Mapping the Continuum of Alcohol Problems in College Students:
A Rasch Model Analysis

Christopher W. Kahler
Brown University

David R. Strong
Butler Hospital and Brown Medical School

Jennifer P. Read
Brown University

Tibor P. Palfai
Boston University

Mark D. Wood
University of Rhode Island

The authors conducted Rasch model (G. Rasch, 1960) analyses of items from the Young Adult Alcohol Problems Screening Test (YAAPST; S. C. Hurlbut & K. J. Sher, 1992) to examine the relative severity and ordering of alcohol problems in 806 college students. Items appeared to measure a single dimension of alcohol problem severity, covering a broad range of the latent continuum. Items fit the Rasch model well, with less severe symptoms reliably preceding more severe symptoms in a potential progression toward increasing levels of problem severity. However, certain items did not index problem severity consistently across demographic subgroups. A shortened, alternative version of the YAAPST is proposed, and a norm table is provided that allows for a linking of total YAAPST scores to expected symptom expression.

Heavy drinking is prevalent on college campuses (O’Malley & Johnston, 2002; Presley, Meilman, & Cashin, 1996; Wechsler, Lee, Kuo, & Lee, 2000) and is associated with myriad negative consequences. These consequences vary in frequency and severity (Perkins, 2002), ranging from mild negative outcomes, such as being late for class, to more severe consequences, such as interpersonal aggression and sexual victimization (Abbey, 2002; Donovan & McEwan, 1995; Perkins, 2002; Presley et al., 1996). Heavy drinking and its consequences among college students have been the focus of much recent attention in both the scientific (Boyd & Faden, 2002; Goldman, 2002) and the popular literature (Thompson, 1998; Winerip, 1998). Yet although it is clear that college is a developmental period of enhanced risk for alcohol-related con-sequences, the actual scope of alcohol-related consequences among college students remains somewhat elusive, in part because of the way in which alcohol problems have been measured, and therefore described, in this population (see Perkins, 2002; Wingo, 1998).

A number of measures have been devised for measuring alcohol problems in college students (cf. Hurlbut & Sher, 1992; Maddock, Laforge, Rossi, & O’Hare, 2001; O’Hare, 1997; Williams & Morrice, 1992), and measures developed and validated in younger adolescents also have been widely used in college populations (e.g., White & Labouvie, 1989). However, psychometric validation of these measures to date has been limited to methods from classical test theory (cf. Nunnally, 1979). Indices of a measure’s integrity based in classical test theory, such as internal consistency analyses, exploratory and confirmatory factor analyses, and concurrent validity analyses using correlations, rely primarily on omnibus statistics that average across levels of individual variation and can be influenced significantly by sample characteristics. These analyses do not directly address the adequacy of the response options used in a scale, the relative severity of individual items (beyond the simple frequency of endorsement), or the quality of the scale as a function of different levels of the construct being measured.

The typical approach to the measurement of alcohol problems in college students is to ask about the frequency of individual alcohol problems occurring in a specified time period and then to sum these frequencies to form a total alcohol problem index. This index is presumed to cover a continuum ranging from no alcohol-related problems to severe alcohol-related problems. In such a scale, all
items contribute equally to the total score, and no consideration is
given to the severity of problems reflected by each item. For
example, being arrested twice for driving under the influence of
alcohol would contribute 2 points to the total score, as would
having twice a hangover after drinking. The former symptom
is likely to denote a far greater level of alcohol problems than the
latter, yet this would not be reflected in the total score. Although
a simple additive approach may be adequate for most research and
clinical purposes, the severity of an individual’s alcohol problems
may not be represented adequately when indexed solely by the
number of symptoms present. Problem severity also is likely to be
related to which particular symptoms are present. Evaluating
the likelihood of specific symptoms as a function of overall severity of
problems can increase researchers’ understanding of alcohol prob-
lems by providing information about the severity of particular
symptoms in relation to one another, the ordering and pattern of
symptoms most likely to be expressed, and the magnitude of
differences reflected by endorsement of additional symptoms.

A wide range of item response models are available that provide
significant advantages over classical methods in exploring the
relations between items and the underlying latent construct they
are designed to represent (see van der Linden & Hambleton, 1997,
for a description of methods). Some approaches begin with a
specific model of how the relationship between the probability of
an item response and an individual’s level of the underlying trait
should look and then use various model-fit statistics to determine
the number of parameters needed to best represent the available
data (cf. Birnbaum, 1968). Other item response model approaches
are much less restrictive, making few assumptions about how item
characteristics should look, and rely on nonparametric methods
(e.g., Ramsay, 2001) to describe how well actual item and option
characteristics conform to assumptions that are formalized in other
models.

The Rasch model (Rasch, 1960) is one particularly appealing
method for analyzing item responses to learn more about the
performance of scales and the underlying constructs being mea-
sured. The Rasch model has recently been used in addictive-
behaviors research to evaluate screens for problem drinking sever-
ity (Cornel, Knibbe, van Zutphen, & Drop, 1994), nicotine
dependence criteria (Strong, Kahler, Ramsey, & Brown, 2003),
benzodiazepine dependence criteria (Kan, Bretelet, van der Ven, &
Zitman, 1998), and gambling consequences and attitudes (Strong,
Lesieur, Breen, Stinchfield, & Lejuez, 2004). The Rasch model,
which has been described as the only method to “transform raw
data from the human sciences into abstract equal-interval scales”
(Bond & Fox, 2001, p. 7), is a logistic item response model that
independently scales the severity of both items and persons along
a theorized underlying latent continuum. The odds of an individual
endorsing a given item is seen as a function of both the individual’s
overall level of problem severity and the severity of that item
(Wright & Masters, 1982). The Rasch model scales both items and
persons using the same metric, an equal interval logit scale (i.e.,
log odds scale—hence the term logistic). Logit units accurately
reflect the true magnitude of the difference between two propor-
tions. For example, an item endorsed by 40% of individuals is only
slightly more difficult (or severe) than one endorsed by 50% of
individuals: odds ratio = 1.5, log odds = .41. By contrast, the
difference in difficulty between an item with 1% endorsement and
one with 11% endorsement is far greater: odds ratio = 12.2, log
 odds = 2.5.

