HHS Public Access

Author manuscript

J Appl Meas. Author manuscript; available in PMC 2018 May 07.

Published in final edited form as:

J Appl Meas. 2007 ; 8(4): 373–387.

Substance Use Disorder Symptoms: Evidence of Differential Item Functioning by Age

Kendon J. Conrad,
University of Illinois at Chicago

Michael L. Dennis,
Chestnut Health Systems

Nikolaus Bezruczko,
Measurement and Evaluation Consulting

Rodney R. Funk, and
Chestnut Health Systems

Barth B. Riley
University of Illinois at Chicago

Abstract

This study examined the applicability of substance abuse diagnostic criteria for adolescents, young adults, and adults using the Global Appraisal of Individual Need's Substance Problems Scale (SPS) from 7,408 clients. Rasch analysis was used to: 1) evaluate whether the SPS operationalized a single reliable dimension, and 2) examine the extent to which the severity of each symptom and the overall test functioned the same or differently by age. Rasch analysis indicated that the SPS was unidimensional with a person reliability of .84. Eight symptoms were significantly different between adolescents and adults. Young adult calibrations tended to fall between adolescents and adults. Differential test functioning was clinically negligible for adolescents but resulted in about 7% more adults being classified as high need. These findings have theoretical implications for screening and treatment of adolescents vs. adults. SPS can be used across age groups though age-specific calibrations enable greater precision of measurement.

There is national concern regarding the adequacy and quality of the U.S. alcohol and drug abuse treatment system, particularly for adolescents (McLellan and Meyers, 2004). In their review, McLellan and Meyers came to the unambiguous conclusion that, although substance abuse is prevalent in most schools, primary care practices, mental health clinics, and criminal justice agencies, there is insufficient training, organization, or reimbursement to screen, assess, and refer those with dependence or abuse disorders to appropriate services.

Accurate and valid assessment is essential for identifying who needs a referral to treatment and for evaluating the effectiveness of treatments (U.S. Preventive Services Task Force, 2004). If the screening process is inaccurate, and people who need the program are denied it,

Requests for reprints should be sent to Kendon Conrad, School of Public Health, University of Illinois at Chicago, 1603 West Taylor Street (MC 923), Chicago, IL 60612 USA kjconrad@uic.edu.

they may fail to improve as a result. Likewise, if persons are screened into the program who do not need it, they will be unlikely to improve since they are already near the desired criterion. From an evaluation research perspective, this lack of improvement due to inaccurate targeting will dilute the observed effectiveness of the program. Among screening alternatives, urine screening is increasingly popular, but it is primarily a measure of metabolite from recent use (not the severity of substance related psychopathology). It also raises civil liberty issues in public schools, is expensive, and has poor sensitivity, i.e., how many people with diagnoses it detects, and specificity, i.e., how many people with use but not abuse/dependence that it correctly rules out (Lennox et al., 2006, in press). Thus it makes sense to shift the focus to more direct measures of the need for treatment.

The American Psychiatric Association (APA, 2000) and the World Health Organization (WHO, 1999) define the need for substance abuse treatment in terms of the substance use disorders (SUD) of dependence or abuse. They both define the term “substance dependence” to indicate a pattern of chronic problems that has been present for over 12 months and is likely to persist if left untreated (e.g., increased tolerance, withdrawal, loss of control, inability to reduce use or abstain, replacing other healthier activities with substance use, and continued use despite persistent related medical or psychological problems). The terms “substance abuse” and “hazardous use” are used by these respective organizations to identify a less severe subgroup of people who do not have dependence but who receive treatment because they report one or more moderately severe symptoms, are at high risk for harming themselves/ others or are at risk of developing dependence. There have been concerns, however, that the SUD symptoms were based on adult models and that there was limited data (less than 100 cases and no break out) on how well they worked with young adults and adolescents (Clark et al., 1998; Kilpatrick et al., 2000; Deas et al., 2000; Coffey et al., 2002). Of specific concern was whether all of the symptoms were present and in similar order of severity across ages. Subsequent data (Chan et al., under review) has shown that the prevalence of any substance use disorder among those presenting to substance abuse treatment increases from 72% at age 12-15 to 87% age 40-65 and that dependence increases from 45% to 77% over the same age groups (cross-sectional prevalence across groups). This suggests the need to understand whether these changes are real or an artifact of how the measure works across age groups.

Rasch (1960) measurement models provide one of the most effective techniques for quantifying the severity of items and the extent to which item severity estimates vary among known client subgroups (Wright, Mead, and Draba, 1976; Wright, 1997; Scheuneman and Subhiyah, 1998). Having the item severity parameter facilitates the formal testing of differential item functioning (DIF) and differential test functioning (DTF) for subgroups of persons that are at the same level on the underlying SUD construct (which typically cannot be observed and is treated here as a latent construct). The fact that certain items in a measure exhibit DIF does not necessarily mean that the test as a whole is biased. Oftentimes, the item differences will go in different directions and balance each other out so that there is no DTF. Both DIF and DTF can be important. Regarding the importance of DIF, a single symptom/ item like withdrawal is often key to treatment planning recommendations (e.g., the need for detox services). On the other hand, regarding DTF, diagnosis and placement tend to focus on the overall pattern of answers, i.e., test score (Lange, Irwin, and Houran, 2000).

Objective

This paper examines whether the SUD symptoms of the SPS: a) fall along a latent dimension and b) whether there is DIF and DTF by age.

Methods

Data Source

The data are from 77 sites across the country which included 7,408 clients (5,366 adolescents, age 10-17, 749 young adults age 18-25, and 1,293 adults age 26-69). All studies measured abuse and dependence with the Substance Problems Scale (SPS) from the Global Appraisal of Individual Needs (GAIN) (Dennis, Titus, White, and Hodgkins, 2003). Table 1 presents more detail on these demographics and selected clinical characteristics since, given the large sample size, all differences by age group were statistically significant. The adolescent and young adult samples were particularly more likely than the adult samples to be male and Caucasian. All three samples typically reported very high (80%) rates of substance use disorders in the past year. Adolescents were less likely than adults to have internalizing disorders (e.g., depression, anxiety, trauma, suicide) and more likely to have externalizing/impulse disorders (e.g., ADHD, conduct disorders), to report crime/violence in the past year – with young adults in between. Adolescents and young adults were less likely to be entering residential treatment or to have been in treatment before and more likely to be currently involved in the juvenile/criminal justice system.

