SUPPLEMENTAL INFORMATION ON MEASURES AND ANALYSES

The following material contains additional details on measures and analytic methods that supplement what is in the main text.

Measures

**Alcohol use quantity.** For 15 out of the 17 studies included in this manuscript, typical weekly drinking was derived from a version of the Daily Drinking Questionnaire (DDQ; Collins et al., 1985), which asks participants to report the number of drinks they consume each day of a typical week. We summed the number of drinks consumed across seven days in a typical week. For studies 15 and 16, baseline values were derived from two single-item measures in which participants were asked to indicate the number of drinks they consumed on a typical drinking occasion, as well as the number of days they consumed alcohol in a typical week. We multiplied these two items to get the total number of drinks consumed in a typical week for these two studies. Follow-up scores for these two studies were derived from weekly online drinking diaries that asked about the number of drinks consumed on each day in the past week for 10 weeks. We summed the number of drinks per week and then averaged these weekly drinking numbers across the four weeks. In sum, all 17 studies (studies 2, 7 [7.1 and 7.2], 8a, 8b, 8c, 9, 10.1, 11, 12, 13, 14, 15, 16, 18, 20, 21, and 22) had typical weekly drinking data.

For peak drinking, data from 15 studies were analyzed. Twelve studies (studies 7, 8a, 8b, 8c, 9, 11, 13, 14, 18, 20, 21, and 22) included a single-item measure capturing the maximum number of drinks consumed on a given occasion, typically within the past month (or past 30 days). Of these 12 studies, studies 8a, 8b, and 8c asked: “Think of the occasion you drank the most this past month. How much did you drink?” Participants answered anywhere from 0 to 25. All other studies asked about the most number of drinks consumed on one occasion in the past
month using an open-ended response format. Studies 15 and 16 utilized the same, single-item question at baseline. At follow-up, however, scores were derived from retrospective diary-style reports described above for studies 15 and 16. For study 2, peak drinking was derived from a modified version of the DDQ where students reported the number of drinks consumed each day during the heaviest drinking week in the past month. The maximum number of drinks during the heaviest drinking week was defined as peak drinking for study 2.

Because of the different response options for peak drinking, we harmonized peak drinking values across studies. Studies 8a, 8b, and 8c restricted responses for the maximum number of drinks to 25 drinks or more. All other studies included in the present analysis did not impose a cap. We placed this cap of 25 drinks across all other studies, which resulted in zero-inflated distributions with generally gradual tails (i.e., minimal right censoring of data) across studies.

Note that study 10.1 also assessed peak drinking but capped responses at nine drinks (i.e., more than eight drinks). Nine drinks or more for peak drinking turned out to be too low a threshold. We thus removed study 10.1 for the analysis for this outcome. No comparable measure of peak drinking was available for study 12. In conclusion, the extent of harmonization at the question level was minimal for peak drinking. As for values, we harmonized response values by capping them at 25 drinks in the present analysis.

**Alcohol problems.** All 17 studies included in the present study had alcohol problems latent trait scores. These scores were derived based on a hierarchical, two-parameter logistic IRT (2-PL IRT) model, which is fully reported in Huo et al. (2014). A more accessible explanation of our IRT analysis can be found in Mun et al. (2014). The six questionnaires used to derive the alcohol problem latent trait (θ) scores were the Rutgers Alcohol Problem Index (RAPI; White
The present analyses used Bayesian multilevel models (MLMs) estimated with Markov chain Monte Carlo (MCMC) sampling to model the three levels of the data (repeated measures, individuals, and studies). Bayesian analysis using MCMC sampling yields vectors of values, the posterior distribution of the regression coefficients and variance terms, instead of a single, point estimate for each parameter in the model. However, to incorporate studies with more than one active treatment condition, we used a three-level data structure but defined the highest level as the within-study randomization group (i.e., a unique condition within a study that is either an active intervention or a control group). The rationale for this is described below. The following model represents the base model for each of the three outcomes (drinks per week, peak drinks, and alcohol problems) evaluated in the present analyses:

\[
\text{Outcome}_{itg} = b_0 + b_1 \text{BaselineOutcome}_{ig} + b_2 \text{Male}_{ig} + b_3 \text{NonWhite}_{ig} + b_4 \text{FirstYear}_{ig} + b_5 \text{Mandated}_{ig} + u_{0g} + r_{0ig} + e_{tig},
\]

where \(t\) indexes repeated measures, \(i\) indexes individuals, and \(g\) indexes unique randomization groups. The outcome variable contains post-baseline observations, whereas baseline values of the outcome are treated as a covariate. Additional covariates included gender, mandated status, race, and whether or not the student was in his/her first year of college. The last three terms on
the right side of the model equation shown above represent variability between randomization groups \((u_{0g})\), between individuals within randomization groups \((r_{0ig})\), and residual error \((e_{tig})\), respectively.

