlatt, 1985), are utilized. The DDQ asks respondents to indicate the number of drinks they consumed on each day of a typical week in the last month. In the present study, the majority of studies utilized the DDQ, which allowed us to create several key alcohol use frequency and quantity measures. Although the usual time frame for the DDQ is past month, a few studies utilized the past 3 months as a referent time frame. We assumed that this different time frame does not bias the self-reported number of drinks consumed *on each day of a typical drinking week* for college students.

Trade off between item overlap and information and limits of harmonization— In the absence of standard measures, however, one needs to weigh a gain in item overlap across studies against a loss of information that can result when trying to find a common denominator for items. In our study, for example, three studies assessed the number of drinking days (frequency of alcohol use) in the past month as an open ended question, and six studies collected daily drinking diaries for a 30-day window, which could then be used to compute the number of drinking days in the past month. In contrast, six other studies assessed the frequency of drinking using the AUDIT, which had the following ordinal response options: 0 = *Never*; 1 = *Monthly or less*; 2 = *2-4 times a month*; 3 = *2-3 times a*

week; and 4 = *4 or more times a week*. In this case, the AUDIT ordinal response format provides the 'lowest common denominator' among response options, but using it would lead to a loss of information for those studies that used more detailed assessments. So, deriving a comparable measure using harmonization required striking an appropriate balance between item overlap across studies and information (i.e., greater overlap but loss of information vs. more information retained for fewer studies). For many of the secondary outcome measures, we derived dichotomous outcome measures (e.g., any driving after three drinks or more in the past year) to ensure the broadest possible measurement coverage across studies.

For some constructs, it was not possible to derive a common measure that could be meaningfully compared across studies. One such example was heavy episodic drinking, a well known and widely used outcome measure in studies of college student drinking. Heavy episodic drinking, or binge drinking, is defined by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) as a pattern of drinking alcohol that brings blood alcohol concentration (BAC) to 0.08 grams percent or above, which corresponds to consuming five or more drinks (men) or four or more drinks (women) in two hours (NIAAA, 2004). Questions actually used in studies were: (1) how many times have you drank 5 drinks or more (for men; 4 or more drinks for women); (2) *how many times have you drank 5 drinks or more* (i.e., regardless of sex); (3) *how many times have you drank 6 drinks or more* (i.e., regardless of sex); (4) how many times have you drank 5 drinks or more in one sitting (or in a row); and (5) how many times have you drank 5 drinks or more within two hours? These questions were also asked for different referent time frames: in the past 2 weeks, 4 weeks, or 1 month. Thus, key differences across all items were the referent time period (2 weeks vs. 4 weeks/1 month), number of drinks (four, five, or six drinks), sex (sex-specific vs. sex-nonspecific), and duration of a drinking episode (unspecified hours, one sitting or in a row, or two hours). Two studies assessed both 2-week and 1-month heavy episodic drinking measures (five or more drinks for men and four or more drinks for women). Examining means and correlations of these items for these studies led us to conclude that heavy episodic drinking in the past 2 weeks could not be multiplied by two to create a measure for the past month (see Table 2). While these two measures were highly correlated ($r > 0.7$), their means were more difficult to map onto a common metric. The 1-month question tended to be an underestimate of the 2-week measure that was multiplied by two, except in one case where it was over-estimated (i.e., women in Study 22). Thus, any between-study differences in heavy episodic drinking would be confounded with the way heavy episodic drinking was asked.

Similar to the heavy episodic drinking measure, we concluded that readiness to change, a key mediator variable, could not be made comparable across studies. Readiness to change refers to the degree to which an individual is motivated to change problematic drinking patterns, and is measured by assessing different stages of cognitive and affective processes that lead to an initial change effort (Carey, Purnine, Maisto, & Carey, 1999). Although this construct was measured in the majority of the studies, each study included only one scale or a single item assessing this construct, and there was little overlap across the studies. Eight studies used the Readiness to Change Questionnaire (RCQ; Heather, Rollnick, & Bell, 1993); four studies used the University of Rhode Island Change Assessment (URICA;

Heesch, Velasquez, & von Sternberg, 2005); and seven studies used different variations of a single-item, readiness to change ruler (LaBrie, Quinlan, Schiffman, & Earleywine, 2005) or contemplation ladder (Herzog, Abrams, Emmons, & Linnan, 2000) that differed in response ranges (1 to 5, 1 to 10, or 0 to 10), as well as anchor points to mark different stages of change. With just one measure for each study and with no overlap in items across scales (and studies), we concluded that any differences in measures would be confounded with between-study differences (e.g., sample/design characteristics). Thus, any analyses using readiness to change will have to be replicated across the different scales that are available rather than using the pooled data set. The steps taken and outcomes from these steps thus far demonstrate that even with latest advances in analytical modeling and well-established measures for key constructs, there are some limits. In the section below, we provide a discussion of how to better design individual studies, especially intervention studies, with IDA in mind.

Lessons Learned Thus Far and Recommendations

One of the most striking lessons that we have learned thus far is that this innovative approach to synthesizing information from multiple studies is very labor-intensive and time-consuming. To meaningfully conduct IDA studies for clinical outcomes, such as intervention efficacy and moderated efficacy, the number of studies included should be sufficiently large to examine study-level, as well as individual-level, differences. However, IDA demands significant time and resources to pool data from studies, clean and check data, and establish commensurate measurement scales. Citing the work by Steinberg et al. (1997) and personal communication with one of the investigators, Cooper and Patall (2009) noted that IPD meta-analysis probably costs 5-8 times more than AD meta-analysis, and takes several years from start to publication in the field of medical research. When standard measures are less commonly used or difficult to establish across studies, which is typical for psychological research, the cost may be even greater than what has been estimated for medical research, in which the primary outcome (e.g., death) can often be clearly defined.

For Project INTEGRATE, we developed new MCMC algorithms to estimate item parameters and latent trait scores across studies, which took an enormous amount of time and effort, because commercially available software programs did not sufficiently meet our needs. Our first-hand experience suggests that the application of IPD meta-analysis may require further methodological developments, whereas AD meta-analysis procedures are fairly well-established at the present time. Furthermore, unlike in medical trials where treatment and control conditions can clearly be defined (i.e., a specific procedure or drug), we learned that treatment and control groups may not be equivalent across studies, which required a closer examination of these groups to ensure that similarly labeled groups in original studies had many critical features in common (Ray et al., in press).

The capabilities to reexamine effect sizes using more appropriate analytical methods and to peruse intervention procedures to appropriately compare treatment groups are important advantages of IDA over single studies or AD meta-analysis. In the context of IPD meta-analysis, multiple RCTs are typically conceptualized as a sample of studies. The findings can then be generalized to a broader population. A recent IPD meta-analysis study that

examined the efficacy of BMIs for Project INTEGRATE is one such research application (Huh et al., 2014). In Huh et al., we utilized Bayesian multilevel, over-dispersed Poisson hurdle models to examine intervention effects on drinks per week and peak drinking, and Gaussian models for alcohol problems. This analytic approach accommodated the sampling, sample characteristics, and distributions of the pooled data while overcoming some of the challenges associated with being an IDA study, one of which was the unbalanced RCT design (i.e., 21 interventions vs. 17 controls across 17 studies) of the pooled data set. Although the study by Huh et al. highlights some of the promises of IDA, for this type of investigation, a large enough number of studies are needed to obtain sufficient precision about point estimates and standard errors. Others have said that at least 10 - 20 studies may be needed for population representation and proper model estimation (e.g., Hussong et al., 2013). As the number of studies included for IDA goes up, however, so does the demand on time and other resources.