Measures

The 77 sites all measured severity of substance use disorders with the GAIN (Dennis et al., 2003) Substance Problem Scale (SPS). The SPS consists of 16 items that assess “recency” (past month, 2-12 months, 1+ years, never) of symptoms of substance related problems: Seven items are based on DSM-IV criteria for substance dependence (tolerance, withdrawal, loss of control, inability to quit, time consuming, reduced activity, continued use in spite of medical/mental problems), four items for substance abuse (role failure, hazardous use, continued use in spite of legal problems, continued use in spite of family/ social problems), two items for substance-induced disorders (health and psychological), and three items for lower severity symptoms commonly used in screeners (hiding use, people complaining about use, weekly use). The latter five items are not used in diagnosis, but help improve the ability of SPS to work as a reliable, unidimensional measure of severity and change.

The recency rating can be used to make symptom counts for the past month, year or lifetime for the 16 item Substance Problem Scale (SPS). Higher scores on all of the overall scales (and subscales) are theorized to represent greater severity of drug problems. The substance problem scale is a measure of both the breadth of the problem measured with 16 symptoms of substance use, abuse, dependence and induced disorders and the recency of these symptoms. Thus if two people had 8 and 4 symptoms in the same time period, the one with 8 would be more severe. If two people each had 8 symptoms, the first all in the past month and the second all more than a year ago – clearly the first is more severe. The raw scores are used to classify people into low/moderate/high severity. Based on DSM-IV criteria

(American Psychiatric Association, 1994), individual items are used to classify people based on a “presumptive” diagnosis of dependence (3 to 7 symptoms of dependence), abuse (1-4 of the abuse symptoms and no dependence), or other (including weekly use, hiding using, complaints about use, 0-2 symptoms of dependence, or substance induced problems—and not meeting dependence or abuse criteria). Generally substance abuse treatment is limited to those with abuse or dependence and residential treatment is further limited to those with higher severity dependence. The SPS has previously been found to have good internal consistency (Cronbach’s alpha of .8 or more), test-retest reliability (Rho = .7 or more) and good test-retest reliability in terms of diagnosis (kappa = .5 or more) (Dennis et al., 2003). Copies of the SPS and full GAIN instruments and descriptions are available at www.chestnut.org/li/gain.

Analysis Procedures

We employed the Rasch measurement model (Rasch, 1960) using WINSTEPS (Linacre, 2005) statistical software to examine differential item and test functioning by age on the SPS (see Conrad and Smith, 2004, for a review; also Wright, Mead, and Draba, 1976). Since all items use a common rating scale, the Rasch rating scale model (Andrich, 1978; Wright and Masters, 1982) was used for the analyses. The Rasch rating scale model estimates the probability that a respondent will choose a particular response category for an item as:

$$\ln \frac{P_{nij}}{P_{ni(j-1)}} = B_n - D_i - F_j,$$

where

- P_{nij} is the probability of respondent n scoring in category j of item i,
- $P_{ni(j-1)}$ is the probability of respondent n scoring in category j–1 of item i,
- B_n is the person measure of respondent n,
- D_i is the difficulty of item i, and
- F_j is the difficulty of category step j.

Rating scale categories are ordered steps on the measurement scale. Completing the j^{th} step can be thought of as choosing the j^{th} alternative over the j−1 category in the response to the item. In this analysis of the SPS, the full rating scale was examined. It consists of three steps: from 0 to 1, from 1 to 2, and from 2 to 3 using the rating scale as follows: 0 = never, 1 = 1+ years ago, 2 = 2-12 months ago, 3 = past month. According to the Rasch model, across all persons and all items, the probability of responding to each ascending step should increase as one goes up the measurement scale.

Rasch analysis places persons (B_n) and items (D_i) on the same measurement scale where the unit of measurement is the logit (logarithm of odds unit). Person reliability estimates the reproducibility of person measures or calibrations on the scale and is comparable to Cronbach’s alpha. Since Rasch places both persons and items on the same scale, reliability can be estimated for items as well as for persons.

Analysis of Persons, Items and Unidimensionality. Following the analytic methods described by Smith, Conrad, Chang, and Piazza (2002), we examined the reliability of persons and items. Then we performed a principal component analysis of item residuals to evaluate unidimensionality. Rasch fit statistics and step calibrations were used to examine expected rating scale functioning in order to make corrections according to Linacre's (1999) procedures. Since changes were needed to the rating scale (reducing it from 4 to 3 levels), the previous analysis of dimensionality was repeated to verify that the changes did not degrade the measure (it did not). Only the final results are reported here. Item fit and person fit were inferred from the consistency between the actual and expected responses and item difficulties. Rather than tailor models to fit the data, the Rasch model holds that the one parameter model fulfills the requirements of fundamental measurement (Wright, 1997) and examines the data, i.e., items and persons, for flaws or problems that are indicated by their failure to fit the model (Smith, 2002; Wright and Stone, 1979; Wright and Masters, 1982). The Rasch model provides two indicators of misfit: infit and outfit. These fit statistics have the form of χ^2 statistics divided by their degrees of freedom. The infit statistic is sensitive to unexpected behavior affecting responses to items near the person ability level and the outfit statistic is outlier sensitive. Mean square fit statistics are defined such that the model-specified uniform value of randomness is 1.0 while the standardized Z statistic, Zstd, provides a significance test where >2.0 is typically regarded as significant. We examined Zstd (>2.0) on person and item mean square statistics as indicators of possible misfitting items (Smith, Schumacher, Bush, 1998). Since the sample was large (over 1,000) Zstd would tend to be overly sensitive so we also used the mean square statistics using <.75 and >1.33 as criteria for misfit (see Linacre, 2002; Smith et al., 1998; and Wilson, 2005, for discussion and examples).

Analysis of Differential Item Functioning (DIF)—In Rasch analysis, DIF concerns whether an item has significantly different calibrations (translate as severities, difficulties or ease/rarity of endorsement) depending upon subgroup membership (Masters, 1988; Smith 1992). This happens when the severity of a given item or symptom varies for people in different subgroups (in this case adolescents, young adults, adults; male/female; racial/ethnic groups) even though they are at the same level on the overall latent construct. The DIF contrast (calculated in Winsteps) is the difference in item calibrations between two subgroups in terms of logit measurement units. Where the goal is to have one common measure (e.g., a school performance test), DIF can be thought of as "bias" and identifies items that should be dropped. However, were the goal is to measure a condition with related but heterogeneous presentations (e.g., a clinical disorder), DIF can be thought of as describing "real" key differences and/or the potential need for subgroup specific calibrations/norms.