The Rasch model is particularly useful in that estimates of item
severities can be made independent of the level of the construct
represented in a sample (cf. Andrich, 1988). Examination of these
severity estimates can reveal gaps along the continuum where
there are no items with a matching severity level. For example, a
scale may contain a cluster of low-severity items and a cluster of
high-severity items but may have few items of moderate severity.
Severity estimates also can reveal how well the items are matched
to a particular sample. For example, a scale may contain primarily
high-severity items but be applied to a population in which few
individuals have such severe problems. In this case, the scale
would not provide much information about relative differences
within the majority of the sample.

The Rasch model also evaluates patterns in endorsement of
symptoms that result in each total score. The assumption is that
with fit to the Rasch model, endorsement of problems should show
a probabilistic linear progression in which less severe symptoms
are likely to be present prior to more severe symptoms. For
example, Symptom A (e.g., getting in trouble or performing poorly
at work because of drinking) may be a precursor to Symptom B
(e.g., getting fired from work because of drinking) so that one
would expect, in most instances, to experience A first, followed by
B, rather than to experience B followed by A. Therefore, the most
usual response pattern would be endorsement of Symptom B but
not Symptom A. A Rasch model analysis provides estimates of the
degree to which response patterns conform to or fit with a pre-
dictable progression from less severe to more severe problems.

Finally, Rasch model analyses also can be used to examine the
extent to which a set of items provides consistent measurement
across populations. An assumption of most measures is that indi-
vidual characteristics do not affect interpretation of total scores,
and thus equal comparisons can be made across individuals from
different subgroups. However, some items may be biased. Items
are considered biased when individuals with the same overall level
of problem severity do not have the same probability of reporting
a particular problem. This bias can result from any number of
factors, including important demographic characteristics of a re-
spondent such as gender, race–ethnicity, regional or geographic
influences, socioeconomic status, and level of education. This
differential item functioning may lead to systematic errors in
placement of individuals along the alcohol problem severity con-
tinuum. Removing items with significant bias can ensure that a
given scale provides an accurate assessment of alcohol problems
across subgroups included in the sample (e.g., men vs. women,
Whites vs. non-Whites).

Given the relatively high rates and heterogeneous distribution of
alcohol-related consequences in the college population, Rasch
model analyses are particularly well suited to extend psychometric
validation of measures of alcohol problems beyond classical test
theory approaches. Furthermore, this analytic approach holds ad-
ditional promise for advancing conceptual understanding of the
nature and interrelations of alcohol-related consequences in an
at-risk population. Accordingly, the purpose of the present study
was to apply a Rasch model analysis to an existing measure of
college alcohol problems, the Young Adult Alcohol Problems
Screening Test (YAAPST; Hurlbut & Sher, 1992). The YAAPST
is one of the most commonly used measures of alcohol problems
in college students (Allen & Columbus, 1995). Using classical test theory approaches, this measure was developed to assess a range of negative consequences commonly associated with college drinking, such as driving while intoxicated, being sick from drinking or having a hangover, and missing work/school as a result of drinking. The YAAPST has demonstrated strong psychometric properties (Hurlbut & Sher, 1992) and has been used to assess college student drinking consequences in prior studies (e.g., Hurlbut & Sher, 1992; Palfai & Wood, 2001; Wood, Read, Palfai, & Stevenson, 2001).

In this study, we used Rasch model analyses to answer the following questions. To what extent do the items of the YAAPST measure a single dimension of alcohol problem severity? What is the relative severity and ordering of the alcohol problems in the measure? Is there significant differential item functioning according to gender, race–ethnicity, and college setting? How many meaningful levels of discrimination can be made with the scale? Prior to conducting these Rasch model analyses, we conducted exploratory analyses to examine the extent to which the use of multiple response options on the YAAPST provided useful levels of discrimination and should be retained in the Rasch model analyses. In prior analyses of items measuring alcohol dependence, dichotomous scoring of multiresponse items appeared most reliable (Kahler, Strong, Hayaki, Ramsey, & Brown, 2003; Kahler, Strong, Stuart, Moore, & Ramsey, 2003), and we expected that dichotomous scoring also may be preferable with the YAAPST.

Method

Participants

Participants were 806 college students, aged 18 to 24, who reported drinking alcohol at least once in the past year. Participants were recruited from introductory undergraduate psychology classes at two universities in the northeastern United States. Sample A consisted of 418 participants drawn from a midsized state university in a rural environment. Sample B was predominantly White (88.5%, n = 369). The remainder of this sample identified themselves as Asian/Pacific Islander (4.1%, n = 17), Hispanic (3.1%, n = 13), Black (2.2%, n = 9), Native American/Alaskan Native (0.5%, n = 2), or "other" (1.7%, n = 7). Women constituted approximately 66% (n = 277) of Sample A, and the mean age of the sample was 18.8 years (SD = 1.0), ranging from 18 to 22. Freshmen (62.6%) and sophomores (25.9%) were overrepresented, and juniors (7.4%) and seniors (2.9%) were underrepresented in the sample. There also was a small number of nonmatriculated/professional students (1.2%).

Sample B consisted of 388 participants drawn from a midsized private university in an urban environment. Although there was more ethnic diversity in Sample B compared to Sample A, participants were predominantly White (74.5%, n = 289). The remainder of the Sample B participants identified themselves as Asian/Pacific Islander (14.7%, n = 57), Hispanic (5.4%, n = 21), Black (1.3%, n = 5), Native American/Alaskan Native (1.8%, n = 7), or "other" (2.3%, n = 9). Approximately 56% (n = 215) of the Sample B participants were women, and the mean age of the sample was 18.8 (SD = 1.0) years, ranging from 18 to 24. Freshmen (60.1%) and sophomores (25.5%) were overrepresented, and juniors (10.3%) and seniors (4.1%) were underrepresented in the sample.

Procedure

Procedures and measures were uniform across both samples. Students were invited by campus postings and in-class announcements to participate in a study of "attitudes, perception, and personality." Participants provided signed informed consent and then completed a battery of questionnaires in exchange for academic credit. Questionnaires were administered during the fall and winter semesters by two of the researchers and research assistants. The data collection sessions lasted approximately 40 min. Participants were given an educational debriefing form, and any additional questions pertaining to the study were addressed.