Having emphasized the need for greater resources for IDA, we remain enthusiastic that IDA is a better research strategy for examining low base rate behaviors, such as marijuana or drug use outcomes (White et al., 2014), and for finding subgroups who may respond to treatment differently (i.e., moderators of treatment outcomes), which is widely considered as one of the most important strengths of IDA (e.g., Brown et al., 2013). Thus, IDA holds special promises for the field. We would also like to note that the resources needed may be highly specific to the research goals of individual IDA studies. Other notable strengths of IDA, compared to single studies, include larger, more heterogeneous samples and more repeated measures for longer observed periods. Depending on the specific research questions, the pooled data set from just two studies may be better than data from a single study, as long as the replicability of measurement models can hold across studies.

Emerging analytical and technological advances may provide more favorable environments for pooling and analyzing IPD in the future. In the present moment, our experiences suggest ways to lower barriers to IDA by planning single intervention studies differently. Below, we make several recommendations for future single, intervention studies.

Increase Overlap in Measures

The simplest option to increase overlap in measures across studies is to use standardized and common measures for a given construct in future single studies. If there is a need to include a newly developed questionnaire or instrument, it would be quite helpful to include other established measures of the same construct to link items from different measures. Note that the overlap needs to exist, not just at the level of the constructs, but at the level of items (and response options). When a concern arises about burdening participants with multiple items, it may be better to administer a portion of items from one measure and a portion of items from other measures (e.g., two versions A-B and B-C administered to two groups), as is done in a planned missingness design (Graham, Hofer, & MacKinnon, 1996). This strategy, a common practice in educational research, is better for IDA because items can be linked across studies. In theory, a single item may be used to provide such a chain. However, the level of precision or trustworthiness of the chain will improve with more shared items across studies. Our experience also suggests that, with more work, item banks may be developed

for key constructs for this college population, which may make it feasible to derive latent trait scores across studies in the future without the needed overlap in items. At present, there is no such known item bank specifically aimed at this population.

Based on our experience, the importance of common, standard items may be greater for single-item measures, such as heavy episodic drinking, which are often utilized in alcohol research. Our experience is by no means unique. Other investigators have also noted the difficulty of harmonizing alcohol measures across studies (e.g., analysis of twin studies; Agrawal et al., 2012; genome-wide association studies; Hamilton et al., 2011). Future investigations could utilize measures from well researched and accessible research tools, such as the Phenotypes and eXposures (PhenX) Toolkit (Hamilton et al. [2011], http://www.Phenxtoolkit.org/), the NIH Toolbox for assessment of neurological and behavioral function (http://www.nihtoolbox.org/Pages/default.aspx), or the Patient-Reported Outcomes Measurement Information System (PROMIS; http://www.nihpromis.org/; see Pilkonis et al., 2013 for the development of item banks for alcohol use, consequences, and expectancies).

Increase Overlap in Follow-ups

The ability to extend the range of observations in terms of the observed time period is one of the advantages of IDA over single studies. However, this can lead to a greater portion of missing data in the combined data set. Two types of missing data exist in IDA: items that were not assessed by study design and, are thus missing at the level of studies; and items that were included but not answered by the participant (Gelman, King, & Liu, 1998). Table 1 provides a glimpse of the sparse nature of pooled data across time, especially at longer-term follow-ups (e.g., 6-12 months post intervention). Table S2 shows available constructs for each study. In both tables, missing data are due to different study designs across studies. Within studies, there were also missing data at the individual level due to omitted responses. Although missing values may be random in nature (i.e., missing at random [MAR]) and ignorable (Schafer, 1997) for this project, the pattern of missingness was unique for some studies, and the overall proportion of missing data was substantial.

The missing data challenge can be mitigated if there is better overlap in follow-up assessments across intervention studies. Overall, the power to detect a group difference goes up with increases in the duration of observations, the number of repeated assessments, and the sample size. Of those, the duration has the greatest effect, and the number of repeated assessments has the smallest effect on power (Moerbeek, 2008). Despite the small effect on power, to capture a change immediately following an intervention and a slower subsequent rebound, one has to have at least four (preferably more) repeated assessments to estimate polynomial growth models without the need to impose restrictive constraints. While it is reasonable to assess outcomes more frequently, for example, in the first 3 months following a BMI, it is also desirable to assess outcome data beyond the initial phase to see whether, and for how long, the intervention effect is sustained. We recommend that future alcohol intervention trials extend the period under observation to intermediate or long-term follow-ups (e.g., 6-12 months post intervention) as this longer-term follow-up is needed from both substantive and methodological perspectives. Assuming that missing data at follow-ups (i.e., dropouts) meet the MAR assumption, the extension of the observed duration should improve

power to detect intervention efficacy and other mediational effects. Similar to the case of increasing item overlap by design, the use of planned missingness may prove to be useful in estimating patterns of change for intervention studies. A design of 1-, 3-, 6-, and 9-month follow-ups for a random half sample and 1-, 2-, 6-, and 12-month follow-ups for the other half, for example, would provide up to seven time points, including baseline, for up to a year, with overlap at baseline, 1-month follow-up, and 6-month follow-up.

Reduce Heterogeneity in Treatment and Control Groups across Trials

Project INTEGRATE includes interventions that varied in, for example, the number and type of content topics covered and the manner in which they were delivered (e.g., in-person one-on-one, in-person group, by mail, etc.) to participants across studies. Therefore, we developed detailed coding procedures for all intervention and control conditions, which allowed us to determine whether similarly labeled groups are indeed equivalent (see Ray et al., in press for detail). Based on the content analysis of these components across conditions and the subsequent analysis of those components, we relabeled some of the groups and removed others from the main data set (see the Online Supplement, Table S1). This observation highlights a need to develop detailed documentation on the proposed mechanisms and protocols for any new treatment and for any new variant of an existing, evidence-based treatment in the future. In designing future single studies, one should also carefully consider a treatment group and a comparison group for their comparability and overlap with other studies.

Improve Transparency and Documentation

In general, it would be helpful to have greater transparency and better documentation in published articles, as well as in unpublished supporting materials. General reporting guidelines, such as the CONSORT statement (Schulz et al., 2010) and the Journal Article Reporting Standards (JARS) by the American Psychological Association Publications and Communications Board Working Group (2008), have provided a minimum reporting standard for various types of studies, including RCTs. ClinicalTrials.gov, an online registry and results database for Phases 2 through 4 intervention studies, provides easy access to some of the critical, scientific information about clinical studies (i.e., participant flow, baseline characteristics, outcome measures, statistical analyses, and adverse events; Tse et al., 2009). However, the required minimum information for ClinicalTrials.gov focuses on the overall efficacy and adverse events of a treatment, and does not go far enough to facilitate future IDA investigations.