Differential Test (DTF)—Differential Test Functioning occurs when the DIF among the items is unbalanced. In this case, if the DIF for adults were greater than the DIF for adolescents, the overall test would tend to be biased. For example, the adults might appear to have less severe substance use disorders than they really had because they would get the same calibrations, e.g., on legal problems, as adolescents. In reality, legal problems are indicative of much greater severity for adults, so adults should have a high calibration for

legal problems whereas adolescents should have low calibrations for the same item. If DIF is balanced among the items, e.g., an item higher in severity for adults is balanced by an item higher for adolescents, the test results are not affected, but the DIF information may still be important theoretically. If DIF is unbalanced, and we still want to measure and compare adolescents and adults on the same scale, then four steps can be taken.

First, obtain pooled calibrations and identify items with and without DIF using Winsteps. In this case, >.6 logit DIF contrast, was considered substantial DIF (explained below).

Second, anchor the item and rating scale steps for the non-DIF items using the pooled calibrations. The anchoring creates a stable yardstick that assures comparability of the two groups even though some unanchored items will be allowed to "float" or achieve their group-specific calibration.

Third, rerun the analyses separately for adolescents and adults. This provides the group-specific, "floating" calibrations for the unanchored items. The non-DIF items will have been anchored thereby creating a common scale for the adjusted results and the non-adjusted results. This common scale enables comparison of adult measures to adolescent measures. The result is adult-specific and adolescent-specific calibrations resulting in person measures that can be compared to each other because they are anchored on a ruler made from the pooled sample using the DIF-free items as the anchors. The result is both item and person measures that are comparable across groups.

Fourth, recalculate the number of persons above and below the clinical cutoff. These numbers should have changed because of the new group-specific recalibrations, i.e., the "truer" age-specific calculations with the bias due to DIF removed.

Effect size—Since *t*-tests are sensitive to sample size, and we had a very large sample, the interpretation of DIF needed to be completed by consideration of effect size. Standards for what is considered an important DIF effect size vary from about .4 to .6 logits (see Longford, Holland, and Thayer; 1993; Paek, 2002; Draba, 1977; Elder, McNamara, and Congdon, 2003; Scheunemann and Subhiyah, 1998; Wang, 2000). In this paper we used the criterion of .6 logit or larger since we believed that most would agree that this is a large DIF contrast. As further context, we note that Norman, Sloan and Wyrwich (2003) stated that half a standard deviation is increasingly becoming the most common standard of clinical effectiveness. In this case, the SPS item standard deviation was .6 logit for a $d = 1.0$, i.e., .6/.6, a large effect. Therefore, by all of the above standards, .6 logit indicates a large and important clinical effect.

Differential test functioning (DTF) impact—In a previous paper we developed and validated cut points on the SPS using Rasch measures such that −2.50 logits = very low cutpoint including many people with low severity; −.75 = low cutpoint; −0.10 = moderate cutpoint; +.70 logits = high cutpoint including only those with high severity (Riley, Conrad, Bezruczko, and Dennis, 2007). To assess clinical impact of the DIF adjustment on DTF, we used those four cutpoints to examine how percentages above the cutpoints would differ between unadjusted measures and measures adjusted for DIF.

Gender and Race Analyses

Since males composed the majority of adolescents and females the majority of adults, we examined the possibility that the age findings were confounded by gender. The underlying assumption of Rasch and other IRT models is that patterns of item response are not influenced by any factor extraneous to the latent variable, i.e., no control for potential confounding influences. Our strategy for examining potential confounding was to examine age DIF within gender and race groups to see if the DIF results for age remained. We also noted that the majority of adolescents were Caucasian while the majority of adults were African American, so we examined whether the age DIF findings were confounded by race/ethnicity using the same approach.

Random split-half cross-validation of DIF contrasts

Since we had a large sample, we chose to examine whether the findings would be robust to chance variation. All analyses were first conducted on a random selection of about half of the subjects. These analyses were subsequently redone on the remaining subjects, i.e., cross-validation sample. Person and item reliabilities were the same for both halves. As further cross-validation, we compared the 48 DIF contrasts among adolescents, young adults, and adults for the two halves. We used a Bonferoni correction for multiple tests, $p < 0.05/48 = p < 0.001$, whereby 3 SE's were significant. Two were different, about 4%. We concluded that the DIF results were largely cross-validated across the random split halves. Therefore, the results reported below are on the full sample of 7,426, with results for all three groups, but we focused on adolescents vs. adults for the DIF analysis. The actual numbers available for analysis after case deletions due to missing data included 7,408 clients (5,366 adolescents, 749 young adults, and 1,293 adults).

Results

Evaluation of SPS Rasch Severity Measure, Items, and Response Rating Scale

The Rasch measures explained 72% of the total variance whereas the first factor of residuals explained 1.9%. We concluded that the results of the principal components analysis of residuals supported unidimensionality. The two items with the highest loadings on the first factor of residuals were: #8 Caused you to have repeated problems with the law, and #9 Get in fights/trouble. The person reliability of the SPS was .84 with person separation of 2.28. The item reliability was 1.00 with separation of 32.10. Figure 1 displays the calibrations, i.e., severity levels, of the items. These were within reasonable theoretical expectations (theoretical issues to be discussed in detail in future work). The measure was reasonably well-targeted with person and item means within a half logit of each other. Also, the high and low SPS rating scale categories (not displayed in Figure 1) covered the range of person measures fairly well, i.e., from about 2 logits to −2 logits in Figure 1. There was a substantial floor effect, but this was expected given that this was a screening assessment. These zero scores were not included in the parameter estimates, i.e., not considered part of the target population.

Rating Scale Analysis—In the original response scale (0 = never, 1 = more than a year ago, 2 = 2 to 12 months ago, 3 = past month), step 1 never achieved the highest probability

of being chosen at any point along the measure. This lack of a highest probability area or threshold is problematic and suggested the need to collapse the 1 category into another (either never and more than a year ago, or more than a year ago and 2 to 12 months ago). We opted to collapse the never and more than a year ago categories in order to still be able to measure change over time. This resulted in the scale 0 = never/not in the past year, 1 = 2-12 months, and 2 = past month. Doing this meant that all three steps had an independent place along the *x*-axis where each was the highest probability event. Based on these results, the category probabilities were properly ordered with this revision so that we used this 012 rating scale in subsequent analyses.

Re-examine Dimensionality, Reliability and Separation after Rating Scale Revision—Using the 3-point rating scale, variance explained by the measurement dimension dropped slightly to 68% of the total variance, and 12% of the residual variance was explained by the first principal component of residuals. Also, the person reliability decreased from .84 to .83 with separation of 2.24. Item reliability was 1.00 with slightly lower separation of 28.48. We regarded this as evidence of reduction in potential confusion and of simplification of the response categories but without significant loss of reliability.

