Published in final edited form as:

Soc Work Public Health. 2015 ; 30(3): 260–271. doi:10.1080/19371918.2014.994725.

Precision Across Race, Age And Gender Of A HIV Risk Screen For Adolescents And Young Adults

Michiel A. van Zyl, Ph.D.,

Associate Dean of Research, University of Louisville, Kent School of Social Work, 109 Oppenheimer Hall, Louisville, KY 40292, Tel: 502-852-2430. Areas of Research: HIV/AIDS; Teenage Pregnancy Prevention; Healthy Relationships; Scale Development

Christina Studts, Ph.D., and

Department of Health Behavioral, University of Kentucky College of Public Health, 335 Bowman Hall, 151 Washington Ave., Lexington, KY 40506-0059. Areas of Research: Externalizing Behavior in Young Children; Item Response Theory

Kathryn Pahl, PhD.

Managing Director, Shout-It-Now, 380 Main Road, Tokai, Cape Town, 7945. Areas of Research: HIV/AIDS; Program Effectiveness; Management of NGO'

Michiel A. van Zyl: mavanz01@louisville.edu; Christina Studts: tina.studts@uky.edu; Kathryn Pahl: Kathy@shout-it-now.org

Introduction

The identification of adolescents with high risk of contracting HIV offers the potential of providing preventative interventions targeted at the high risk HIV-negative adolescent population. A 12-item measure to identify adolescents at high risk for HIV infection was administered in 2011 by a non-governmental organization (NGO) as part of innovative voluntary testing and counseling (VCT) intervention to a group of adolescents (N = 3 872) in South Africa. NGO staff subsequently shortened the measure from 12 items to 7 items. This study retrospectively investigates the validity and reliability of the risk screening measures and proposes a new, briefer 5-item measure.

Promising results of a pilot study (Van Zyl, Barney & Pahl, In Press) of an innovative HIV prevention program targeted at adolescents, Shout-It-Now, led to wide scale implementation of the program in South Africa. Computers and internet access were made available to schools and community settings by the Shout-It-Now program and its sponsors, allowing adolescents to individually access online program content related to HIV prevention. Adolescents participated by viewing an online video of South African celebrities talking about issues related to HIV/AIDS prevention, including (a) condom usage, (b) Voluntary Counseling and Testing (VCT), (c) safer sex, and (d) responsible decision-making. Celebrities representing a variety of South African race groups provided information in English, interspersed with local vernacular. The styles of the videos were similar to MTV television programs, using popular music and attention-retaining visual material. During the

video a number of ‘pop-up’ questions appeared to reinforce messages being conveyed in the video program. The video took about 12 minutes to view, in line with anticipated attention spans and efforts to reduce participant burden.

Following the video presentation, adolescents were invited to participate in VCT. No coercion was placed on those who declined. Participants who had questions or concerns regarding the HIV test were invited to speak with a trained counselor. All who agreed to be tested for HIV were given a confidential one-on-one counseling session with a trained and accredited VCT counselor, followed by an HIV screening test. In cases of positive or undefined results, a confirmatory test was conducted. Testing was conducted in accordance with UN guidelines and regulations of the South African government. On the same day as testing, participants were given their results within the framework of a confidential post-test counseling session. During this session, risk reduction strategies were discussed. Participants who tested positive for HIV were referred to appropriate treatment centers, care and support services, and were invited to access the program’s 24-hour Helpline. Following the post-test counseling session, participants were given compensation such as music and cell phone airtime.

In implementing the Shout-It-Now intervention, program staff perceived the need to identify those at higher levels of risk and to offer additional services to them. To meet this need, a risk questionnaire was developed by program staff. Questions focused on three types of risk behaviors: risky sexual behaviors (condom use, number of sexual partners and being forced to have intercourse), alcohol and drug use, and absenteeism from school. These risk behaviors are related: Du Rant et al. (1999) described strong correlations among adolescent health risk behaviors, specifically early onset smoking and use of other substances (alcohol and drugs) with absenteeism and poor academic performance. These associations were evident across socio-demographic groups. Similarly, Guttmacher et al. (2002) identified correlations between school absenteeism and other adolescent risk taking behaviors. In South Africa, education is compulsory from the age of seven (grade 1) to age 15, or the completion of grade 9. Also, the surveillance system of the Center for Disease Control for sexual behaviors that contribute to unintended pregnancy and sexually transmitted diseases, including HIV infection, monitor two of the three types of risk as indicators of high risk behavior: (i) drug and alcohol use before sexual intercourse, and (ii) risky sexual behaviors regarding condom use and number of sexual partners (CDC, 2010). Although there is agreement in the literature about drug, alcohol and condom use as risk factors, no validated risk measure with known psychometric characteristics is currently available for use in the diverse South African population.