We recommend that any additional outcome measures and covariates at each assessment point, follow-up schedules (beyond post-treatment), and any additional groups (treatment arms) be publicly accessible if they are omitted in published articles. This supplementary information, which could be publicly accessible and searchable, would facilitate IDA studies in the future by helping to select studies for IDA or determining feasibility of such investigations. More detailed and accurate documentation will decrease the need, for example, to pore over codebooks, questionnaires, and data to examine the nature of variation in key outcome measures and covariates. Making this information publicly available may

also help to increase awareness among investigators as to the potential overlap with other studies when planning a single study.

Conclusions

Project INTEGRATE was launched to generate robust statistical inference on the efficacy of BMIs for college students, and to examine theory-supported mechanisms of behavior change. The detailed account outlined in this paper illustrates both the promises and challenges of this particular IDA project and of IDA in general. The promises of IDA are attractive in the current research environment where limited resources are maximized by taking advantage of more efficient designs and analyses. Moreover, IDA investigations are well positioned to confront current outcries about replication failures and potentially overstated treatment benefits in the era of evidence-based treatment decision making. At the same time, these notable promises are coupled with significant challenges. IDA is not a single analytic technique per se. Rather, it is a set of advanced methods that can be tailored and implemented to address specific goals and challenges of each IDA study, which can be seen clearly in the present article. Our strategies and procedures differed from those of others (e.g., Hussong et al., 2013), which can be attributed to the different data characteristics and different assumptions made about item performance in our study. More methodological research is needed to test these assumptions and to develop guidelines for IDA research, which is expected to increase in the future. Nonetheless, the specific recommendations that we have for single intervention studies may be helpful not only for more robust research practice but also for large-scale research synthesis, such as IPD meta-analysis and IDA.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The project described was supported by Award Number R01 AA019511 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAAA or the National Institutes of Health. We would like to thank Lisa A. Garberson, Caressa Slocum, and Yue Feng for their helpful comments on earlier drafts of this paper and for their help with data management.

References

Agrawal A, Freedman ND, Cheng Y-C, Lin P, Shaffer JR, Sun Q, Bierut LJ. Measuring alcohol consumption for genomic meta-analyses of alcohol intake: Opportunities and challenges. The American Journal of Clinical Nutrition. 2012; 95:539–547. doi: 10.3945/ajcn.111.015545. [PubMed: 22301922]

Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19:716–723. doi:10.1109/TAC.1974.1100705.

American College Health Association. National College Health Assessment ACHA-NCHA reliability and validity analyses. American College Health Association; Baltimore, MD: 2001.

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed.. Author; Washington, DC: 1994.

American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards. Reporting standards for research in psychology: Why do we need them? What might they be? (2008). American Psychologist. 2008; 63:839–851. doi: 10.1037/0003-066x.63.9.839. [PubMed: 19086746]

Baer JS, Kivlahan DR, Blume AW, McKnight P, Marlatt GA. Brief intervention for heavy-drinking college students: 4-year follow-up and natural history. American Journal of Public Health. 2001; 91:1310–1318. [PubMed: 11499124]

Bakker M, van Dijk A, Wicherts JM. The rules of the game called psychological science. Perspectives on Psychological Science. 2012; 7:543–554. doi: 10.1177/1745691612459060.

Barnett NP, Murphy JG, Colby SM, Monti PM. Efficacy of counselor vs. computer-delivered intervention with mandated college students. Addictive Behaviors. 2007; 32:2529–2548. doi: 10.1016/j.addbeh.2007.06.017. [PubMed: 17707594]

Bauer DJ, Hussong AM. Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. Psychological Methods. 2009; 14:101–125. doi: 10.1037/a0015583. [PubMed: 19485624]

Bechger, TM.; Gunter, M.; Verstralen, HHFM. Measurement and Reserch Department Reports. Cito; Arnhem, The Netherlands: 2010. A different view on DIF..

Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012; 483(7391):531–533. doi:10.1038/483531a. [PubMed: 22460880]

Benton SL, Downey RG, Glider PJ, Benton SA. College students' norm perception predicts reported use of protective behavioral strategies for alcohol consumption. Journal of Studies on Alcohol and Drugs. 2008; 69:859–865. [PubMed: 18925344]

Borenstein, M.; Hedges, LV.; Higgins, JPT.; Rothstein, H.; R.. Introduction to meta- analysis. Wiley; New York, NY: 2009.

Brown CH, Sloboda Z, Faggiano F, Teasdale B, Keller F, Burkhart G, Prevention Science and Methodology Group. Methods for synthesizing findings on moderation effects across multiple randomized trials. Prevention Science. 2013; 14:144–156. doi: 10.1007/s11121-011-0207-8. [PubMed: 21360061]

Byrne BM, Shavelson RJ, Muthén BO. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin. 1989; 105:456–466. doi: 10.1037/0033-2909.105.3.456.

Carey KB, Purnine DM, Maisto SA, Carey MP. Assessing readiness to change substance abuse: A critical review of instruments. Clinical Psychology: Science and Practice. 1999; 6:245–266. doi: 10.1093/clipsy.6.3.245.

Carey KB, Scott-Sheldon LA, Carey MP, DeMartini KS. Individual-level interventions to reduce college student drinking: A meta-analytic review. Addictive Behaviors. 2007; 32:2469–2494. doi: S0306-4603(07)00145-1 [pii]10.1016/j.addbeh.2007.05.004. [PubMed: 17590277]

Carey KB, Scott-Sheldon LAJ, Elliott JC, Garey L, Carey MP. Face-to-face versus computer-delivered alcohol interventions for college drinkers: A meta-analytic review, 1998 to 2010. Clinical Psychology Review. 2012; 32:690–703. doi: 10.1016/j.cpr.2012.08.001. [PubMed: 23022767]

Cimini MD, Martens MP, Larimer ME, Kilmer JR, Neighbors C, Monserrat JM. Assessing the effectiveness of peer-facilitated interventions addressing high-risk drinking among judicially mandated college students. Journal of Studies on Alcohol and Drugs. 2009; (Supplement (16)):57–66. [PubMed: 19538913]

Clarke N, Kim S-Y, White HR, Jiao Y, Mun E-Y. Associations between alcohol use and alcohol-related negative consequences among Black and White college men and women. Journal Studies on Alcohol and Drugs. 2013; 74:521–531.