Determine Whether There Were Any Misfitting Items and Persons—As recommended by Wilson (2005), Rasch fit statistics (Table 2) indicated unacceptable infit and outfit for two items, #1 “Tried to hide when using AOD” and #8 “Caused you to have repeated problems with the law.” We also noted that #8 was also the principal item on the first principal component of residuals that we discussed previously. Regarding person fit, in the most misfitting persons, we looked for patterns that would indicate non-compliance or misunderstanding of the task and found none. Rather, there were some outliers that would naturally be expected, e.g., people with high SPS severity who unexpectedly did not endorse a low severity item or two such as “complaints about use.” As a result, no person data were dropped from analyses.

Determine Whether There Was Any Differential Item Functioning (DIF) and Assess Its Impact—In Table 3, we display the DIF results for three age groups: adolescents, $n = 5{,}366$; young adults, $n = 749$; and adults, $n = 1{,}293$. The items with DIF contrasts above .6 logit are bolded. We can see item #8, “Caused you to have repeated problems with the law” was the item with the largest DIF contrast for age at −1.57 logits contrast for adolescents vs. adults followed by #1, “Hide when using AOD” at −1.02. Both items were easier for adolescents to endorse. The existence of DIF was pronounced between adolescents and adults since there were eight large (>.6) DIF contrasts, four items being easier for each group.

As we noted earlier, item #8 was also one of the two misfitting items (Table 2), and was the lead item on the first factor of residuals, and was the item with the highest DIF contrast for gender at −1.09 where it was easier for males to endorse and for race at −1.2 where white was compared with all others, and it was easier for whites to endorse. Since #8 was the only DIF item for gender and race, we viewed the differential test functioning (DTF) as insubstantial both clinically and theoretically for gender and race. Therefore, since the two items with the highest DIF were both easiest to endorse by adolescents and since there was a

clear pattern of differences for adolescents vs. adults, these patterns were worthy of further examination for DTF.

Since the largest DIF was between adolescents and adults with only three large DIF contrasts involving young adults, i.e., one involving a contrast between adolescents and young adults on Hiding Use and two involving contrasts between young adults and adults on Trouble/fights and Caused Repeated Legal Problems, we focused on the adolescent vs. adult contrasts to examine the impact of DIF (Table 3). The position of young adult measures between adolescents and adults on most symptoms is important theoretically since it displays the progression of symptomatology by age. While this analysis focused on the DIF between adolescents and adults, it is also important clinically to note this progression.

A higher probability of endorsing an item is indicated by a lower position on the scale, i.e., an “easier” item. We found then that it was more likely or easier for adults to endorse most symptoms when compared to adolescents, i.e., nine out of sixteen. However, the contrasts representing adults’ most prevalent or “easier” symptoms were for the dependence symptoms, i.e., Time Consuming, Role Failure, Loss of Control, Substance Use Disorder Induced Mental Health Problems, Can’t Stop, Tolerance, Despite Physical and Mental Health Problems, Give Up Activities, Withdrawal. On the other hand, the symptoms that were most prevalent or “easier” for adolescents were the abuse symptoms, i.e., Complaints About Use, Trouble/fights, Hiding Use, Caused Repeated Legal Problems.

To see if these results were sensitive to the unequal sample sizes of the adolescent vs. adult groups, we did a sensitivity analysis using a random subsample of adolescents to equal the number of adults. With equal numbers of adults and adolescents, using the joint SE, a very sensitive criterion, only 2 of 32 item calibrations (16 adolescent and 16 adult) differed by slightly more than two SE from the calibrations obtained using the full sample. None of the 16 DIF contrasts differed by more than two SE. Therefore, we concluded that the unequal sample sizes did not affect the findings.

Age within gender and race—We found that the within-group age DIF results for males and females were very similar to the findings for the full sample. This ruled out gender as a confounder for age. Likewise, race was ruled out as a confounder since we found that within-group age DIF patterns were also very similar for African Americans, Caucasians and Hispanics.

Differential Test Functioning—As noted earlier, adolescents and adults displayed large (>.6 logit) or “practically significant” DIF on eight items. This finding led to an examination of the clinical impact, or differential test functioning, of the SPS to address the question: Would adjustment for DIF make a difference in screening results for adolescents and adults?

Figure 2 is a scatter plot of the 16 items as they were calibrated in the separate analyses for adolescents and for adults. The dots/item numbers on the diagonal are the eight anchored items, i.e., those items on which both adolescent and adult samples had similar calibrations, so the calibrations for the pooled (adolescent and adult) groups were used. This, along with anchored calibrations for the rating scale categories created the common ruler. Above the

diagonal, we see the four items that were more difficult for adults to endorse: #1, Hide when using AOD; #2, Parent complained; #8, Caused repeated legal problems; #9, Fights and trouble. Below the diagonal are the four items that were more difficult for adolescents to endorse; #4, Depressed, nervous; #11, Withdrawal, illness; #13, Unable to cut down AOD; #16 Use in spite of physical and mental health problems.

If we only look horizontally at the adolescent calibrations on the *x*-axis, we see that the DIF items are well-balanced on each side of the 0.00 logit line. Specifically, if we add the negative calibrations of adolescent items 1, 2, and 9 we get −1.2 + −.7 + −.6 = −2.5. This is balanced by adding the calibrations of the five items with positive calibrations, i.e., 4, 11, 13, 16, and 8 where the sum is .7 + .4 + .4 + .3 + .2 = 2.0. Since these items balance fairly well for adolescents, they cancel each other out so that we would expect a negligible clinical effect of DIF for adolescents.

On the other hand, for adults, looking vertically on the *y*-axis, the four items above zero, 1, 8, 9, and 11 sum to 3.4 and the four below zero sum to −.7. These do not cancel each other out, but instead indicate that, by taking DIF into account, the measure has become significantly more difficult for adults. Taking DIF into account will mean that the greater weight given to items 1, 8, 9, and 11 will result in generally higher measurements for adults calibrated separately than when adults were co-calibrated with adolescents.

Clinical Impact—To test this expectation, we examined the percentages of adults and adolescents at various cutoff points, i.e., −2.5, −.75, −0.10, +0.70 logits before and after adjusting for DIF (Table 4). Cutoff scores of 0, 1, and 2 logits would be in the areas of increasingly high need which would result in assignment to more intense treatment such as residential care at the highest level. Since the DIF adjustment made the items more difficult for adults, adults should score higher on the adjusted score so that more adults would be over the cutoff. This was the case where, above −0.10 logits, over 7% more adults would be assigned to high need treatments. For adolescents, however, the unadjusted vs. adjusted differences were trivial at around 0.1%. Therefore, the differential test functioning or clinical impact would have been reasonably substantial for adults. After DIF adjustment, at the high need levels, over 90 adults would have been switched into high need instead of a lower level of care.