An initial set of 20 questions derived from the three types of risk behaviors was compiled by the trained NGO HIV counsellors and reviewed in focus group discussions with adolescents aged 14 to 18 years old. Twelve of the twenty items emerged with consistently shared meaning in discussions and included items measuring all three types of risk behaviors. As indicated in Table 1, the 12 items had response options on semantic differential scales of varying lengths (i.e., several questions had 4 response options, others had 3 response options, etc.). The 12 items were included as a screening measure in the online program and administered to 3,872 adolescents and young adults in two cities (Johannesburg and Cape

Town) and one rural area (Burgersfort, Limpopo) in South Africa. The NGO program staff subsequently decided to use only 7 of the 12 items that they regarded, based on face validity, as the most important to determine risk. Different weights were allocated to response options according to the clinical team's perceptions of relative risk associated with each response option. The summed score on the questionnaire was used to identify those with high risk for contracting HIV. No formal investigations of the psychometric properties of the 12-item or 7-item measures were conducted.

Study Aims

The purposes of this study were two-fold: first, to determine the reliability and construct validity of the 7-item HIV Risk measure used by the NGO; and second, to determine if a brief risk measure with better psychometric properties could be developed from the initial 12-item questionnaire. The research was conducted to inform the NGO about the reliability and validity of the measure being used to determine which adolescents were at high risk for HIV infection.

Limitations of the original risk measure

Development of the original risk measure had two primary limitations. First, determination of items to be included in the measure was guided primarily by the face validity of questions. Face validity alone is insufficient to justify wide scale implementation of an instrument measuring a high stakes construct such as risk for HIV infection.

Second, the 7-item measure developed by the clinical team used a weighted total score to identify high risk adolescents. There are several problems associated with assigning weights to different response options. Cognitive bias in this scoring approach may be a problem. For example, one may believe or even have evidence that a certain drug (e.g., methamphetamine) is more detrimental to health than another substance (e.g., alcohol), and therefore regard users of methamphetamine as having higher risk for HIV infection compared to alcohol users. However, there is a cognitive bias in this perception, related to generalizing one type of health risk to another. Giving a higher weight on a risk scale to a question that asks about the use of drugs as opposed to alcohol may be intuitively appropriate, but not empirically sound. Another problem with the weighted scoring approach used for the 7-item risk questionnaire stems from possible range compression, related to the clinical team's subjective assessments of the amount of risk associated with each question's response options. Range compression occurs when response options are limited to a small number of outcomes or possibilities, when in fact a much wider range of options are possible or likely. Consequently, there is a significant loss of precision in the assessment.

To address these limitations, a formal psychometric assessment of the risk measure was conducted. Traditional psychometric analyses were complemented with item response theory (IRT) analyses to (1) obtain detailed item- and test-level information about the performance of the risk measures across race, age and gender groups, and (2) investigate the possibility of developing a briefer, psychometrically sound risk measure from the original pool of 12 items.

Method

This study was conducted with existing data provided by the NGO. The data were from a sample of 3,872 adolescents in two cities (Johannesburg and Cape Town) and one rural area (Burgersfort, Limpopo) in South Africa. Adolescents aged 12 to 19 and young adults aged 20 to 25 were eligible to participate, based upon the legal age for consenting to an HIV test (12 years) in South Africa. Participants were recruited in two ways. First, students in grades 8 to 12 from six public secondary schools in lower and lower-middle socio-economic communities were invited to participate in the intervention. These schools were identified in collaboration with the department of education and voluntarily participated in Shout-It-Now's program during a four-month time period in 2011. The NGO reported a 93% participation rate of all students present on the day the Shout-It-Now program was delivered. Second, respondents were also recruited by program outreach efforts in the community, as the Shout-it-Now program includes active on-site marketing and recruitment of participants. Shout-it-Now staff were deployed to work in non-school settings on days outside the regular school calendar. On these days, they worked in shopping malls, at sports events, and at other venues where it was expected that they would encounter high numbers of adolescents and young adults. Staff approached adolescents and young adults at each venue, described the program, and invited them to participate. This strategy added to the diversity of the sample, ideal for validation studies where the aim is not to describe the characteristics of a specific population, but rather to determine the psychometric qualities of an instrument. Fingerprint identification (by means of the integrated confidential biometric identifier - DigitalPersona™) was used to ensure that each participant was screened only once. The refusal rate for fingerprint scanning was less < .005%. Approval and oversight to conduct research using these de-identified data was obtained from the University of Louisville's Institutional Review Board.