Collins RL, Parks GA, Marlatt GA. Social determinants of alcohol consumption: The effects of social interaction and model status on the self-administration of alcohol. Journal of Consulting and Clinical Psychology. 1985; 53:189–200. doi: 10.1037/0022-006x.53.2.189. [PubMed: 3998247]

Cooper H, Patall EA. The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. Psychological Methods. 2009; 14:165–176. doi: 10.1037/a0015565. [PubMed: 19485627]

Cronce JM, Larimer ME. Individual-focused approaches to the prevention of college student drinking. Alcohol Research & Health. 2011; 34:210–221. [PubMed: 22330220]

Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. Psychological Methods. 2009; 14:81–100. doi: 10.1037/a0015914. [PubMed: 19485623]

Curran PJ, Hussong AM, Cai L, Huang W, Chassin L, Sher KJ, Zucker RA. Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. Developmental Psychology. 2008; 44:365–380. doi: 10.1037/0012-1649.44.2.365. [PubMed: 18331129]

Curran PJ, McGinley JS, Bauer DJ, Hussong AM, Burns A, Chassin L, Zucker R. A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. Multivariate Behavioral Research. 2014; 49:214–231. doi: 10.1080/00273171.2014.889594.

D'Amico EJ, Fromme K. Health risk behaviors of adolescent and young adult siblings. Health Psychology. 1997; 16:426–432. doi: 10.1037/0278-6133.16.5.426. [PubMed: 9302539]

DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986; 7:177–188. doi: 10.1016/0197-2456(86)90046-2. [PubMed: 3802833]

Dimeff, LA.; Baer, JS.; Kivlahan, DR.; Marlatt, GA. Brief alcohol screening and intervention for college students. Guilford; New York, NY: 1999.

Doornik, JA. Object-oriented matrix programming using Ox (Version 3.1) [Computer software]. Timberlake Consultants; London, UK: 2009.

Fortier I, Burton PR, Robson PJ, Ferretti V, Little J, L'Heureux F, Hudson TJ. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. International Journal of Epidemiology. 2010; 39:1383–1393. doi: 10.1093/ije/dyq139. [PubMed: 20813861]

Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC, Fawcett J. Antidepressant drug effects and depression severity: A patient-level meta-analysis. Journal of the American Medical Association. 2010; 303:47–53. doi: 10.1001/jama.2009.1943. [PubMed: 20051569]

Fromme K, Corbin W. Prevention of heavy drinking and associated negative consequences among mandated and voluntary college students. Journal of Consulting and Clinical Psychology. 2004; 72(6):1038–1049. doi: 10.1037/0022-006X.72.6.1038. [PubMed: 15612850]

Gelman A, King G, Liu C. Not asked and not answered: Multiple imputation for multiple surveys: Rejoinder. Journal of the American Statistical Association. 1998; 93(443):869–874. doi: 10.2307/2669825.

Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kim JB, Kupfer DJ. Development of a computerized adaptive test for depression. Archives of General Psychiatry. 2012; 69:1104–1112. doi: 10.1001/archgenpsychiatry.2012.14. [PubMed: 23117634]

Graham JW, Hofer SM, MacKinnon DP. Maximizing the usefulness of data obtained with planned missing value patterns: an application of maximum likelihood procedures. Multivariate Behavioral Research. 1996; 31:197–218. doi: 10.1207/s15327906mbr3102_3.

Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, Haines J. The PhenX Toolkit: Get the most from your measures. American Journal of Epidemiology. 2011; 174:253–260. doi: 10.1093/aje/kwr193. [PubMed: 21749974]

Hardin, JW. The sandwich estimate of variance.. In: Fomby, TB.; Hill, RC., editors. Maximum likelihood estimation of misspecified models: Twenty years later. Emerald Group Publishing Limited; Bingley, UK: 2003. p. 45-73.doi: 10.1016/S0731-9053(03)17003-X

Heather N, Rollnick S, Bell A. Predictive validity of the readiness to change questionnaire. Addiction. 1993; 88:1667–1677. doi: 10.1111/j.1360-0443.1993.tb02042.x. [PubMed: 8130706]

Heesch KC, Velasquez MM, von Sternberg K. Readiness for mental health treatment and for changing alcohol use in patients with comorbid psychiatric and alcohol disorders: Are they congruent? Addictive Behaviors. 2005; 30:531–543. doi: 10.1016/j.addbeh.2004.08.003. [PubMed: 15718069]

Herzog TA, Abrams DB, Emmons KM, Linnan L. Predicting increases in readiness to quit smoking: A prospective analysis using the contemplation ladder. Psychology & Health. 2000; 15:369–381. doi: 10.1080/08870440008401999.

Huh D, Mun E-Y, Larimer ME, White HR, Ray AE, Rhew I, Atkins DC. Brief motivational interventions for college student drinking may not be as powerful as we think: An individual participant-level data meta-analysis. 2014 (under review).

Huo Y, de la Torre J, Mun E-Y, Kim S-Y, Ray AE, Jiao Y, White HR. A hierarchical multi-unidimensional IRT approach for analyzing sparse, multi-group data for integrative data analysis. Psychometrika. Ad. 2014 online pub. doi:10.1007/s11336-014-9420-2.

Hurlbut SC, Sher KJ. Assessing alcohol problems in college students. Journal of American College Health. 1992; 41:49–58. doi: 10.1080/07448481.1992.10392818. [PubMed: 1460173]

Hussong AM, Curran PJ, Bauer DJ. Integrative data analysis in clinical psychology research. The Annual Review of Clinical Psychology. 2013; 9:61–89. doi:10.1146/annurev-clinpsy-050212-185522.

Hussong AM, Wirth RJ, Edwards MC, Curran PJ, Chassin LA, Zucker RA. Externalizing symptoms among children of alcoholic parents: Entry points for an antisocial pathway to alcoholism. Journal of Abnormal Psychology. 2007; 116:529–542. doi: 10.1037/0021-843X.116.3.529. [PubMed: 17696709]

Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005; 2(8):e124. doi: 10.1371/journal.pmed.0020124. [PubMed: 16060722]

Kahler CW, Strong DR, Read JP. Toward efficient and comprehensive measurement of the alcohol problems continuum in college students: the Brief Young Adult Alcohol Consequences Questionnaire. Alcoholism: Clinical and Experimental Research. 2005; 29:1180–1189. doi: 10.1097/01.alc.0000171940.95813.a5.

Kahler CW, Strong DR, Read JP, Palfai TP, Wood MD. Mapping the continuum of alcohol problems in college students: a Rasch model analysis. Psychology of Addictive Behaviors. 2004; 18:322–333. doi: 10.1037/0893-164x.18.4.322. [PubMed: 15631604]

Kim ES, Yoon M. Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. Structural Equation Modeling: A Multidisciplinary Journal. 2011; 18:212–228. doi: 10.1080/10705511.2011.557337.