As a *post hoc* analysis we also examined whether the age DIF were causing the misfit in items 8 and 9, the most misfitting items. If it were, then these items would not misfit when fit was calculated within each sample. We found that within adults and within adolescents, the misfit disappeared for item #9 and was less, but still substantial for item #8.

Discussion

We found that the items of the SPS were useful in assessing the construct of interest, Substance Problems. Two misfitting items (8-Legal and 9-Trouble) were also two of the principal symptoms involved in DIF between adolescents and adults. Based on DIF analysis, we concluded that the items were useful for both groups, but that control for DIF might be needed to improve the validity of screening results for adolescents and adults. We concluded

that DIF concerns were not substantial for young adults. Young adults tended to be between adolescents and adults on item calibrations with only three substantial DIF contrasts. The contrasts involving young adults differed in both directions, i.e., one with adolescents and two with adults, rather than systematically contrasting with one group only. These findings were indicative of progressive change across the three age categories.

Since the item calibrations differed most between adolescents and adults, the screening results were examined for potential clinical impact. For adolescents this was negligible since only a few adolescents were affected, but for adults it was more substantial with about 90 adults moving from below to above the high need threshold. While this was only about 1% of the full sample and 7% of the adults, the significance of misclassification for the individuals involved can be very high.

Limitations

While the study has several strengths (e.g., diverse large sample, advanced measurement model and analytic methods), it is important to acknowledge some key limitations. The sample was not nationally representative (in particular having more adolescents and minority adults in residential care than typical of the U.S. public treatment system). All of the data are from self-report whereas ideally we would also have collateral information and/or the final clinical diagnoses to compare the results to. The differential item functioning analysis identifies differences in practice regardless of cause. Thus differences on repeated problems with the law (8-legal) could be due to developmental age, but could also be due to our culture's intolerance of repeated legal infractions (e.g., three strikes and you're out laws). The latter is likely to have systematically reduced the number of adults with this symptom and hence made it appear to be "rarer" than if we had included people in all settings (e.g., households, jail, other facilities). The differential test functioning is valid, but the common ruler may be slightly sensitive to the composition of the three age groups and ideally should be replicated in another sample.

Conclusion

We concluded that the SPS is a useful, unidimensional measure, although one item, "Caused you to have repeated problems with the law" was found to be especially poorly fitting. This item was also involved in some minor DIF by gender and race groups. We compared item calibrations for various subgroups and found that the major issues concerned eight items that functioned differently for adolescents vs. adults. From a theoretical perspective, this was evidence of the symptoms that account for the differences in self-reported severity. In other words, the DIF results made sense theoretically and clinically, the ultimate test of their validity. For adolescents, it was much easier to endorse symptoms of abuse: hide when using AOD; parent complained; caused legal problems; fights and trouble. For adults, it was much easier to endorse symptoms of dependence: depressed, nervous; withdrawal, illness; unable to cut down AOD; and use in spite of physical and mental health problems. Young adults tended to be between the younger and older groups which indicated a progression of change of symptom severity. The obvious implication is that it is necessary to take these differences

into account in order to screen, triage, and treat adolescents, young adults, and adults most appropriately.

To address the screening issue, Rasch analysis enabled us to control for the observed differential item functioning while maintaining the separate group measures on a common ruler so that the clinical impact of this control could be examined. More specifically, items that were similar in severity for the two groups were anchored as was the rating scale that was used. This placed both groups on the same ruler. Before anchoring on the common ruler, it was only possible to use the pooled calibrations so that adult endorsements would receive the same calibrations as adolescent endorsements. This made the adults “look like adolescents” in terms of severity, i.e., their measures were artificially low. However, the anchoring to establish a common ruler for both groups allowed the estimate of unique severities for adolescents, e.g., “legal problems” would be common/easy/low severity. Likewise, unique estimates for adults could be obtained, e.g., “legal problems” would be more rare, more difficult to endorse, and higher in severity. Using the common ruler, between-group comparisons were more valid.

In an examination of impact on clinical screening over a range of possible cutoffs, this difference would have changed the triage decision for over 1% of all persons in the sample and 7% of the adults in the higher need range. This was evidence that improved measurement using the linear, interval scale provided by Rasch analysis may have useful implications for improved substance abuse screening, triage, and treatment.

Acknowledgments

The authors are grateful to Dr. Ya-Fen Chan and Dr. Karen M. Conrad for their reviews. The development of this paper was supported by the Center for Substance Abuse Treatment (CSAT), Substance Abuse and Mental Health Services Administration (SAMHSA) via Westat under contract 270-2003-00006 to Dr. Dennis at Chestnut Health Systems in Bloomington, Illinois using data provided by the following grants and contracts from CSAT (TI11320, TI11324, TI11317, TI11321, TI11323, TI11874, TI11424, TI11894, TI11871, TI11433, TI11423, TI11432, TI11422, TI11892, TI11888, TI013313, TI013309, TI013344, TI013354, TI013356, TI013305, TI013340, TI130022, TI03345, TI012208, TI013323, TI14376, TI14261, TI14189,TI14252, TI14315, TI14283, TI14267, TI14188, TI14103, TI14272, TI14090, TI14271, TI14355, TI14196, TI14214, TI14254, TI14311, TI15678, TI15670, TI15486, TI15511, TI15433, TI15479, TI15682, TI15483, TI15674, TI15467, TI15686, TI15481, TI15461, TI15475, TI15413, TI15562, TI15514, TI15672, TI15478, TI15447, TI15545, TI15671, TI11320, TI12541, TI00567; Contract 207-98-7047, Contract 277-00-6500), the National Institute on Alcohol Abuse and Alcoholism (NIAAA) (R01 AA 10368), the National Institute on Drug Abuse (NIDA) (R37 DA11323; R01 DA 018183), the Illinois Criminal Justice Information Authority (95-DB-VX-0017), the Illinois Office of Alcoholism and Substance Abuse (PI 00567), the Interventions Foundation’s Drug Outcome Monitoring Study (DOMS), and the Robert Wood Johnson Foundation’s Reclaiming Futures (45054, 45059, 45060, 45053, 047266). The opinions are those of the author and do not reflect official positions of the contributing project directors or government.

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th. Washington, DC: Author; 2000.

Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978; 43:561–573.