Participants provided demographic information and information about the quantity and frequency of their drinking in the past year. In addition, all participants completed the YAAPST (Hurlbut & Sher, 1992). The YAAPST contains 27 items measuring potential negative consequences of alcohol use occurring in the past 12 months. Eight of the items are measured on a 10-point scale that ranges from 0 (No, never) and 1 (Yes, but not in the past year) to 9 (40 or more times). Eleven items are measured with a 5-point scale that ranges from 0 (No, never) and 1 (Yes, but not in the past year) to 4 (3 or more times in the past year). Finally, 6 items are measured with a 3-point scale that includes 0 (No, never), 1 (Yes, but not in the past year), and 2 (Yes, in the past year). The YAAPST has been shown to have a single-factor structure; good internal consistency; and good concurrent validity with indices of drinking, alcohol expectancies, and alcohol abuse and dependence symptoms (Hurlbut & Sher, 1992). For this analysis, we collapsed the responses 0 (No) and 1 (Yes, but not in the past year) into one negative response category, as our primary interest was in measuring alcohol problems in the past year.

Analyses

We conducted analyses in a planned progression to determine (a) the optimal number of response options to be retained in the Rasch model analyses, (b) the unidimensionality of the scale, (c) the relative severity of items on an equal-interval scale, (d) the fit of the observed ordering of items to the Rasch model, and (e) the presence of differential item functioning.

Before conducting Rasch model analyses, we examined the multiple response options on the YAAPST to determine whether the use of these options was adding meaningful levels of discrimination. Given the exploratory nature of this analysis and minimal prior analyses of YAAPST item characteristics, we had no basis on which to assume how YAAPST items would behave. Also, because some items have nine response options, it might be difficult to derive stable parameter estimates for all possible response options given that some responses might be rarely endorsed. Thus, we examined these option responses using a graphical, nonparametric technique that requires minimal assumptions about item characteristics.

For these exploratory purposes, we used a nonparametric kernel smoothing method and software (TestGraf) developed by Ramsay (Ramsay, 2001) to estimate option characteristic curves (OCC) for each item. These methods have been used in previous research on the performance of scales measuring depression (Santor & Coyne, 2001; Santor, Ramsay, & Zuroff, 1994; Santor, Zuroff, Ramsay, Cervantes, & Palacios, 1995), nicotine dependence (Strong, Brown, Ramsey, & Myers, 2003), and alcohol dependence (Kahler, Strong, Hayaki, et al., 2003; Kahler, Strong, Stuart, et al., 2003). To create the OCCs, individuals are ranked according to their total scores on the instrument. OCCs are then constructed by estimating the probability of endorsing each item option at specified evaluation points within these scores using a local average (Ramsay, 1991). We inspected the OCCs to determine whether individual response options followed a consistent pattern in which (a) each option has a distinct region on the continuum in which it is the most likely option to be endorsed and (b) more severe options become more likely to be endorsed than less severe options as problem severity increases. On the basis of these inspections, we collapsed response categories that were not contributing to the item’s discrimination ability.

After adjusting response options for optimal item performance, we conducted Rasch model analyses of the items. A primary assumption of the Rasch model is that responses to items are a function of individual
variation along a single underlying dimension (i.e., unidimensionality). Second, responses to a given item should be independent from responses to other items (i.e., locally independent). This second assumption is similar to but goes beyond the assumption of unidimensionality. For example, asking about the same symptom twice will not affect an index’s dimensionality, but individuals would be expected to report the same answer to both questions, and thus the identical items would not be locally independent (Wright, 1996). Therefore, before interpreting the fit of the data to the Rasch model, we conducted a principal-components analysis of residual variation after the Rasch model was fit. Additional factors accounting for greater than 1.5 units of variance are considered significant, and symptoms loading on these factors are examined further (Linacre, 1998; Smith & Miao, 1994).

Following the principal-components analysis of residuals, we examined the relative severity of each item and examined infit and outfit statistics for each item to determine how well the data fit the Rasch model (Wright & Masters, 1982). Infit and outfit statistics are sensitive to violations of assumptions of the Rasch model, such as unexpected variation in response patterns (e.g., a more severe item is endorsed by a number of individuals who do not endorse less severe items) or unequal slopes (i.e., item discrimination) across item response functions (Linacre & Wright, 1994). The outfit statistic is based on conventional sum of squared standardized residuals and is affected by outlying and off-target responses (e.g., responses to severe items by individuals with low levels of alcohol problems). The infit statistic is an information-weighted sum and is affected by unexpected patterns among on-target observations (e.g., the level of problem severity reflected in the item is similar to the level of alcohol problems expected in the individual). When the ordering of symptom severities fits the Rasch model, values of observed minus expected scores (infit and outfit) will fall within an acceptable range of 0.6–1.4 (Linacre & Wright, 1994). Infit statistics are generally preferred, as they are weighted locally and thus are less susceptible to outlier influences (Bond & Fox, 2001).

Following these initial analyses in the full sample, we conducted analyses to determine whether there was significant differential item functioning (DIF) across demographic subgroups. The estimation of DIF involves comparing analyses conducted separately within each group (Holland & Wainer, 1993). If items behave similarly across groups, then severity parameters estimated independently in different samples will fall within an acceptable range of agreement (e.g., 95% confidence interval). In this study, we compared women to men, non-White students to White students, and students from University B to students from University A. We examined items that had significantly different severity estimates across subgroups. Of these, we considered differences greater than 1 logit a clinically meaningful difference, equal to about one half of the sample standard deviation of item severities, that is, a medium effect size. We eliminated items showing DIF of this magnitude.

We reran Rasch model analyses using only items that fit the Rasch model well and did not show large DIF across subgroups. On the basis of these results, we examined the range of alcohol problem severity covered by the YAAPST and the extent to which the items target the level of alcohol problems that are likely to be found among college student drinkers. We used BIGSTEPS (Wright & Linacre, 1998) for all Rasch model analyses. Finally, we ran correlation and regression analyses to examine whether the relationship between alcohol consumption variables and YAAPST scores differed once the YAAPST scoring was empirically modified.