Kraemer H, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. Archives of General Psychiatry. 2006; 63:484–489. doi: 10.1001/archpsyc.63.5.484. [PubMed: 16651505]

LaBrie JW, Huchting K, Tawalbeh S, Pedersen ER, Thompson AD, Shelesky K, Neighbors C. A randomized motivational enhancement prevention group reduces drinking and alcohol consequences in first-year college women. Psychology of Addictive Behaviors. 2008; 22:149–155. doi: 2008-01797-017 [pii]10.1037/0893-164X.22.1.149. [PubMed: 18298242]

LaBrie JW, Huchting KK, Lac A, Tawalbeh S, Thompson AD, Larimer ME. Preventing risky drinking in first-year college women: further validation of a female-specific motivational-enhancement group intervention. Journal of Studies on Alcohol and Drugs. Suppl. 2009; 16:77–85. [PubMed: 19538915]

LaBrie JW, Hummer JF, Neighbors C, Pedersen ER. Live interactive group-specific normative feedback reduces misperceptions and drinking in college students: a randomized cluster trial. Psychology of Addictive Behaviors. 2008; 22:141–148. doi: 2008-01797-016 [pii] 10.1037/0893-164X.22.1.141. [PubMed: 18298241]

LaBrie JW, Lamb TF, Pedersen ER, Quinlan T. A group motivational interviewing intervention reduces drinking and alcohol-related consequences in adjudicated college students. Journal of College Student Development. 2006; 47:267–280. doi: 10.1016/j.addbeh.2007.05.014.

LaBrie JW, Pedersen ER, Lamb TF, Quinlan T. A campus-based motivational enhancement group intervention reduces problematic drinking in freshmen male college students. Addictive Behaviors. 2007; 32:889–901. doi: 10.1016/j.addbeh.2006.06.030. [PubMed: 16876963]

LaBrie JW, Quinlan T, Schiffman JE, Earleywine ME. Performance of alcohol and safer sex change rulers compared with readiness to change questionnaires. Psychology of Addictive Behaviors. 2005; 19:112–115. doi: 10.1037/0893-164X.19.1.112. [PubMed: 15783287]

LaBrie JW, Thompson A, Huchting K, Lac A, Buckley K. A group motivational interviewing intervention reduces drinking and alcohol-related negative consequences in adjudicated college

women. Addictive Behaviors. 2007; 32:2549–2562. doi: 10.1016/j.addbeh.2007.05.014. [PubMed: 17628347]

Larimer ME, Cronce JM. Identification, prevention and treatment: A review of individual-focused strategies to reduce problematic alcohol consumption by college students. Journal of Studies on Alcohol, Supplement. 2002; 14:148–163. [PubMed: 12022721]

Larimer ME, Cronce JM. Identification, prevention, and treatment revisited: individual-focused college drinking prevention strategies 1999-2006. Addictive Behaviors. 2007; 32:2439–2468. doi: 10.1016/j.addbeh.2007.05.006. [PubMed: 17604915]

Larimer ME, Lee CM, Kilmer JR, Fabiano PM, Stark CB, Geisner IM, Neighbors C. Personalized mailed feedback for college drinking prevention: A randomized clinical trial. Journal of Consulting and Clinical Psychology. 2007; 75:285–293. doi: 10.1037/0022-006X.75.2.285. [PubMed: 17469886]

Larimer ME, Turner AP, Anderson BK, Fader JS, Kilmer JR, Palmer RS, Cronce JM. Evaluating a brief alcohol intervention with fraternities. Journal of Studies on Alcohol. 2001; 62:370–380. [PubMed: 11414347]

Lazarsfeld, PF.; Henry, NW. Latent structure analysis. Houghton Mifflin; Boston, MA: 1968.

Lee, CM.; Kaysen, DL.; Neighbor, C.; Kilmer, JR.; Larimer, ME. Feasibility, acceptability, and efficacy of brief interventions for college drinking: Comparison of group, individual, and web-based alcohol prevention formats. Department of Psychiatry and Behavioral Sciences, University of Washington; Seattle, Washington: 2009. Unpublished manuscript

Lord, FM.; Novick, MR. Statistical theories of mental test scores. Addison-Wesley; Reading, MA: 1968.

Markon KE. Information utility: Quantifying the total psychometric information provided by a measure. Psychological Methods. 2013; 18:15–35. doi: 10.1037/a0030638. [PubMed: 23477605]

Marlatt GA, Baer JS, Kivlahan DR, Dimeff LA, Larimer ME, Quigley LA, Williams E. Screening and brief intervention for high-risk college student drinkers: results from a 2-year follow-up assessment. Journal of Consulting and Clinical Psychology. 1998; 66:604–615. [PubMed: 9735576]

Martens MP, Ferrier AG, Cimini MD. Do protective behavioral strategies mediate the relationship between drinking motives and alcohol use in college students? Journal of Studies on Alcohol and Drugs. 2007; 68:106–114. [PubMed: 17149524]

Martens MP, Ferrier AG, Sheehy MJ, Corbett K, Anderson DA, Simmons A. Development of the protective behavioral strategies survey. Journal of Studies on Alcohol. 2005; 66:698–705. [PubMed: 16329461]

Martens MP, Kilmer JR, Beck NC, Zamboanga BL. The efficacy of a targeted personalized drinking feedback intervention among intercollegiate athletes: A randomized controlled trial. Psychology of Addictive Behaviors. 2010; 24:660–669. doi: 10.1037/a0020299. [PubMed: 20822189]

Martens MP, Martin JL, Hatchett E, Fowler RM, Fleming KM, Karakashian MA, Cimini M. Protective behavioral strategies and the relationship between depressive symptoms and alcohol-related negative consequences among college students. Journal of Counseling Psychology. 2008; 55:535–541. doi: 10.1037/a0013588. [PubMed: 22017560]

Martin CS, Chung T, Kirisci L, Langenbucher JW. Item response theory analysis of diagnostic criteria for alcohol and cannabis use disorders in adolescents: Implications for DSM V. Journal of Abnormal Psychology. 2006; 115:807–814. doi: 10.1037/0021-843x.115.4.807. [PubMed: 17100538]

Moerbeek M. Powerful and cost-efficient designs for longitudinal intervention studies with two treatment groups. Journal of Educational and Behavioral Statistics. 2008; 33:41–61. doi: 10.3102/1076998607302630.

Moreno SG, Sutton AJ, Turner EH, Abrams KR, Cooper NJ, Palmer TM, Ades AE. Novel methods to deal with publication biases: Secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. BMJ. 2009; 339:b2981. doi: 10.1136/bmj.b2981. [PubMed: 19666685]

Muraki E. A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement. 1992; 16:159–176. doi: 10.1177/014662169201600206.

Murphy JG, Benson TA, Vuchinich RE. A comparison of personalized feedback for college student drinkers delivered with and without a motivational interview. Journal of Studies on Alcohol. 2004; 65:200–203. [PubMed: 15151350]

Murphy JG, Duchnick JJ, Vuchinich RE, Davison JW, Karg RS, Olson AM, Coffey TT. Relative efficacy of a brief motivational intervention for college student drinkers. Psychology of Addictive Behaviors. 2001; 15:373–379. [PubMed: 11767271]

Muthén, LK.; Muthén, BO. Mplus user's guide (version 7.2). Muthén and Muthén; Los Angeles: 1998–2014.