The original 7-item risk measure was scored using a summative model, adding the weighted scores for each item response to obtain a total score. This scoring approach assumes that the measure is unidimensional. The unidimensionality assumption was tested using exploratory factor analysis (EFA) on the 7-item scale. Reliability and validity of the 7-item scale were also assessed, including each item's mean corrected item-total correlation and the measure's overall internal consistency reliability (i.e., Cronbach's alpha).

Next, the full set of 12 items was analyzed to determine whether a brief risk measure with good psychometric properties could be developed. First, the factor structure of the 12 questions was determined using EFA. An iterative process of determining factor structure and interpreting factors was employed to identify a unidimensional construct, which was then assessed for reliability and validity.

This process facilitated the identification of a 5-item risk measure. Item response theory (IRT) analyses were employed to determine the psychometric properties of the newly developed brief risk measure. Item response theory provides more detailed psychometric information than classical test theory and is suitable to address the hypotheses generated in response to the second research question. First, unidimensionality of the 5-item measure was investigated using EFA. Next, the IRT assumption of local independence (i.e., the

requirement that items should be statistically independent from one another after controlling for the level of risk (Steinberg & Thissen, 1996; Wainer & Thissen, 1996; Yen, 1993)) was assessed by inspecting the absolute values of residual correlations for each pair of items. A criterion of $|r| \leq .20$ (Reeve et al., 2007) was used to determine violation of local independence. Following the testing of IRT assumptions, maximum-likelihood estimation procedures were used to fit a two-parameter logistic model to the data: MULTILOG 7.03 (Thissen et al., 2003) software was used to fit Samejima's (1968) Graded Response Model and obtain item parameter estimates and estimates of participants' levels of HIV risk. The amount and precision of measurement information provided by the newly developed 5-point scale were assessed using test information, item information, and item parameter estimates.

Once these psychometric properties were established, differential item functioning (DIF) analysis was conducted to determine if each item and the scale performed consistently across groups categorized by race, gender, and age. The R package *lordif* (Choi et al, 2011) was used in these analyses. The *lordif* package relies upon ordinal logistic regression for uniform and nonuniform DIF detection, employing Monte Carlo procedures (algorithms relying on repeated random sampling) to identify thresholds indicating whether items exhibit DIF, minimizing Type I error. The use of Monte Carlo procedures were preferable in detecting DIF given the differences in sample size of the various cohorts in the comparison. In this approach, the impact of DIF on IRT parameter estimates is assessed by comparing model fit between nested ordinal logistic regression models with and without group terms (i.e., for race, gender, and age). Main effects of group are included to test for uniform DIF, while interactions between group and risk level are included to test for non-uniform DIF. Significant DIF is identified using likelihood-ratio (LR) tests between models. In the ordinal logistic regression approach to DIF-detection, an iterative approach is used in which group-specific parameter and trait estimates are updated and re-estimated until consistent identification of items with DIF over subsequent iterations is achieved (Choi et al., 2011). Monte Carlo simulations were used to determine if empirical threshold values systematically deviated from the nominal level. Two additional DIF analyses were also employed, assessing the magnitude of (a) changes in pseudo- R^2 values, and (b) differences in parameter estimates between groups of interest.

Results

The majority of the 3 872 participants were female (54.6%). Race categories similar to those used in official census surveys were used to describe the racial composition of the sample. The largest racial group was Black (87.5%), followed by Coloured (10.8%), White (1%), Indian (0.3%), and Other (0.3%). The mean age was 17.1 years ($SD = 4.3$), and most participants were in grade 10 (25.6%), with grade 8 (22.5%) and grade 11 (20.2%) also well represented in the sample. 15.7% were in grade 12 and no grade was reported for .7% of the sample. Nearly all (97.0%) of participants were seen in a school setting, with the remainder (3.0%) seen in a shopping mall. Most participants (86.4%) were from the wider Cape Town area, with 10.7% from Johannesburg and 3.0% from Burgersfort in the Limpopo Province. The sample was from lower and middle socio-economic areas and was equally split between those whose families did versus did not own a car.