Results

Examination of Response Options

An examination of the OCC for each of the 27 YAAPST items revealed that the use of multiple response options did not improve discrimination of the items. In all cases, except those where there were extremely few positive responses (< 5%), a negative response (i.e., a response indicating that the respondent had not experienced the given problem) showed a consistent decline in probability of endorsement with increasing level of alcohol problem severity, as would be expected. However, the remaining response options showed uneven patterns of endorsement, with the highest frequency option often being more likely to be endorsed than the lower nonzero responses across a wide range of the continuum.

The left-hand panel of Figure 1 shows three examples of option characteristic curves: Items 1, 2, and 13. Item 1 (driving while intoxicated) has nine response options. Although Response Option 0 (i.e., not in the past year) becomes less likely with greater severity, the other options are not clearly differentiated, suggesting that the options may be adding noise to the assessment. Although Item 2 (a less severe item: having a hangover after drinking) shows better ordering of response options, no option other than 0 is ever endorsed by more than one third of the sample at any point along the continuum. Finally, Item 13 (neglecting social obligations for 2 or more consecutive days because of drinking) has four response options. In this relatively more severe item, the only response options that are clearly the most likely to be endorsed in a given region of the continuum are Option 0 (never) and Option 3 (3 or more times). The endorsement of the other two options does not show a clear pattern in which more severe options become substantially more likely to be endorsed than less severe options as severity increases. On the basis of our examination of these OCCs, and the utility and interpretational simplicity of having all items share a common metric, we chose to dichotomize all items as either present or absent in the past 12 months. The right-hand panel of Figure 1 shows examples of the improved OCC when Items 1, 2, and 13 are dichotomized. With only two response options, the point at which the two lines intersect in each OCC represents the region of the alcohol problem severity continuum that a specific problem has a 50% probability of being reported.

Unidimensionality

Principal-components analysis of residual variance after fitting the Rasch model to the 27 dichotomized YAAPST items did not fully support the unidimensionality of this set of items, as the first residual factor accounted for 1.84 units of variance, greater than the target of 1.50 units. The second residual factor was within limits, accounting for 1.43 units of variance. Examination of loadings on the first factor indicated that Item 4 (late for work or school because of drinking) and Item 5 (missed work or classes because of drinking) loaded at .76 and .77, respectively, with no other item loading higher than .21. This result suggests that these items violate the assumption of local independence (i.e., they are too correlated relative to the correlations among the other items). Because this is likely due to content overlap, we decided to eliminate Item 4, which was less commonly endorsed than Item 5 even though being late should conceptually be a less severe problem than being absent.

We reran the Rasch model analysis with Item 4 deleted and again found that the first of the residual principal components accounted for more than 1.50 units of variance, with a value of 1.53. This factor again included two items with very similar content: (a) Item 10 (partner or near relative complained about
Figure 1. Probability of endorsing each item response option for Young Adult Alcohol Problems Screening Test (YAAPST) Items 1 (drove car while intoxicated), 2 (hangover after drinking), and 13 (neglected obligations) as a function of standardized expected total YAAPST score. The left panels show the characteristics of the options when dichotomized as absent (0) versus present (1).
drinking; loading of .41) and (b) Item 11 (drinking caused problems with partner or near relative; loading of .42). We chose to drop Item 10 from the scale because Item 11 had a more general wording that more clearly denotes negative consequences. The principal-component analysis of the residual variance after the Rasch model was fit for the 25 remaining items indicated that the first component now accounted for only 1.50 units of variance, suggesting that problems of local dependence within the measure were minimized.

Item-Level Estimates of Problem Severity

The Rasch model provides estimates of the level of alcohol problems reflected in each symptom (Andrich, 1988). Table 1 lists the item-level estimates of the severity of the 25 retained YAAPST items listed in descending order of severity, which is indicated by the measure parameter for each item. The frequency of endorsement of each of the YAAPST items is also presented in this table. Items 4 (late for work or school because of drinking) and 10 (partner or near relative complained about drinking) do not have entries for Rasch model parameters because these were eliminated previously on the basis of analysis of model residuals (see above).

The severity estimates, which are expressed in equal-interval logit units, indicate the region along the latent problem-drinking continuum where that individual symptom is making discriminations. The estimates of each item’s severity (typically called the measure estimate) in a Rasch analysis is standardized so that the average severity of items is given a value of 0. Items with positive severity estimates are targeting alcohol problems that are relatively more severe than the mean level of severity targeted by the set of items, whereas those with negative estimates are targeting problems that are relatively less severe than the mean level of severity targeted by the items. The order of Rasch estimates generally mirrors the raw frequencies of endorsement of each item. However, the logit scale better reflects the relative distance between items than simple frequencies. For example, a difference in the frequency between symptoms closer to the center of the distribution of 7.8% (e.g., between driving while intoxicated [Item 1] and receiving a worse grade because of drinking [Item 15]) is equal to 0.62 logits, whereas at the highest levels a raw difference of 5.4% (e.g., going to someone for help [Item 25] and experiencing trouble at school or work [Item 7]) is equal to 2.22 logits. Table 1 also lists the standard errors of the severity estimates. The most severe items