National Institute on Alcohol Abuse and Alcoholism. Binge drinking defined. 2004. Retrieved from http://pubs.niaaa.nih.gov/publications/Newsletter/winter2004/Newsletter_Number3.htm

Neal DJ, Corbin WR, Fromme K. Measurement of alcohol-related consequences among high school and college students: Application of item response models to the Rutgers Alcohol Problem Index. Psychological Assessment. 2006; 18:402–414. doi: 10.1037/1040-3590.18.4.402. [PubMed: 17154761]

Nosek BA, Spies JR, Motyl M. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science. 2012; 7:615–631. doi: 10.1177/1745691612459058.

Pilkonis PA, Yu L, Colditz J, Dodds N, Johnston KL, Maihoefer C, McCarty D. Item banks for alcohol use from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Use, consequences, and expectancies. Drug and Alcohol Dependence. 2013; 130(1–3):167–177. doi: 10.1016/j.drugalcdep.2012.11.002. [PubMed: 23206377]

Ray AE, Kim S-Y, White HR, Larimer ME, Mun EY, Clarke N. The Project INTEGRATE Team. (in press). When less is more and more is less in brief motivational interventions: Characteristics of intervention content and their associations with drinking outcomes. Psychology of Addictive Behaviors. doi: 10.1037/a0036593.

Saunders JB, Aasland OG, Babor TF, De La Fuente JR, Grant M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on early detection of persons with harmful alcohol consumption-II. Addiction. 1993; 88:791–804. doi: 10.1111/j.1360-0443.1993.tb02093.x. [PubMed: 8329970]

Schafer, JL. Analysis of incomplete multivariate data. Chapman & Hall/CRC; New York, NY: 1997.

Schwarz GE. Estimating the dimension of a model. Annals of Statistics. 1978; 6:461–464.

Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. PLoS Med. 2010; 7(3):e1000251. doi: 10.1371/journal.pmed.1000251. [PubMed: 20352064]

Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011; 22:1359–1366. doi: 10.1177/0956797611417632. [PubMed: 22006061]

Skinner HA, Allen B. Alcohol dependence syndrome: Measurement and validation. Journal of Abnormal Psychology. 1982; 91:199–209. doi: 10.1037/0021-843X.91.3.199. [PubMed: 7096790]

Skinner, HA.; Horn, JL. Alcohol Dependence Scale: Users guide. Addiction Research Foundation; Toronto, Canada: 1984.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B. 2002; 64:583–639. doi: 10.1111/1467-9868.00353.

Steenkamp J-BEM, Baumgartner H. Assessing measurement invariance in cross-national consumer research. Journal of Consumer Research. 1998; 25:78–107. doi: 10.1086/209528.

Steinberg KK, Smith SJ, Stroup DF, Olkin I, Lee NC, Williamson GD, Thacker SB. Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. American Journal of Epidemiology. 1997; 145:917–925. doi: 10.1093/oxfordjournals.aje.a009051. [PubMed: 9149663]

Thissen DS, Steinberg L, Gerrard M. Beyond group-mean differences: The concept of item bias. Psychological Bulletin. 1986; 99:118–128. doi: 10.1037/0033-2909.99.1.118.

Tse T, Williams RJ, Zarin DA. Reporting “basic results” in clinicaltrials.gov. CHEST. 2009; 136:295–303. doi: 10.1378/chest.08-3022. [PubMed: 19584212]

Tversky A, Kahneman D. Belief in the law of small numbers. Psychological Bulletin. 1971; 76:105–110. doi: 10.1037/h0031322.

Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. New England Journal of Medicine. 2008; 358:252–260. doi:10.1056/NEJMsa065779. [PubMed: 18199864]

Walters ST, Roudsari BS, Vader AM, Harris TR. Correlates of protective behavior utilization among heavy-drinking college students. Addictive Behaviors. 2007; 32:2633–2644. doi: 10.1016/j.addbeh.2007.06.022. [PubMed: 17669596]

Walters ST, Vader AM, Harris TR. A controlled trial of web-based feedback for heavy drinking college students. Prevention Science. 2007; 8:83–88. doi: 10.1007/s11121-006-0059-9. [PubMed: 17136461]

Walters ST, Vader AM, Harris TR, Field CA, Jouriles EN. Dismantling motivational interviewing and feedback for college drinkers: A randomized clinical trial. Journal of Consulting and Clinical Psychology. 2009; 77:64–73. doi: 2009-00563-015 [pii]10.1037/a0014472. [PubMed: 19170454]

Walters ST, Vader AM, Harris TR, Jouriles EN. Reactivity to alcohol assessment measures: an experimental test. Addiction. 2009; 104:1305–1310. doi: 10.1111/j.1360-0443.2009.02632.x. [PubMed: 19624323]

Werch CE, Gorman DR. Relationship between self-control and alcohol consumption patterns and problems of college students. Journal of Studies on Alcohol. 1988; 49:30–37. [PubMed: 3347074]

White HR, Labouvie EW. Towards the assessment of adolescent problem drinking. Journal of Studies on Alcohol. 1989; 50:30–37. [PubMed: 2927120]

White HR, Mun E-Y, Morgan TJ. Do brief personalized feedback interventions work for mandated students or is it just getting caught that works? Psychology of Addictive Behaviors. 2008; 22:107–116. doi:10.1037/0893-164x.22.1.107. [PubMed: 18298236]

White HR, Mun E-Y, Pugh L, Morgan TJ. Long-term effects of brief substance use interventions for mandated college students: sleeper effects of an in-person personal feedback intervention. Alcoholism: Clinical and Experimental Research. 2007; 31:1380–1391. doi: ACER435 [pii]10.1111/j.1530-0277.2007.00435.x.

White HR, Jiao Y, Ray AE, Huh D, Atkins DC, Larimer ME, Mun E-Y. Are there secondary effects on marijuana use from brief alcohol interventions for college students?. 2014 (under review).

Wood MD, Capone C, Laforge R, Erickson DJ, Brand NH. Brief motivational intervention and alcohol expectancy challenge with heavy drinking college students: a randomized factorial study. Addictive Behaviors. 2007; 32:2509–2528. doi: 10.1016/j.addbeh.2007.06.018. [PubMed: 17658696]

Wood MD, Fairlie AM, Fernandez AC, Borsari B, Capone C, Laforge R, Carmona-Barros R. Brief motivational and parent interventions for college students: a randomized factorial study. Journal of Consulting and Clinical Psychology. 2010; 78:349–361. doi: 10.1037/a0019166. [PubMed: 20515210]

Yoon M, Millsap RE. Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. Structural Equation Modeling: A Multidisciplinary Journal. 2007; 14:435–463. doi: 10.1080/1070551070130167.