A notable unique specification in the formulation of the model above is that *study by treatment* (i.e., unique randomization group) was defined as the highest level in the model, as opposed to study. A randomization group is the group to which a participant was randomly assigned, so if a parallel arm randomized controlled trial evaluated two interventions it would contain a total of three randomization groups (2 interventions and 1 control). The present data contained a total of 38 randomization groups (21 interventions and 17 controls) across 17 studies. By defining study by treatment (i.e., 38 randomized groups) as the highest level of the MLM, our model is a derivation of a typical MLM where study is defined at the highest level of the model and study-specific intervention effects are estimated by including a slope of treatment. The advantage of our current model is that non-existent study by treatment combinations can be dropped from the model. This can be conceptualized through the analogy of converting a two-way ANOVA into an equivalent one-way ANOVA. Estimating study-specific treatment slopes via MLM analysis is analogous to a two-way ANOVA where the between-subjects factors are study and treatment type. Thus, the present study can be seen as a \(17 \times 4\), study by treatment, factorial design with actual observations only in 38 out of possible 68 conditions. However, all of the unique study by treatment combinations of the two-way ANOVA can be defined as separate groups in an equivalent one-way ANOVA.

Another thing to note is that no fixed effect for treatment was specified in the equation since the samples from the posterior distributions of the random effects \((u_{0g})\) for the 38 randomization groups were used to estimate intervention effects within each study. Specifically,
the posterior distribution of the difference in random effects between each intervention group and its corresponding control group (e.g., $u_0^{\text{intervention}} - u_0^{\text{control}}$) within studies provided an effect size estimate and its confidence interval for each randomized intervention group shown in a forest plot (Figure 3) and Table 2, by using the mean and highest probability density interval of the corresponding posterior distribution, respectively. An effect was statistically significant if the 95% posterior probability density interval did not include zero.
SUPPLEMENTAL REFERENCES


### Supplemental Table. Covariate Estimates for Outcome Models Aggregated across Follow-ups 1 to 12 months Post Baseline

<table>
<thead>
<tr>
<th></th>
<th>Drinks per Week</th>
<th>Peak Drinks</th>
<th>Alcohol Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Logit</td>
<td>Count</td>
<td>Logit</td>
</tr>
<tr>
<td></td>
<td>OR 95% CI</td>
<td>RR 95% CI</td>
<td>OR 95% CI</td>
</tr>
<tr>
<td>No BL alcohol use vs. BL alcohol use</td>
<td>0.01 0.01 0.02 0.75 0.70 0.80 0.04 0.03 0.05 0.99 0.94 1.05</td>
<td>– – –</td>
<td>– – –</td>
</tr>
<tr>
<td>BL alcohol quantity</td>
<td>4.78 3.82 5.87 1.59 1.55 1.62 4.04 3.21 5.01 1.44 1.42 1.47</td>
<td>– – –</td>
<td>– – –</td>
</tr>
<tr>
<td>BL alcohol problems</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
</tr>
<tr>
<td>Men vs. women</td>
<td>1.08 0.85 1.40 1.31 1.25 1.37 1.05 0.94 1.21 1.02 1.01 1.03</td>
<td>0.06 0.02 0.09</td>
<td></td>
</tr>
<tr>
<td>Non-white vs. white</td>
<td>0.67 0.52 0.85 0.87 0.83 0.91 0.58 0.45 0.76 0.89 0.86 0.92</td>
<td>-0.03 -0.07 0.01</td>
<td></td>
</tr>
<tr>
<td>First-year vs. non first-year</td>
<td>0.72 0.54 0.93 1.11 1.05 1.17 0.79 0.60 1.07 1.07 1.03 1.10</td>
<td>0.04 0.00 0.08</td>
<td></td>
</tr>
<tr>
<td>Mandated vs. volunteer</td>
<td>0.48 0.15 1.45 0.74 0.60 0.93 0.73 0.24 2.72 0.89 0.77 1.03</td>
<td>-0.37 -0.54 -0.21</td>
<td></td>
</tr>
</tbody>
</table>

*Note. BL alcohol quantity measures were drinks per week and peak drinks at baseline for their respective outcome model. Results highlighted in bold are statistically significant.*