<table>
<thead>
<tr>
<th>Item no.</th>
<th>Item content</th>
<th>% endorsed</th>
<th>Item severity</th>
<th>SE</th>
<th>Infit</th>
<th>Outfit</th>
<th>Gender</th>
<th>Race/ethnicity</th>
<th>College</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.</td>
<td>Arrested for DWI/DUI</td>
<td>0.5</td>
<td>3.50</td>
<td>0.50</td>
<td>1.11</td>
<td>0.89</td>
<td>−0.11</td>
<td>−1.94</td>
<td>1.88</td>
</tr>
<tr>
<td>26.</td>
<td>Attended Alcoholics Anonymous</td>
<td>0.6</td>
<td>3.27</td>
<td>0.45</td>
<td>1.02</td>
<td>0.93</td>
<td>1.29</td>
<td>−0.54</td>
<td>0.34</td>
</tr>
<tr>
<td>27.</td>
<td>Sought professional help</td>
<td>0.6</td>
<td>3.26</td>
<td>0.45</td>
<td>1.04</td>
<td>1.24</td>
<td>2.22</td>
<td>0.39</td>
<td>0.37</td>
</tr>
<tr>
<td>8.</td>
<td>Fired or suspended from school</td>
<td>0.7</td>
<td>3.08</td>
<td>0.41</td>
<td>0.95</td>
<td>0.93</td>
<td>0.62</td>
<td>0.57</td>
<td>0.04</td>
</tr>
<tr>
<td>25.</td>
<td>Gone to someone for help</td>
<td>0.9</td>
<td>2.91</td>
<td>0.38</td>
<td>0.85</td>
<td>0.27</td>
<td>2.59</td>
<td>−0.1</td>
<td>−0.17</td>
</tr>
<tr>
<td>17.</td>
<td>Arrested for drunken behavior</td>
<td>6.3</td>
<td>1.90</td>
<td>0.24</td>
<td>0.95</td>
<td>0.62</td>
<td>1.72</td>
<td>1.76</td>
<td>0.42</td>
</tr>
<tr>
<td>12.</td>
<td>Lost friends due to drinking</td>
<td>3.5</td>
<td>1.40</td>
<td>0.20</td>
<td>1.07</td>
<td>1.93</td>
<td>−0.27</td>
<td>1.56</td>
<td>0.3</td>
</tr>
<tr>
<td>20.</td>
<td>Need drink upon awakening</td>
<td>4.7</td>
<td>1.05</td>
<td>0.17</td>
<td>0.98</td>
<td>0.80</td>
<td>0.71</td>
<td>0.12</td>
<td>−0.28</td>
</tr>
<tr>
<td>24.</td>
<td>Doctor said drinking harmful</td>
<td>5.0</td>
<td>0.99</td>
<td>0.17</td>
<td>1.11</td>
<td>1.16</td>
<td>0.69</td>
<td>−0.04</td>
<td>0.38</td>
</tr>
<tr>
<td>7.</td>
<td>In trouble at school or work</td>
<td>6.3</td>
<td>0.69</td>
<td>0.15</td>
<td>1.02</td>
<td>1.20</td>
<td>0.87</td>
<td>0.53</td>
<td>0.17</td>
</tr>
<tr>
<td>19.</td>
<td>Had the shakes</td>
<td>7.0</td>
<td>0.57</td>
<td>0.15</td>
<td>0.97</td>
<td>1.09</td>
<td>−0.32</td>
<td>0.66</td>
<td>−0.22</td>
</tr>
<tr>
<td>22.</td>
<td>Felt dependent on alcohol</td>
<td>10.3</td>
<td>0.06</td>
<td>0.13</td>
<td>0.99</td>
<td>1.02</td>
<td>−0.37</td>
<td>−0.16</td>
<td>0.17</td>
</tr>
<tr>
<td>6.</td>
<td>Physical fights while drinking</td>
<td>11.5</td>
<td>−0.10</td>
<td>0.12</td>
<td>0.94</td>
<td>0.81</td>
<td>1.05</td>
<td>0.52</td>
<td>−0.10</td>
</tr>
<tr>
<td>9.</td>
<td>Damaged property</td>
<td>13.2</td>
<td>−0.28</td>
<td>0.11</td>
<td>0.97</td>
<td>0.89</td>
<td>1.83</td>
<td>0.48</td>
<td>−0.18</td>
</tr>
<tr>
<td>13.</td>
<td>Neglected obligations</td>
<td>15.5</td>
<td>−0.53</td>
<td>0.11</td>
<td>0.89</td>
<td>0.66</td>
<td>0.06</td>
<td>0.76</td>
<td>−0.59</td>
</tr>
<tr>
<td>11.</td>
<td>Problems with partner/relative</td>
<td>17.4</td>
<td>−0.70</td>
<td>0.10</td>
<td>1.01</td>
<td>0.94</td>
<td>−0.10</td>
<td>0.72</td>
<td>−0.24</td>
</tr>
<tr>
<td>15.</td>
<td>Worse grade due to drinking</td>
<td>17.9</td>
<td>−0.75</td>
<td>0.10</td>
<td>0.98</td>
<td>0.84</td>
<td>−0.46</td>
<td>0.58</td>
<td>−1.05</td>
</tr>
<tr>
<td>10.</td>
<td>Partner/relative complained</td>
<td>19.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Gender difference reflects measures from women compared with men, race/ethnicity difference reflects measures from non-Whites compared with Whites, and college differences reflect measure estimates from Sample A compared with Sample B.

* Items with large differences across groups, but level of bias is not statistically significant due to large error in estimates in this region of the continuum.  
* Items eliminated due to significant gender bias.  
* Item eliminated due to significant college sample bias.  
* Items eliminated due to local dependence with other items, Rasch estimates not derived.  
* Item eliminated due to significant race/ethnicity bias.
have the largest standard errors, because relatively few participants in the sample were endorsing these items, providing less precise estimates of these item severities. By contrast, items in the middle and low ranges of severity have small standard errors because estimates were derived by examining patterns among the large proportion of individuals who populate this region of the continuum.

The 25 analyzed YAAPST items (see Table 1) appeared to fit a unidimensional Rasch model well, with infit values ranging from 0.85 to 1.31, falling well within the target range of 0.60 to 1.40 (Linacre & Wright, 1994). Although outfit statistics for Item 23 (feeling guilty about drinking) and Item 12 (lost friends because of drinking) fell outside the range of 1.40, the infit statistics for these items suggest undue influence by extreme values, and therefore these items are considered to fit the model. Item severity estimates ranged from −4.45 to 3.50, suggesting that the items cover a broad range of severity of alcohol problems. The severity estimates show not only the interitem distance reflected by endorsement of each symptom but also the relative position of each symptom. For example, we expect that if an individual endorses having gotten into a sexual situation he or she later regretted (Item 14), then he or she also is likely to report having had a hangover (Item 2) and feeling very sick after drinking (Item 3).