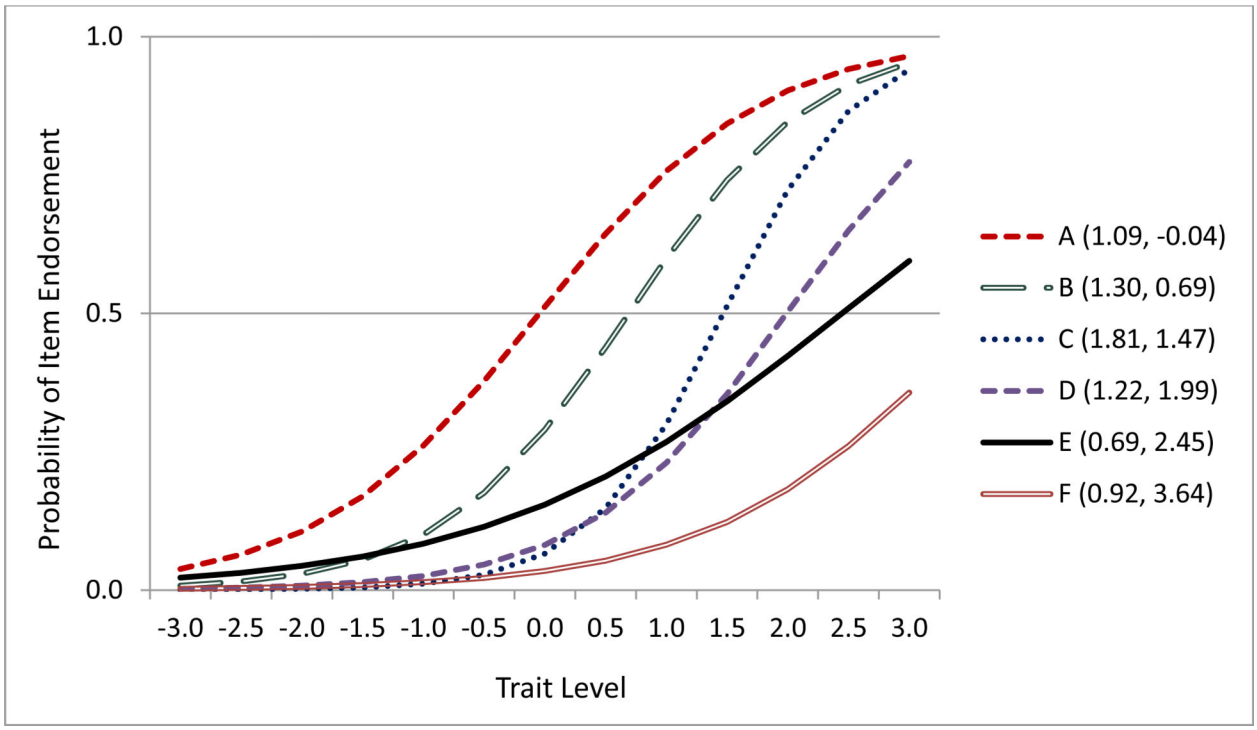

Figure 1.

Item characteristic curves of several items in a two-parameter logistic (2-PL) item response theory (IRT) model.

A = While drinking, I have said or done embarrassing things; B = Said things while drinking that you later regretted; C = The quality of my work or school work has suffered; D = Told by a friend or neighbor to stop or cut down on drinking; E = Gotten into trouble at work or school; F = Almost constantly think about drinking alcohol. Numbers in parenthesis indicate item discrimination and severity parameters, respectively.

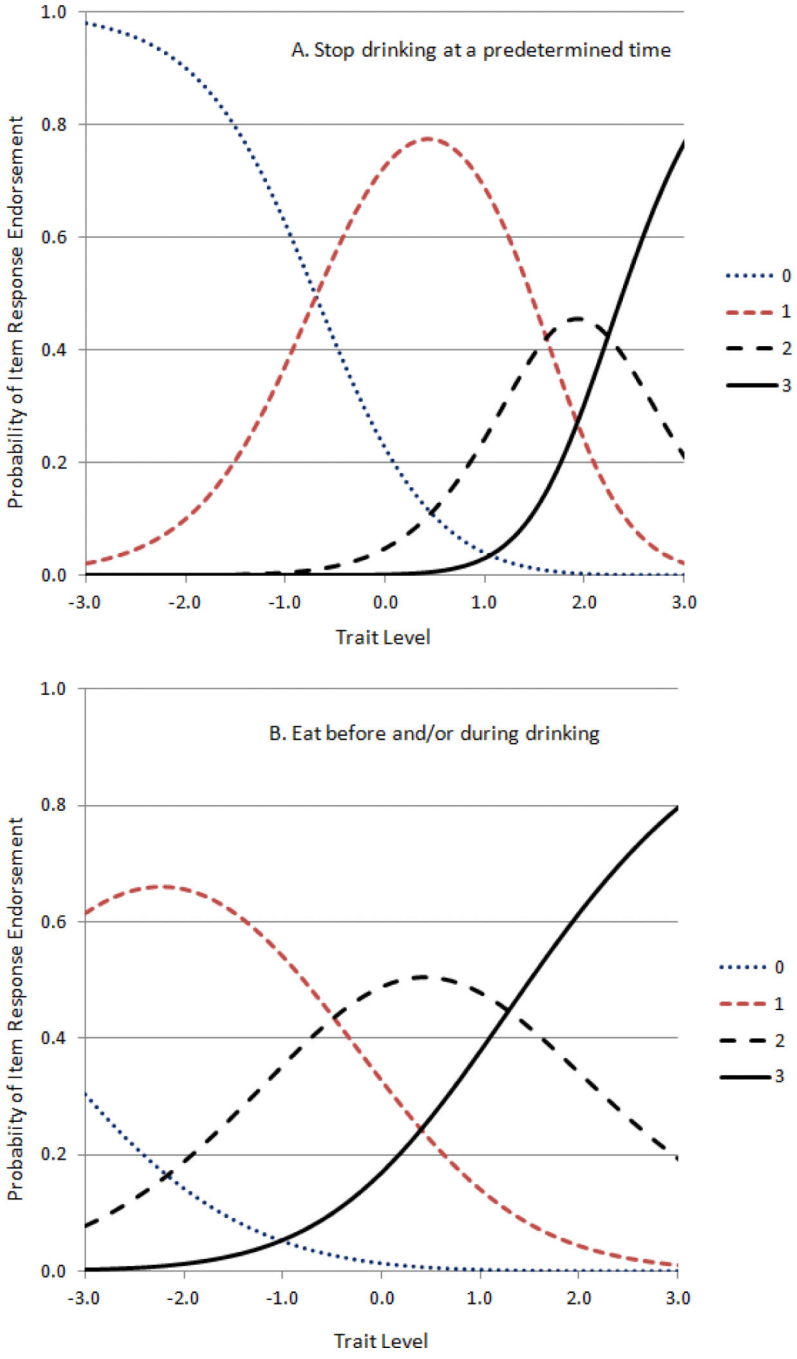

Figure 2.

Category response curves of two items (Figure 2A and Figure 2B) of protective behavioral strategies from the generalized partial credit IRT model.

Item A ("*Stop drinking at a predetermined time*") in Figure 2A had a slope parameter of 0.99 with item step difficulty parameters of –0.69 (from 0 to 1), 1.62 (from 1 to 2), and 2.24 (from 2 to 3), respectively. Item B ("*Eat before and/or during drinking*") in Figure 2B had a

slope parameter of 0.49 with item step difficulty parameters of –3.85 (from 0 to 1), –0.48 (from 1 to 2), and 1.29 (from 2 to 3), respectively.