The 7-Item Measure

An EFA of the 7-item risk measure (See Table 1) yielded one factor that explained 40.6% of the variance. The 7 items loaded on one factor with an Eigenvalue of 2.84. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was .83, indicating an adequate sample size for the analysis. Bartlett's test of sphericity was highly significant ($p < .001$). The factor loadings were between .33 and .78. Internal consistency reliability as measured by Cronbach's alpha was .74. Only three items had relatively high correlations with the total scale score ($> .50$). The corrected mean item-total correlation was .46 ($SD = .14$). Applying a 90th percentile cut-off score to determine high risk using the total weighted score approach, 496 (12.8%) were identified as falling into the high risk category.

The Original 12-Item Measure

The original set of 12 items (See Table 1), when subjected to a principle component analysis and varimax rotation, yielded three factors that explained 52% of the variance. The first factor consisted of 6 items, with loadings between .44 and .73. The second factor had 4 items with loadings between .54 and .67, and the third factor had only 2 items with loadings of .78 and .84. The third factor focused on drug use ("How often do you smoke dagga?" and "How often do you use any other drugs (e.g. tik, crack, cocaine, sniff glue, etc)?"). The first two factors included 2 cross-loading items, and distinguishing between these factors was difficult, as each had items related to having sex, alcohol use, and missing school. Given the difficulty in interpreting the first two factors, it was decided to follow a different approach in deriving a unidimensional measure from the 12 items.

A 5-Item Measure

In reviewing the 12 items for content, 6 items were identified as focusing on the conditions associated with having sex. No central theme could be determined for the remaining 6 items. The 6 items associated with conditions when having sex were subjected to factor analysis. In a principal component analysis of these 6 items, one factor was extracted that explained 48% of the variance. The internal consistency reliability (Cronbach's alpha) of the 6-item instrument was .77, but one of the items ("Have you ever been forced to have sex when you did not want to?") correlated poorly with the total scale score (.30). After this item was removed, Cronbach's alpha of the 5-item risk measure increased to .79, all items correlated at .50 or higher with the total scale score, and the mean corrected item-total correlation was .57 (See Table 1). The 5-item measure in comparison to the 12-item and 7-item measures, offers an unidimensional factor structure that explained the most variance, as well as the highest coefficients of validity and reliability, and the only measure with psychometric properties within a preferred range. Although some of the differences in coefficients of the measures may not seem to be substantial, item diagnostics in scale development are often used incrementally to arrive at the best possible item combinations, particularly in the case of screening instruments that require as few items as possible. Also, if all diagnostics point in the same direction such as in this case, without for example, having to justify why a slight increase in validity is traded for a decrease in reliability, it is difficult to explain why items should not be deleted as indicated by the psychometric characteristics.

IRT Analysis of the 5-Item Measure

Unidimensionality and local independence of the 5-item risk measure were supported by EFA results: A single factor was extracted with an eigenvalue of 2.71 that accounted for 54% of the variance, and absolute values of residual correlations for each pair of items ranged from .00 to .08. Fitting the Graded Response Model yielded four parameter estimates for each item: a (discrimination), b_1 (difficulty threshold between option 1 and option 2), b_2 (difficulty threshold between option 2 and option 3), and b_3 (difficulty threshold between option 3 and option 4). High values of a suggest that items are able to distinguish between participants at similar levels of risk. Item parameter estimates and standard errors for each item are presented in the first column of Table 2.

Three of the five items (items 1, 3, and 4) demonstrated high discrimination ($a = 1.43$ to 1.53) (Baker, 1985), and one item (item 2) very high discrimination ($a = 2.56$). Only item 5 had moderate discrimination ($a = .77$). The lowest difficulty parameter, b_1 , clustered around one-third of a standard deviation above the mean risk level ($M = .28$). The lowest b_1 parameter estimate was for item 1 ($b_1 = -.21$, $se = .05$), while item 5 exhibited the highest b_1 parameter estimate ($b_1 = 1.12$, $se = .07$). This range of values suggests that risk levels near the mean were associated with selecting option 1 rather than option 0 on all 5 items.