**DIF**

To assess DIF, we ran Rasch model analyses in each subgroup of interest (i.e., men and women, Whites and non-Whites, Sample A and Sample B) and then created scatter plots of item-severity estimates obtained in separate subgroup analysis. If item-severity estimates are similar across subgroups, then plots of these estimates should fall on a 45° angle within bounds of the joint standard errors of the estimates from each analysis (e.g. 95% confidence interval). An example of one of these plots is displayed in Figure 2 for the comparison of item-severity estimates obtained for men and women. Any item that falls outside of the standard error lines is considered to have significant DIF. However, we chose to examine closely only those items for which significant DIF was equal to or greater than 1, to ensure that only items with large and potentially meaningful DIF were eliminated. In the

![Figure 2](image)

*Figure 2.* Plot of item severity estimates and 95% confidence intervals (CI) from differential item functioning analyses conducted within women and men. Items that fall below the lower 95% CI reflect more severe problems for women compared with men. Items 6 (physical fights while drinking), 9 (damage property while drinking), and 17 (arrested for drunken behavior) show significant and meaningful levels of bias.
comparison of severity estimates for men and women, Items 6 (physical fights while drinking), 9 (damage property while drinking), and 17 (arrested for drunken behavior) were significantly more severe items for women than for men; that is, men were more likely to endorse these problems than women at the same level of alcohol problem severity. The differences in logit values for these items were greater than 1, suggesting these items could introduce bias when attempting to place men relative to women on the alcohol problem severity continuum.

In comparing Whites and non-Whites, only Item 5 (missed work or classes because of drinking) showed a significant DIF that was greater than 1 logit. The direction of the DIF indicated that this item was relatively less severe for White students compared to non-White students. Finally, for the Sample A–Sample B comparison, only Item 15 (received a lower grade on an exam or paper because of drinking) showed a significant DIF that was greater than 1 logit, indicating this was a more severe item in Sample B relative to Sample A.

**Raw Score Distribution**

We reran Rasch model analyses with the 20 items that fit the model well and did not show substantial and significant DIF. From these analyses, we obtained a norm table that shows the raw scores on the measure (ranging from 0 to 20), the estimated level of alcohol problem severity associated with each score, the standard error of this estimate, and the frequency and cumulative frequency of each score. This information is presented in Table 2. The severity estimate represents the estimated level of problem severity, expressed in equal-interval logit units, that would be associated with a given raw score based on the Rasch model (Wright & Masters, 1982). For example, with these data a raw score of 4 is converted to a severity estimate of –2.72. Individuals with a score of 4 would have a .50 probability of endorsing an item that has an estimated severity of –2.72. They would be increasingly likely to endorse items at progressively lower severity and increasingly unlikely to endorse items of progressively higher severity. A comparison of the mean person-level estimates of alcohol problem severity to the mean item-severity estimates (standardized to 0) revealed that this combined sample of college student alcohol drinkers fell at –2.71 logit units (SE = 0.80), indicating that the YAAPST items are targeting a level of alcohol problems that are considerably more severe than those typically experienced among college students who have experienced at least some alcohol problems.

**Correlates**

The empirically derived raw score we created by summing the 20 dichotomized YAAPST items was highly correlated with the original YAAPST scoring, \( r(806) = .88 \). However, the distribution of the empirically scored YAAPST was superior, with skewness of 0.71 and kurtosis of 0.51, compared to skewness of 1.47 and kurtosis of 2.28 for the original YAAPST scoring. This improvement reflects primarily a reduction in the spread of scores due to dichotomization, as the sum of all 27 YAAPST items when dichotomized has a skewness of 0.85 and a kurtosis of 0.50.

The empirically scored YAAPST correlated .61 with frequency of drinking, .55 with average quantity of drinking, and .65 with frequency of drinking five or more drinks \( (p < .0001) \). Correlations between the original YAAPST and these measures were .64, .53, and .68, respectively. To examine the potential effects of removing items with a significant DIF, we examined whether relationships between consumption variables and alcohol problems

<table>
<thead>
<tr>
<th>Raw score</th>
<th>Severity estimate</th>
<th>SE</th>
<th>Frequency (n)</th>
<th>% of sample</th>
<th>Cumulative frequency (n)</th>
<th>% of sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>–5.95</td>
<td>1.52</td>
<td>101</td>
<td>12.5</td>
<td>101</td>
<td>12.5</td>
</tr>
<tr>
<td>1</td>
<td>–5.10</td>
<td>1.15</td>
<td>84</td>
<td>10.4</td>
<td>185</td>
<td>23.0</td>
</tr>
<tr>
<td>2</td>
<td>–4.07</td>
<td>0.91</td>
<td>102</td>
<td>12.7</td>
<td>287</td>
<td>35.6</td>
</tr>
<tr>
<td>3</td>
<td>–3.34</td>
<td>0.81</td>
<td>108</td>
<td>13.4</td>
<td>395</td>
<td>49.0</td>
</tr>
<tr>
<td>4</td>
<td>–2.72</td>
<td>0.75</td>
<td>113</td>
<td>14.0</td>
<td>508</td>
<td>63.0</td>
</tr>
<tr>
<td>5</td>
<td>–2.18</td>
<td>0.72</td>
<td>83</td>
<td>10.3</td>
<td>591</td>
<td>73.3</td>
</tr>
<tr>
<td>6</td>
<td>–1.69</td>
<td>0.69</td>
<td>78</td>
<td>9.7</td>
<td>669</td>
<td>83.0</td>
</tr>
<tr>
<td>7</td>
<td>–1.23</td>
<td>0.67</td>
<td>51</td>
<td>6.3</td>
<td>720</td>
<td>89.3</td>
</tr>
<tr>
<td>8</td>
<td>–0.79</td>
<td>0.66</td>
<td>34</td>
<td>4.2</td>
<td>754</td>
<td>93.5</td>
</tr>
<tr>
<td>9</td>
<td>–0.36</td>
<td>0.65</td>
<td>22</td>
<td>2.7</td>
<td>776</td>
<td>96.3</td>
</tr>
<tr>
<td>10</td>
<td>0.05</td>
<td>0.64</td>
<td>16</td>
<td>2.0</td>
<td>792</td>
<td>98.3</td>
</tr>
<tr>
<td>11</td>
<td>0.47</td>
<td>0.65</td>
<td>7</td>
<td>0.9</td>
<td>799</td>
<td>99.1</td>
</tr>
<tr>
<td>12</td>
<td>0.89</td>
<td>0.65</td>
<td>2</td>
<td>0.2</td>
<td>801</td>
<td>99.4</td>
</tr>
<tr>
<td>13</td>
<td>1.33</td>
<td>0.67</td>
<td>3</td>
<td>0.4</td>
<td>804</td>
<td>99.8</td>
</tr>
<tr>
<td>14</td>
<td>1.78</td>
<td>0.68</td>
<td>0</td>
<td>0.0</td>
<td>804</td>
<td>99.8</td>
</tr>
<tr>
<td>15</td>
<td>2.26</td>
<td>0.70</td>
<td>1</td>
<td>0.1</td>
<td>805</td>
<td>99.9</td>
</tr>
<tr>
<td>16</td>
<td>2.77</td>
<td>0.72</td>
<td>0</td>
<td>0.0</td>
<td>805</td>
<td>99.9</td>
</tr>
<tr>
<td>17</td>
<td>3.32</td>
<td>0.76</td>
<td>1</td>
<td>0.1</td>
<td>806</td>
<td>100.0</td>
</tr>
<tr>
<td>18</td>
<td>3.95</td>
<td>0.85</td>
<td>0</td>
<td>0.0</td>
<td>806</td>
<td>100.0</td>
</tr>
<tr>
<td>19</td>
<td>4.85</td>
<td>1.09</td>
<td>0</td>
<td>0.0</td>
<td>806</td>
<td>100.0</td>
</tr>
<tr>
<td>20</td>
<td>5.64</td>
<td>1.47</td>
<td>0</td>
<td>0.0</td>
<td>806</td>
<td>100.0</td>
</tr>
</tbody>
</table>
were relatively more equal across groups when using the empirically scored YAAPST. To test this possibility, we ran regression analyses predicting both original 27-item and revised YAAPST scores including a main effect of one alcohol use variable (frequency, quantity, or frequency of five or more drinks) and one background variable (sex, race, or school) and a term carrying the interactions between these two main effects. When predicting original YAAPST scores, seven of nine Background Interaction effects between these two main effects. When predicting original YAAPST scores, seven of nine Background × Alcohol Use Variable interactions were significant at \(p < .05\). By contrast, only two of nine interactions (college sample × drinking frequency and college sample × average quantity) were significant when using the empirically derived 20-item scale. In all cases, interaction effects were smaller in magnitude when using the 20-item YAAPST, and no interaction accounted for more than 1% of the variance in the 20-item YAAPST scores.