Table 1

Study Design Characteristics (N = 12,630)

Study	Representative Reference	Intervention[1]	College Campus	1st Year (%)	Men (%)	White (%)	*N*	*N* at Follow-up				
								1 mo.	2 mo.	3-4 mo.	6 mo.	9-12 mo.
Mandated College Students												
1	White, Mun, Pugh, and Morgan (2007)	BMI, WF	Large Public U. in the Northeast US	62	60	73	348			319		219
2	White, Mun, and Morgan (2008)	WF, Delayed WF	Large Public U. in the Northeast US	63	71	69	230		199		106[2]	
3	Barnett, Murphy, Colby, and Monti (2007)	BMI, AE	Large Private U. in the Northeast US	67	49	66	225			206		211
4	Cimini et al. (2009)	Group MI, Group Theatrical Presentation, AE	Large Public U. in the Northeast US	49	62	80	682				471	430
5	LaBrie, Lamb, Pedersen, and Quinlan (2006)	Group BMI	Mid-sized Private U. in the Southwest US	71	60	74	167	158	148	139	125	
6	LaBrie, Thompson, Huchting, Lac, and Buckley (2007)	Group BMI	Mid-sized Private U. in the Southwest US	49	0	58	115	110	110	110		
7.1	Fromme and Corbin (2004)	Peer/Professional LMC, Control	Large Public U. in the Southern US	58	76	75	124	106			61[3]	
Volunteer College Students												
7.2	Fromme and Corbin (2004)	Peer/Professional LMC, Control	Large Public U. in the Southern US	38	59	59	452	332			221	
8a	Larimer et al. (2007)	Feedback, Control	Mid-sized Public U. in the Northwest US	40	35	86	1,486					1,122
8b	Larimer et al. (2007)	Feedback, Control	Large Public U. in the Northwest US	37	41	64	2,155					1,618
8c	Larimer et al. (2007)	Feedback, Control	Small Public C. in the Northwest US	22	34	83	600					304

Study	Representative Reference	Intervention[1]	College Campus	1st Year (%)	Men (%)	White (%)	*N*	*N* at Follow-up				
								1 mo.	2 mo.	3-4 mo.	6 mo.	9-12 mo.
	Volunteer 1st-year or Incoming College Students											
10.2	Baer, Kivlahan, Blume, McKnight, and Marlatt (2001)	Non-High-Risk Control	Large Public U. in the Northwest US	100	41	78	87					81
11	Walters, Vader, and Harris (2007)	Feedback, Control	Large Public U. in the Southern US	100	59	64	383		272	288		
15	LaBrie, Huchting, et al. (2008)	Group BMI, Control	Mid-sized Private U. in the Southwest US	100	0	56	263	261	260	258		
16	LaBrie et al. (2009)	Group BMI, Control	Mid-sized Private U. in the Southwest US	100	0	57	287	282	277	268	250	
17	LaBrie, Pedersen, Lamb, and Quinlan (2007)	Group BMI	Mid-sized Private U. in the Southwest US	100	100	65	120	110	105	90	56	
22	Wood et al. (2010)	BMI, BMI + PBI, Control	Large Public U. in the Northeast US	100	43	87	758					687
	Volunteer Heavy Drinking College Students											
9	Lee, Kaysen, Neighbor, Kilmer, and Larimer (2009)	AE, ASTP, BASICS, Choices, Web BASICS, Control	Large Public U. in the Northwest US	100	38	71	604			504	485	
10.1	Baer et al. (2001)	BASICS, Control	Large Public U. in the Northwest US	100	46	84	348					322
12	Wood et al. (2007)	EC, BMI, Feedback + BMI, Control	Large Public U. in the Northeast US	4	47	91	335	276		257	258	
13	Murphy, Benson, and Vuchinich (2004)	BMI, Feedback	Large Public U. in the Southern US	13	32	94	54				51	
14	Murphy et al. (2001)	AE, BMI, Control	Large Public U. in the Southern US	41	46	94	84			79		79
21	Walters, Vader, Harris, Field, and Jouriles (2009); Walters, Vader, Harris, and Jouriles (2009)	BMI, BMI without Feedback, Feedback, Control	Mid-sized Private U. in the Southern US	41	35	84	288			261	252	251

Study	Representative Reference	Intervention[1]	College Campus	1st Year (%)	Men (%)	White (%)	*N*	*N* at Follow-up				
								1 mo.	2 mo.	3-4 mo.	6 mo.	9-12 mo.
Intercollegiate Student Athletes or Fraternity, Sorority, and Service Organization Members												
18	Martens, Kilmer, Beck, and Zamboanga (2010)	Targeted Feedback, Standard Feedback, AE	Two Small Private C. in the Northwest and Northeast US, Large Public U. in the Midwest US	32	26	85	329	289			259	
19	LaBrie, Hummer, Neighbors, and Pedersen (2008)	Group-specific Feedback, Control	Mid-sized Private U. in the Southwest US	19	31	67	1,178	966	922			
20	Larimer et al. (2001)	BASICS, Control	Large Public U. in the Northwest US	78	52	84	928					631

Notes. U. = University. C. = College

[1]Intervention is labeled as originally described in the published study (see Table S1 for new labels for intervention groups; see Ray et al., in press for the content analysis of BMIs). BMI = Brief Motivational Interviewing; WF = Written Feedback; AE = Alcohol Education; LMC = Lifestyle Management Class; BASICS = Brief Alcohol Screening and Intervention for College Students; ASTP = Alcohol Skills Training Program; EC = Expectancy Challenge; PBI = Parent-based Intervention. For ease of comparison, many conditions were relabeled based on key design features, and this information is provided in the Online Supplement.

[2]The delayed WF group (n = 119) received feedback at 2 months post baseline and thus their follow-up data at 6 months post baseline were excluded

[3]Mandated students who were in the control group (n = 24) received LMC at 1 month post baseline and their follow-up data at 6 month post baseline were excluded. Follow-up sample sizes were based on selective alcohol use measures and could differ from those reported in published articles.

Table 2

Heavy Episodic Drinking (HED) Variable as an Example when Harmonization was not Feasible

		2-Wk. HED	1-Mo. HED	Correlation
Study 21	Response option	0 = Never to 6 = 6 or more	0 = Never to 10 = 11 or more	
	Men (*n* = 102)	1.95 (0.84)	2.49 (0.82)	0.77
	Women (*n* = 186)	1.66 (0.83)	2.29 (0.88)	0.71
Study 22	Response option	0-31	0-31	
	Men (*n* = 305)	1.68 (2.40)	2.89 (3.65)	0.74
	Women (*n* = 424)	1.15 (1.91)	2.69 (3.48)	0.74

Notes. HED = Five drinks or more for men; four drinks or more for women at one sitting. Data at baseline were reported as an example.