Notably, the b_3 difficulty parameter estimates were extremely high ($M = 11.07$ standard deviations above the mean). A total of 563 (14.5%) respondents endorsed option 3 on at least one item. Item 4 ($b_3 = 15.77$) was the most difficult of the set, requiring extremely high levels of risk for participants to select the highest response option on this item. However, the two items with the lowest difficulty levels for their upper thresholds were item 2 ($b_3 = 8.18$) and item 3 ($b_3 = 8.88$), also requiring very high levels of risk for endorsement of option 3 versus 2. The extreme standardized scores have large standard errors, and the values should be interpreted relative to the mean of 0 and other large parameter estimates, and not in terms of absolute values.

Information for measuring risk with the total number of 5 items was higher than the standard error (i.e., most precise) from approximately 1.0 standard deviation below the mean to about 2.6 standard deviations above the mean. The test information curve peaks from 0.2 to 1.6 standard deviations above the mean, a range appropriate for precise measurement in a screening instrument, as screening should accurately assess those with low levels of risk (such as at one standard deviation below the mean) and those with high risk (such as those with more than 1 standard deviations above the mean).

As some items offered more information than others at similar levels of risk, some items could be omitted. Other aspects of item performance need to be considered before such a decision is made, including the degree to which an item exhibits measurement bias, or DIF.

DIF Analysis of the 5-Item Measure

Racial differences—Differences in item parameter estimates by racial groups were investigated using Black versus Other, in which Other included all groups other than Black (i.e., Coloured, White, Indian, and Other). With *lordif*'s default settings, the program terminated in two iterations. All five items were identified as potentially having DIF.

Sparseness due to all items being flagged limited the range of diagnostics possible for detailed DIF analysis. However, it was apparent that the mean slope of the true score functions for all 5 items was substantially lower for Blacks than for Others (1.55 vs. 2.15), indicating non-uniform DIF. The LR χ^2 test for uniform DIF, comparing Model 1 and Model 2, was significant for items 1, 2, 4, and 5 ($p < .001$). This was also true for the 2-*df* test of non-uniform DIF (comparing Models 1 and 3, $p < .001$) for the same items. The overall 1-*df* test was significant for items 1 and 5. The non-uniform component of DIF revealed by the LR χ^2 test can also be observed in the substantial group differences of the slope parameter estimates for item 1 (2.29 vs. 1.50), item 2 (3.90 vs. 2.51), and item 5 (1.07 vs. .76). When weighted by the focal group trait distribution, the expected impact of DIF as reflected by McFadden's pseudo- R^2 measures varied across items from .02 to .08 with a mean of .05 for R^2_{I3} . The impact apparent in the R^2_{23} was less and varied from .00 to .03 with a mean of .01. The percent change in β_I for all 5 items tended to be small ($M = 4\%$), with a maximum difference of 11% noted for item 2, which is higher than the frequently used criterion of 10% change in β_I to conclude that DIF exists (Crane et al., 2004).

The mean Monte Carlo probability threshold values associated with the χ^2 statistic across items were .008, .01 and .01 for χ^2_{12} (testing for uniform DIF), χ^2_{13} (testing for non-uniform DIF), and χ^2_{23} (testing for DIF overall while controlling Type I error), respectively. On average, the empirical threshold values for the probability associated with the χ^2 statistic were close to the nominal α level. The Monte Carlo simulation results confirmed that the LR χ^2 test maintains the Type I error adequately in this dataset.

Gender and age—Similar analyses were conducted for two other comparison groups, categorized by gender and age. For these analyses, age was categorized as younger (< 19 years) or older (≥ 20 years). The program terminated for gender in two iterations, flagging all items. Similarly, the program terminated for age in five iterations, flagging all items.

The mean slope for the true score functions was lower for males in comparison to females (1.54 vs. 1.64) and substantially lower for younger compared to older subjects (1.51 vs. 1.82), indicating non-uniform DIF. The LR χ^2 test for uniform DIF, comparing Model 1 and Model 2 (χ^2_{12}), was significant for all items for gender and items 1, 2, 4, and 5 for age. The 2-*df* test (χ^2_{13}) for non-uniform DIF (comparing Models 1 and 3, $p < .001$) was significant for all items for both gender and age. The overall 1-*df* test (χ^2_{23}) was significant for items 1, 3, 4 and 5 in the case of gender, and items 1, 2 and 5 for age. The non-uniform component of DIF revealed by the LR χ^2 test can also be observed in the difference of the slope parameter estimates; for example, in the case of age for the younger vs. older groups, respectively, for item 2 (2.44 vs. 3.16), item 3 (1.44 vs. 1.97), and item 5 (1.51 vs. 1.82). The McFadden's pseudo- R^2 measures varied across items from .0025 to .0231, with a mean of .01 for R^2_{I3} and .0037 to .0195 with a mean of .013 for R^2_{I3} respectively for gender and age. The impact apparent in the R^2_{23} was also small, with a mean of .003 for gender and age. When aggregated over all the items in the test, differences in item characteristic curves may become small due to canceling of differences in opposite directions. However, this does not mean that the impact on trait estimates is not of concern.