Clark DB, Kirisci L, Targer RE. Adolescent versus adult onset and the development of substance use disorders in males. Drug and Alcohol Dependence. 1998; 49:115–121. [PubMed: 9543648]

Coffey C, Carlin JB, Degenhardt L, Lynskey M, Sanci1 L, Patton GC. Cannabis dependence in young adults: an Australian population study. Addiction. 2002; 97:187–194. [PubMed: 11860390]

Conrad KJ, Smith EV Jr. Applications of Rasch analysis in health care. Medical Care. 2004; 2004(Suppl I):42.

Deas D, Riggs P, Langenbucher J, Goldman M, Brown S. Adolescents are not adults: developmental considerations in alcohol users. Alcoholism: Clinical and Experimental Research. 2000; 24:232–237.

Dennis M, Godley SH, Diamond G, Tims FM, Babor T, Donaldson J, et al. The Cannabis Youth Treatment (CYT) study: Main findings from two randomized trials. Journal of Substance Abuse Treatment. in press.

Dennis ML, Scott CK, Funk R. An experimental evaluation of recovery management checkups (RMC) for people with chronic substance use disorders. Evaluation and Program Planning. 2003; 26:339–352.

Dennis, ML., Titus, JC., White, MK., Unsicker, J., Hodgkins, D. Global appraisal of individual needs: Administration guide for the GAIN and related measures. Bloomington, IL: Chestnut Health Systems; 2003. Retrieved DATEXXX, from http://www.chestnut.org/li/gain

Draba, RE. The identification and interpretation of item bias. Chicago: Statistical Laboratory, Department of Education, University of Chicago; 1977. Research Memorandum No. 26

Elder C, McNamara T, Congdon P. Rasch techniques for detecting bias in performance assessment: An example comparing the performance of native and non-native speakers on a test of academic English. Journal of Applied Measurement. 2003; 4:181–197. [PubMed: 12748409]

Kilpatrick DG, Acierno R, Saunders B, Resnick HS, Best CL, Schnurr PP. Risk factors for adolescent substance abuse and dependence. Journal of Consulting and Clinical Psychology. 2000; 68:19–31. [PubMed: 10710837]

Lange R, Irwin HJ, Houran J. Top-down purification of Tobacyk's revised paranormal belief scale. Personality and Individual Differences. 2000; 29:131–156.

Lennox RD, Dennis ML, Scott CK, Funk RR. Combining psychometric and biometric measures of substance use. Drug and Alcohol Dependence. 2006; 83:95–103. [PubMed: 16368199]

Linacre JM. Investigating rating scale category utility. Journal of Outcome Measurement. 1999; 3:103–122. [PubMed: 10204322]

Linacre JM. What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions. 2002; 16:878.

Linacre, JM. A user's guide to WIN-STEPS. Chicago: Mesa Press; 2005.

Longford, NT., Holland, PW., Thayer, DT. Stability of the MH D-DIF statistics across populations. In: Holland, PW., Wainer, H., editors. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum; 1993. p. 67-113.

Masters GN. Item discrimination: When more is worse. Journal of Educational Measurement. 1988; 24:15–29.

McLellan AT, Meyers K. Contemporary addiction treatment: A review of systems problems for adults and adolescents. Biol Psychiatry. 2004; 56:764–70. [PubMed: 15556121]

Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. Medical Care. 2003; 41:582–592. [PubMed: 12719681]

Paek, I. Unpublished doctoral dissertation. University of California; Berkeley: 2002. Investigations of differential Item functioning: Comparisons among approaches, and extension to a multidimensional context.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research; 1960. Expanded edition, 1980. Chicago: University of Chicago Press

Riley BB, Conrad KJ, Bezruczko N, Dennis M. Relative precision, efficiency and construct validity of different starting and stopping rules for a computerized adaptive test: The GAIN substance problem scale. Journal of Applied Measurement. 2007; 8(1)

Scheuneman JD, Subhiyah RG. Evidence for the validity of a Rasch technique for identifying differential item functioning. Journal of Outcome Measurement. 1998; 2:33–42. [PubMed: 9661730]

Smith EV Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied Measurement. 2002; 3:205–231. [PubMed: 12011501]

Smith EV Jr, Conrad KM, Chang K, Piazza J. An introduction to Rasch measurement for scale development and person assessment. Journal of Nursing Measurement. 2002; 10:189–206. [PubMed: 12885145]

Smith, RM. Applications of Rasch measurement. Chicago: MESA Press; 1992.

Smith RM, Schumacher RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. Journal of Outcome Measurement. 1998; 2:66–78. [PubMed: 9661732]

U.S. Preventive Services Task Force. Screening and behavioral counseling interventions in primary care to reduce alcohol misuse: recommendation statement. Annals of Internal Medicine. 2004 Apr 6.140(7):164.

Wang WC. Modeling effects of differential item functioning in polytomous items. Journal of Applied Measurement. 2000; 1:63–82. [PubMed: 12023558]

Wilson, M. Constructing measures: An item response modeling approach. Mahwah, NJ: Lawrence Erlbaum; 2005.

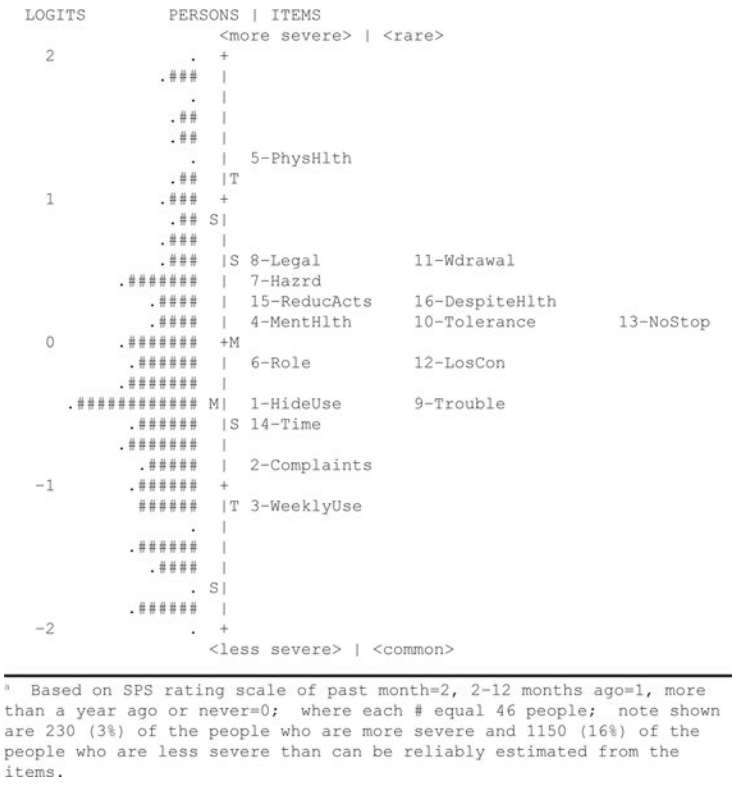
World Health Organization (WHO). The international statistical classification of diseases and related health problems, tenth revision (ICD-10). Geneva, Switzerland: World Health Organization; 1999. Retrieved from www.who.int/whosis/icd10/index.html

Wright BD. A history of social science measurement. Educational Measurement: Issues and Practice. 1997 Winter;:33–52.