**Discussion**

Results of item response analyses of the 27 YAAPST items indicated that a reliable scale can be formed with 20 dichotomized items that retains the vast majority of the meaningful variance in the scale and that has superior distributional properties (i.e., reduced skewness and kurtosis). These items appear to measure a single dimension of alcohol problem severity among college students and reflect a broad range of severity of problems. The observed patterns of symptom expression were consistent with a Rasch model, with endorsement of more severe items being associated with a greater likelihood of endorsing less severe items. As we detail next, results are informative about the measurement, relative severity, and ordering of alcohol problems in college populations and ultimately may inform alcohol interventions with college students.

**Scale Refinement**

As a first step in our analysis plan, we used exploratory techniques to examine whether the use of multiple response options in the YAAPST contributed to the scale’s ability to differentiate levels of alcohol problem severity. We saw no evidence to suggest that the use of frequency scales added to the items’ ability to make these discriminations. This finding makes intuitive sense. Unlike attitudes, alcohol-related problems are discrete events. It is not clear how the frequency of an event will translate in terms of overall problem severity. For example, how many times would one expect an individual to get sick after drinking before that individual feels he or she needs a drink after wakening? As dichotomies, alcohol problems are much more clearly interpretable. We know that someone who needs a drink after wakening is likely to have had a number of problems in the past. By contrast, knowing that an individual has been sick after drinking tells us that the individual is not at the lowest level of alcohol problems but does not give much information about whether he or she has more severe problems. In the absence of evidence of frequency assessments increasing a scale’s discrimination ability, researchers may be better served by relying on a graded set of dichotomies ranging from low-severity problems to high-severity problems. Analyses of items from the Alcohol Dependence Scale (Skinner & Allen, 1982) in treatment-seeking alcoholics and high-risk drinkers have led to similar conclusions (Kahler, Strong, Hayaki, et al., 2003; Kahler, Strong, Stuart, et al., 2003).

Measurement bias has been suggested to affect estimations of alcohol-associated consequences in college men and women, inflating problems experienced by men or underestimating those experienced by women (Lo, 1996; Perkins, 1992). Our data support this notion. There was evidence of DIF across demographic subsamples for five YAAPST items. Although it may be reasonable to retain these items on the scale given that they may be of clinical relevance, such items should not be used when calculating a total problem severity score to ensure that the measurement of alcohol problems is equivalent across subgroups (i.e., free of bias). For example, regression analyses suggested that differential associations across subgroups between alcohol consumption and alcohol problems (i.e., YAAPST scores) were reduced when the revised scale rather than the original scale was used as a dependent variable.

Items relating to physical fights, damaging property, and being arrested for drunken behavior were endorsed at lower levels of alcohol problem severity among men than they were among women. This is not a surprising result as men would be expected to engage in more externalizing behavior than women even in the absence of drinking. In regard to ethnicity, one item—missing work or class because of drinking—showed significant and meaningful DIF. Endorsement of this item was associated with relatively more severe alcohol problems among non-White students compared to White students. Finally, receiving a lower grade on an exam or paper because of drinking was a relatively more severe item in Sample B compared to Sample A. These two school-related items may reflect differences in norms regarding the importance of school performance across ethnic groups and universities; that is, within an ethnic group or university in which schoolwork is given higher priority, only those students with severe alcohol problems may allow drinking to affect their school performance. Such interpretations are, of course, speculative. Nonetheless, these results show the importance of considering whether items function equivalently across subpopulations.

**Problem Severity and Ordering**

Rasch model analyses of the dichotomized YAAPST items indicated that these items measure a single dimension of alcohol problems. However, two items appeared to be redundant with other items on the scale (i.e., locally dependent) and were dropped from further analysis. Infit and outfit statistics suggested that the data fit a Rasch model well, providing evidence for a progression of symptoms that follows a developmental pathway in which less severe alcohol problems reliably precede more severe problems. The severity estimates for each item (see Table 1) indicate where along the alcohol problem continuum certain types of problems are likely to occur and which problems are relatively similar in severity. The probabilistic ordering of these problems is consistent with the findings of Langenbucher and Chung (1995), which indicated that certain symptoms of alcohol abuse and dependence are likely to occur earlier in the development of dependence than others.