The theta values (i.e., levels of risk) with the highest information for the total sample and the various groups are reported in Table 3. Maximum item information estimates reflected by theta values of .2, 1.6 were most common in all groups for items 1, 2 3 and 4. Exceptions included lower values (−1.4, .0) for item 1 in the case of younger participants, and for item 2 for the total sample, as well as for older participants in the case of item 3. Information for item 3 for the racial “other” group was highest in the 1.8, 3.0 range.

Total non-weighted mean and percentile scores (90, 93, 95 and 99) on the 5-item risk scale for eight different groups (age by race by gender) are presented in Table 4. Scores varied widely between the groups. For example, the mean score across all items of the 5-item measure for older female subjects of Other race is 5.10, but for younger females in this race group, the mean score is only 0.41. The 90th percentile scores for Other race ranges from 1.00 for younger females to 11.00 for older males and females. For older Blacks, the impact of gender on scale total scores appears small, but for younger Blacks, gender has a significant impact ($M = 3.12$ for males vs. 2.08 for females; 90th percentile = 8.00 vs. 5.00 , respectively).

Using the age-, race- and gender-specific 90th percentile as a criterion to determine high risk with the 5-item measure, 499 (12.9%) adolescents were identified as falling into the high risk group. This proportion is equivalent to the 496 (12.8%) identified as high risk using the summative weighted scoring approach for the 7-item measure. However, only 313 (8.1%) cases were identified by both scales as high risk, and the inter-rater agreement between the two scales was significant, but not high (Cohen’s kappa = 0.57 ; $p < 0.001$). Although the percentage of high risk groups identified out of the total sample by the 5-item and 7-item measures is almost similar, the composition of the groups differs substantially. The need for accurate assessment and precision in measurement is highlighted by this finding.

Discussion

The 7-item risk measure used at present is unidimensional with acceptable reliability for group administration, but its internal consistency is inadequate for differentiating between individuals (.65 is required at the group comparison level and .80 at the individual level; Nunnally, 1994). In addition, the validity of the 7-item scale as measured by corrected mean item-total correlation (.45) was relatively low; a corrected mean item-total correlation of at least .50 is desired (Hudson, 1982).

In contrast, the 5-item risk measure was unidimensional, valid (mean corrected item-total correlation .50), and reliable enough for use in differentiating individual levels of risk (Cronbach’s alpha = .79). Further, the 5-item scale was 2 items shorter and more valid (.57 vs. .46) and reliable (.79 vs. .74) than the 7-item instrument. A reduced number of items selected from the 12-item HIV Risk Questionnaire was therefore incorporated into a new instrument with improved psychometric characteristics in comparison to the 7-item measure currently in use.

All 5 items of the newly developed measure discriminated sufficiently between subjects at different levels of risk. For all 5 items, extremely high levels of risk were necessary for a

subject to select the highest option on the four-point scale. The extremely high level of risk required for the highest scores to be endorsed is an indication of range compression. An additional response option between “A few times” and “Many times”—for example, “Several times”—may add to the measure’s precision. The range of precision of the measure is appropriate for a screening instrument.

All items in the 5-point risk measure were flagged for measurement bias or DIF. Both uniform and non-uniform DIF were identified as problems. However, the percentage of subjects with salient score changes and the minimal clinically important difference (MCID) for the risk measure have not yet been determined. Age-, race- and gender-specific norms are therefore recommended. The mean scores and percentiles between these groups vary substantially (see Table 4). Of note, the actual number of older subjects of Other race in the sample is very low. Consequently, percentile scores are not an appropriate way to identify high risk individuals in this group, because a few subjects with high score skew the distribution. Lower thresholds for older adolescents of Other race, such as those presented for the older Black adolescents, are recommended.

Limitations

Several study limitations should be considered. This study focused on items used in practice to determine risk and were supported by previous studies (mainly in North America) and the CDC’s surveillance system. Information on items used in the South African context was not available. Also, the sample size of one cohort for those 20 years and older was small and suggested risk-cutting scores for this group could not be developed. The most important limitation is the lack of longitudinal data inclusive of HIV status to use as a criterion in determining the predictive validity of the measures analyzed. More information on the history of participant HIV+ status and reported risk behavior at the time of contracting HIV, will add to the value of studies aimed at determining predictive validity.