Wright, BD., Masters, GN. Rating scale analysis. Chicago: MESA Press; 1982.

Wright, BD., Mead, R., Draba, R. Detecting and correcting test item bias with a logistic response model. MESA Psychometric Laboratory; 1976. MESA Research Memorandum Number 22

Wright, BD., Stone, MH. Best test design. Chicago: MESA Press; 1979.



[a] Based on SPS rating scale of past month=2, 2–12 months ago=1, more than a year ago or never=0; where each # equal 46 people; note shown are 230 (3%) of the people who are more severe and 1150 (16%) of the people who are less severe than can be reliably estimated from the items.

Figure 1.
Substance Problem Scale (SPS) Rasch Person/Item “Wright” Mapa

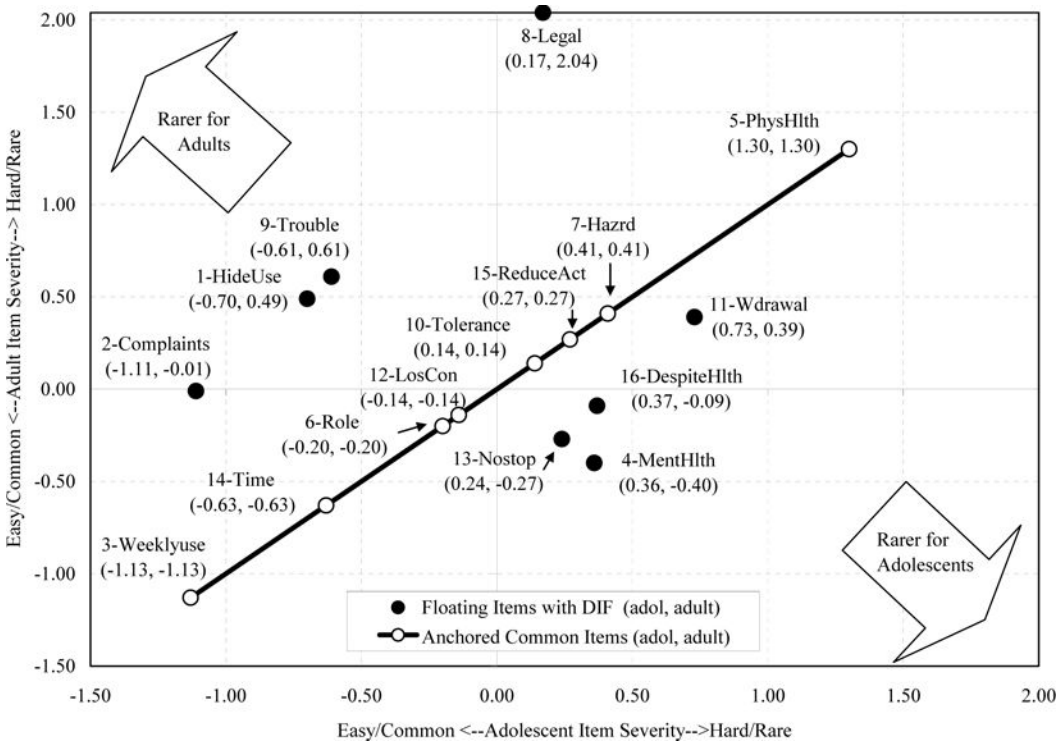


Figure 2.
Substance Problem Scale (SPS) DIF Between Adolescents and Adults

Table 1

Sample Characteristics

	Adolescents Aged <18 (*n* = 5,366)	Young Adult Aged 18-25 (*n* = 749)	Adults Aged 26+ (*n* = 1,293)	Total (*n* = 7408)
Male	73%	63%	48%	67%
Caucasian	48%	55%	30%	45%
African American	17%	25%	62%	26%
Hispanic	14%	6%	2%	11%
Mixed/Other	21%	14%	5%	18%
Average Age	15.6	20.1	37.6	19.9
Substance Use Disorder	86%	80%	90%	86%
Internalizing Disorder	55%	63%	69%	58%
Externalizing Disorder	65%	50%	38%	59%
Crime/Violence	65%	52%	35%	59%
Residential Tx	33%	51%	76%	42%
Current legal involvement	70%	71%	45%	66%

Note: All significant, $p < .001$

J Appl Meas. Author manuscript; available in PMC 2018 May 07.

Table 2

Substance Problem Scale Items and Rasch Statistics

		Rasch		INFIT[c]		OUTFIT[d]	
Item-Label	**Substance Problem Scale (SPS) questions\b. When was the last time…**	**Measure[a]**	**Rank[b]**	**MNSQ[e]**	**ZSTD[f]**	**MNSQ**	**ZSTD**
1–Hideuse	you tried to hide that you were using alcohol or drugs?	−0.45	13	1.42	9.9	1.78	9.9
2–Complaints	your parents, family, partner, co–workers, classmates or friends, complained about your alcohol or drug use?	−0.89	15	1.2	9.9	1.39	9.9
3–Weeklyuse	you used alcohol or drugs weekly?	−1.13	16	0.93	−4.3	1.04	1.5
4–MentHlth	your alcohol or drug use caused you to feel depressed, nervous, suspicious, uninterested in things, reduced your sexual desire or caused other psychological problems?	0.19	7	0.95	−3.4	0.9	−4.3
5–PhysHlth	your alcohol or drug use caused you to have numbness, tingling, shakes, blackouts, hepatitis, TB, sexually transmitted disease or any other health problems?	1.29	1	1.26	9.9	1.13	3.3
6–Role	you kept using alcohol or drugs even though you knew it was keeping you from meeting your responsibilities at work, school, or home?	−0.19	11	0.78	−9.9	0.75	−9.9
7–Hazrd	you used alcohol or drugs where it made the situation unsafe or dangerous for you, such as when you were driving a car, using a machine, or where you might have been forced into sex or hurt?	0.37	4	1.06	3.7	1.09	3.3
8–Legal	your alcohol or drug use caused you to have repeated problems with the law?	0.55	3	1.53	9.9	1.75	9.9
9–Trouble	you kept using alcohol or drugs even though it was causing social problems, leading to fights, or getting you into trouble with other people?	−0.39	12	1.08	5.1	1.16	7.3
10–Tolerance	you needed more alcohol or drugs to get the same high or found that the same amount did not get you as high as it used to?	0.14	8	0.84	−9.9	0.77	−9.8
11–Wdrawal	you had withdrawal problems from alcohol or drugs like shaking hands, throwing up, having trouble sitting still or sleeping, or that you used any alcohol or drugs to stop being sick or avoid withdrawal problems?	0.63	2	1.1	5.8	0.95	−1.9
12–LosCon	you used alcohol or drugs in larger amounts, more often or for a longer time than you meant to?	−0.14	10	0.77	−9.9	0.75	−9.9
13–Nostop	you were unable to cut down or stop using alcohol or drugs?	0.12	9	0.93	−4.8	0.88	−4.7
14–Time	you spent a lot of time either getting alcohol or drugs, using alcohol or drugs, or feeling the effects of alcohol or drugs (high, sick)?	−0.63	14	0.78	−9.9	0.75	−9.9
15–ReducActs	your use of alcohol or drugs caused you to give up, reduce or have problems at important activities at work, school, home or social events?	0.27	5	0.8	−9.9	0.79	−9.9
16–DespiteHlth	you kept using alcohol or drugs even after you knew it was causing or adding to medical, psychological or emotional problems you were having?	0.26	6	0.86	−9.3	0.89	−9.9