Knowing an individual’s responses on the YAAPST, one can make reasonable predictions about the next most severe symptoms that are likely to emerge if drinking problems progress. For example, on the basis of item severity estimates, an individual who
has recently begun to neglect social obligations because of drinking may begin to feel dependent on alcohol if alcohol problems worsen. Needing a drink on waking, and seeking help, would be less likely next steps in the progression of problems because these represent far more severe symptoms. Of course, longitudinal data ultimately are necessary to determine the extent to which symptoms unfold in this expected developmental pattern.

**Score Distribution and Expected Symptoms**

After elimination of the five items with significant DIF, as well as those two that showed local dependence, raw YAAPST scores could range from 0 to 20 based on the remaining dichotomized items. In this sample of alcohol-drinking college students, 87.5% indicated that they had experienced at least one alcohol problem in the past year. This high proportion suggests that the YAAPST does contain some items that tap relatively common consequences of alcohol use. However, the majority of items target a level of alcohol problems that are considerably more severe than those typically experienced among college students. Only 3.7% of the sample scored 10 or higher on the scale, and only 0.2% scored higher than 14. This indicates that the YAAPST contains more severe items than are needed to make useful discriminations among college students. Collapsing or dropping items about help seeking, being arrested, and being fired/suspended may improve the efficiency of the instrument with general college populations.

About 50% of the sample received scores between 1 and 4, suggesting that scores in this range reflect relatively normative levels of drinking-related problems in college students. Using the norm table and the item-severity estimates, we can determine which symptoms are likely to be expressed in this range of scores. At a score of 1, all items on the scale have less than a 50% chance of being endorsed. However, individuals with a score of 4 are likely to have experienced a number of significant signs of excessive drinking, including hangover, vomiting, and memory loss. They have a slightly less than 50% chance of displaying tolerance to alcohol, and significantly more severe problems remain unlikely. Thus, this range of scores involves primarily excessive drinking with few significant consequences experienced. Individuals with scores in the 1-to-4 range may be considered reasonable targets for lower cost, widely disseminable universal preventive interventions such as social marketing campaigns (e.g., Haines & Spear, 1996).

One third of individuals fell in the range of scores from 5 to 9. As can be seen in the norm table (Table 2), there is a rapid increase in severity from a score of 5 (~2.18) to a score of 9 (~0.36). At the low end of this range of scores, certain risky behaviors (e.g., sexual situations that are later regretted, drinking and driving) are likely, and at the high end, school, work, and social obligations also are likely to be affected negatively. Thus, this region of the continuum could be considered to reflect the emergence of problem drinking. Students scoring in this range have experienced some significant consequences and may be at risk for developing alcohol dependence if their problems continue to progress. These individuals may be particularly good candidates for brief, tailored interventions to reduce harm caused by drinking (e.g., Baer, Kivlahan, Blume, McKnight, & Marlatt, 2001; Borsari & Carey, 2000; Collins, Carey, & Sliwinski, 2002; Marlatt et al., 1998). Data from the YAAPST can be used in feedback with these students to indicate the relative severity of symptoms they have endorsed and the types of symptoms that typically occur among individuals with more severe problems.

Scores of 10 and higher occurred in only 3.7% of this sample. At a score of 10, feeling dependent on alcohol has a 50% likelihood, and the next most severe problem is having the shakes. Thus, scores of 10 and higher reflect probable dependent drinking. Individuals with scores higher than 11 are likely to have serious problems with friends, work, and school and may seek help. These individuals are likely to require relatively intensive and specialized interventions to treat their significant alcohol problems and likely alcohol dependence.

**Limitations**

We conducted Rasch model analyses using a relatively large sample of college students from two universities. However, these universities were both located in the Northeast and had limited representation of many ethnic groups. Future research on measures of college student drinking should sample from universities in other geographic regions and with better representation of minority students. The high number of White students in our sample required us to examine DIF using a very crude split of White versus non-White.

We also were limited by the measure of alcohol problems that we used. Although the YAAPST is a relatively well known instrument with adequate psychometric properties, it provides only a subsample of a large number of items that could be used to assess alcohol problems in college students. An item pool that draws from multiple measures of alcohol problems may provide more complete and diversified coverage of the alcohol problem continuum. On the basis of our results, combining items from different instruments is feasible given that most problems can be usefully dichotomized as either present or absent. Additional construct validation analyses are needed in future studies, including examination of alcohol abuse and dependence diagnoses across levels of YAAPST scores. Given the high prevalence of alcohol abuse (31.6%) and dependence (6.3%) suggested by a recent national study of college drinkers (Knight et al., 2002), this information would be extremely valuable as an assessment and referral tool (e.g., in university counseling centers).

**Conclusions**

This analysis represents a first step in attempting to refine the measurement of alcohol problems in college and increase understanding of the relative severity and ordering of these problems. Studies that track the development of alcohol problems longitudinally are a next logical step and will help to delineate further the progression of alcohol consequences over time. Continued application of Rasch model analyses as well as other item response models will result in refined, unbiased, and streamlined measures that allow for the most efficient and accurate assessment of alcohol consequences. For example, having a bank of items with known severities can allow one to create computer-adaptive testing programs that present items of differing severity based on the participant’s responses, maximizing testing efficiency. Also, knowledge of the severity of symptoms can allow investigators to select items that best target the level of problem severity within a population of
interest. Finally, the establishment of stable severity estimates for particular alcohol-related problems can allow different instruments to be linked so that cross-instrument comparisons can be made on a similar metric. For example, if neglecting social obligations suggests a particular level of problem severity, then questions targeting this symptom would suggest whether an individual is above or below this level of severity irrespective of the manner in which it is asked (e.g., interview, self-report, manner of phrasing). Thus, item response analyses can inform one not only about the specific instrument being studied but also about specific symptoms and their relation to the underlying construct being measured.

References


Received April 30, 2003
Revision received July 15, 2003
Accepted July 23, 2003

---

**E-Mail Notification of Your Latest Issue Online!**

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at http://watson.apa.org/notify and you will be notified by e-mail when issues of interest to you become available!