Conclusions

These findings, combined with the fact that a less reliable 7-item measure is currently used to measure risk for HIV infection, have immediate implications for practice. A 90th percentile cutoff score for high risk resulted in similar numbers (respectively 499 and 496 for the 5-item and 7-item scales) being identified as high risk, but the agreement between the scales in categorizing adolescents’ risk levels was only moderate. In addition, the extremely high level of risk required for the highest score option on the four-point scale to be endorsed (as reflected by the mean standardized score of 11.07) is alarming. For most domains of behavior, the range of standardized scores is between −3 and 3. In total, 563 participants selected the highest response option on at least one item. This means that in a sample primarily comprised of high school students of low and middle socioeconomic status in South Africa, 14.5% indicated that they engaged in behavior much more risky than the behavior of their peers. Although the measurement error on any single item is higher than the measurement error for the total scale, it is interesting that the percentage of high risk individuals using the 90th percentile cut-off (12.9%) is close to the percentage of participants who endorsed the highest response option on at least one item (14.5%).

VCT is not done routinely in schools in South Africa. In light of the relatively high percentage of adolescents that engage in risky behavior for HIV infection, this policy should be revisited. The newly developed 5-item measure offers a way to identify adolescents at high risk, a first step in providing targeted preventive interventions to the high risk adolescent population. Use of the 5-item screener, including age-, race-, and gender-specific percentile-based norms, may dramatically improve prevention effectiveness, and will enable additional validation studies and further refinement of this and other risk measures. In addition, the behavior dimension included in the 5-item measure is not different from other approaches used to determine risk for HIV infection. Consequently, it is possible that other measures may also include item bias for various groups. Item bias should therefore be investigated for instruments or questions used to determine HIV risk.

References

- Baker, FB. The Basics of Item Response Theory. Portsmouth, NH: Heinemann Educational Books; 1985.
- Centers for Disease Control and Prevention. Surveillance Summaries. MMWR. 2010; 59(SS-5)
- Crane PK, Gibbons LE, Jolley L, Van Belle G. Differential Item functioning analysis with ordinal Logistic Regression Techniques: DIF Detect and *difwithpar*. Medical Care. 2006; 44(11 Supp 3):S115–S123. [PubMed: 17060818]
- Crane PK, Van Belle G, Larson EB. Test bias in a cognitive Test: Differential Item Functioning in the CASI. Statistics in Medicine. 2004; 23:241–256. [PubMed: 14716726]
- Choi SW, Gibbons LE, Crane PK. *lordif*: An R Package for Detecting differential Item functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. Journal of Statistical Software. 2011; 39 (8):1–28. [PubMed: 21572908]
- DeVellis, RF. Scale Development - Theory and Applications. 3. Los Angeles: Sage; 2012.
- Du Randt RH, Smith JA, Kreiter SR, Krowchuk DP. The Relationship Between Early Age of Onset of Initial Substance Use and Engaging in Multiple Health Risk Behaviors Among Young Adolescents. Pediatric Adolescent Medicine. 1999; 153:286–291.
- Guttmacher S, Weitsman B, Kapadia F, Weinberg S. Classroom-based Surveys of Adolescent Risk Taking Behaviors: Reducing the Bias of Absenteeism. American Journal of Public Health. 2002; 92(2):235–237. [PubMed: 11818298]
- Nunnally, JC.; Bernstein, IH. Psychometric Theory. 3. New York: McGraw-Hill; 1994.
- Hudson, WW. The Clinical Measurement Package: A Field Manual. Homewood, IL: Dorsey Press; 1982.
- Reckase MD. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics. 1979; 4:207–230.
- Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of Health-Related Quality of Life item banks plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care. 2007; 45(5 Suppl 1):S22–S31. [PubMed: 17443115]
- Samejima F. Calibration of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement. 1969; 17
- Steinberg L, Thissen D. Uses of item response theory and the testlet concept in the measurement of psychopathology. Psychological Methods. 1996; 1:81–97.
- Thissen, D.; Chen, WH.; Bock, RD. MULTILOG 7.03 [computer software]. Lincolnwood, IL: Scientific Software International; 2003.
- Van Zyl MA, Barney R, Pahl K. VCT and Celebrity Based HIV/AIDS Prevention Education: A Pilot Program Implemented in Cape Town Secondary Schools. Journal of HIV/AIDS & Social Services. In Press.