[a]Measure is the Rasch estimated logits of a given item from the mean of items on the latent SPS dimension.

[b]Rank is the rank order (low to high) of items.

[c]Item INFIT is the overall agreement of the item's with the model prediction in this data where high values suggest more random than expected and low values suggest everyone answers the item the same (e.g., all the same).

[d] Item OUTFIT is a measure of how precisely the item severity measure predicts individuals' answers to the item, where high values means that item response curve is flatter than average and/or crosses over other item curves and low values mean that the item is closer to a step function that provides little information just below or above the cut point to help the overall estimate of a continuous dimension.

[e] MNSQ is the average squared deviation of the difference between individual responses and the predict response for the item.

[f] ZSTD is the MNSQ standardized into Z–score based on the mean/variance of MNSQ for the items in the table.

Note: SPS Items reproduced with permission of copyright holder (Chestnut Health Systems), see www.chestnut.org/li/gain for complete copy of the GAIN

J Appl Meas. Author manuscript; available in PMC 2018 May 07.

Table 3

Analysis of Differential Item Functioning (DIF) between Adolescents, Young Adults, and Adults

	Rasch Calibration of Item for:				Pair-wise Difference[a]			Pair-wise *T*-Tests[b]		
Item–label\a	**Total (*n* = 7408)**	**Adolescent (*n* = 5,366)**	**Yount Adult (*n* = 749)**	**Adult Adult (*n* = 1,293)**	**Adol – Young Adult**	**Adol – Adult**	**Young Adult – Adult**	**Adol vs. Young Adult**	**Adol vs. Adult**	**Young Adult vs Adult**
1–HideUse	−0.45	−0.71	−0.08	0.31	**−0.63**	**−1.02**	−0.39	**−6.62**	**−14.60**	**−3.57**
2–Complaints	−0.89	−1.08	−0.65	−0.22	−0.42	**−0.86**	−0.43	**−4.45**	**−12.00**	**−3.91**
3–WeeklyUse	−1.13	−1.15	−1.11	−1.16	−0.04	0.01	0.05	−0.40	0.15	0.43
4–MentHlth	0.19	0.40	−0.10	−0.55	0.50	**0.95**	0.45	**5.23**	**12.70**	**3.99**
5–PhysHlth	1.29	1.30	1.09	1.60	0.21	−0.30	−0.51	1.89	**−3.61**	**−4.05**
6–Role	−0.19	−0.10	−0.13	−0.50	0.03	0.40	0.38	0.31	**5.53**	**3.37**
7–Hazrd	0.37	0.33	−0.01	0.54	0.34	−0.21	−0.55	**3.50**	**−3.00**	**−4.99**
8–Legal	0.55	0.16	0.74	1.72	−0.58	**−1.57**	**−0.99**	**−5.65**	**−19.80**	**−8.14**
9–Trouble	−0.39	−0.58	−0.43	0.34	−0.14	**−0.92**	**−0.78**	−1.52	**−13.20**	**−7.09**
10–Tolerance	0.14	0.24	0.20	−0.12	0.04	0.36	0.32	0.44	**4.99**	**2.82**
11–Wdrawal	0.63	0.78	0.39	0.17	0.40	**0.61**	0.21	**3.99**	**8.39**	1.89
12–LosCon	−0.14	−0.01	0.00	−0.55	−0.01	0.55	0.55	−0.06	**7.43**	**4.92**
13–NoStop	0.12	0.29	−0.02	−0.42	0.31	**0.71**	0.40	**3.23**	**9.60**	**3.53**
14–Time	−0.63	−0.58	−0.54	−0.85	−0.04	0.27	0.31	−0.46	**3.51**	**2.73**
15–ReducActs	0.27	0.33	0.41	−0.01	−0.09	0.34	0.43	−0.89	**4.74**	**3.79**
16–DespiteHlth	0.26	0.42	0.18	−0.18	0.24	**0.60**	0.36	**2.52**	**8.30**	**3.20**

Note: See Table 2 for full questions.

[a] Difference in Rasch Calibration of Item (first group minus second group) with those differences of .6 or more **bolded**

[b] *t*–test of respective pair wise difference (difference divided by pooled standard error), with those *t*–tests with $p < .05$ **bolded**.

Table 4

Percentage above clinical cutoffs on adjusted and unadjusted measures

	Percentage Adolescent* (*n* = 5,366)			Percentage Adult** (*n* = 1,293)		
Clinical Cutoffs	**Unadjusted Measure**	**Adjusted for DIF**	**% Change**	**Unadjusted Measure**	**Adjusted for DIF**	**% Change**
−2.5 = very low	86.2%	86.2%	0.0	91.6%	91.6%	0.0
−0.75 = low	50.5%	50.5%	0.0	76.8%	76.9%	0.1
−0.10 = moderate	27.3%	27.2%	−0.1	59.2%	66.5%	7.3
+.70 = high	11.1%	11.1%	0.0	34.0%	42.0%	8.0

*Adolescent unadjusted mean (logits) = −.8975, adjusted for DIF mean = −.9014 ($t = -4.7$, $p < .001$)

**Adult unadjusted mean (logits) = .0217, adjusted for DIF mean = .2086 ($t = 31.4$, $p < .001$)

Note: Due to the large numbers of cases, even small differences were statistically significant. Therefore, statistical criteria were not regarded as meaningful.