- Wainer H, Thissen D. How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*. 1996; 15:22–29.
- Yen W. Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*. 1993; 30:187–213.

Table 1

Risk Assessment Questions and corresponding 7-item scale and 5-item scale items

Original 12-item questionnaire			7-Item Scale Items Cronbach alpha = .74	5-item Scale Items* Cronbach alpha = .79
#	Item	Response options and Weights		
1	If you have a boyfriend or a girlfriend, please tell us:	0 I don't have a boyfriend or girlfriend 1 She or he is not more than 5 years older than me 3 She or he is more than 5 years older than me		
2	How often do you have three or more drinks at one time?	0 Never 1 Less than monthly 2 Monthly 3 Weekly 4 Daily or almost daily	Item 1	
3	Have you ever been forced to have sex when you did not want to?	0 No, never. 2 Yes, but only once or twice 4 Yes, it happens often		
4	How many days did you bunk school in the last year?	0 Never 1 Once or twice 2 Between 3 and 5 days 3 Between 5 and 10 days 4 More than 10 days	Item 3	
5	How do your parents feel about you bunking school?	0 They don't allow it at all 1 They allow it occasionally 3 They allow it whenever I want to bunk 4 They don't care if I bunk school		
6	How many times in the last year did you have sex without a condom?	0 Never 2 Once 3 A few times 4 Many times	Item 4	Item 1
7	How many times in the last year did you have sex while drunk?	0 Never 2 Once 3 A few times 4 Many times	Item 2	Item 2
8	How many times in the last year did you have sex while high on drugs?	0 Never 2 Once 3 A few times 4 Many times		Item 3
9	How many times in the last year did you have sex for gifts or favours?	0 Never 2 Once 3 A few times 4 Many times	Item 5	Item 4
10	How many people have you had sex with in the past 6 months?	0 None 2 One 3 Two or three 4 More than 3 people	Item 6	Item 5
11	How often do you smoke dagga?	0 never 2 occasionally 3 daily		
12	How often do you use any other drugs (e.g. tik, crack, cocaine, sniff glue, etc)	0 never 2 occasionally 3 daily	Item 7	

* Never = 0, Once = 1, A few times = 2, Many times = 3

Table 2
Graded Response Model Item Parameter Estimates for Total Sample and Subgroups

		Parameter Estimate (SE)						
Item	Parameter	Total Sample (N=3,872)	Male (N=1,757)	Female (N=2,115)	Black (N=3,387)	"Other" (N =487)	Age 19 (N=3,483)	Age 20 (N=388)
1	a	1.53 (-0.1)	1.53 (-0.14)	1.77 (0.14)	1.5 (0.09)	2.29 (0.62)	1.51 (0.1)	1.56 (0.21)
	b_1	-0.21 (-0.05)	-0.11 (-0.07)	-0.27 (0.07)	-0.19 (0.05)	-0.5 (0.23)	-0.12 (0.06)	-0.73 (0.19)
	b_2	1.32 (-0.08)	1.32 (0.14)	1.23 (0.1)	1.33 (0.09)	1.15 (0.28)	1.38 (0.1)	1.03 (0.23)
	b_3	10.08*	11.91	8.92	10.78	7.98	10.76	8.92
2	a	2.56	2.3	2.79	2.51	3.9	2.44	3.16
		-0.18	0.21	0.29	0.18	1.3	0.19	0.55
	b_1	0.13	0.07	0.19	0.13	-0.02	0.2	-0.25
		-0.04	0.06	0.07	0.05	0.16	0.05	0.11
	b_2	1.42	1.36	1.51	1.41	1.53	1.44	1.31
		-0.07	0.1	0.11	0.07	0.26	0.08	0.15
	b_3	8.18	8.05	9.45	8.34	9.85	8.48	8.99
	3	a	1.53	1.48	1.59	1.53	1.98	1.44
		-0.23	0.36	0.31	0.17	1.28	0.27	0.55
b_1		0.07	0.02	0.11	0.09	-0.43	0.13	-0.22
		-0.19	0.12	0.2	0.11	0.53	0.18	0.31
	b_2	1.5	1.49	1.48	1.49	1.76	1.59	1.14
		-0.28	0.18	0.26	0.16	0.89	0.4	0.29
	b_3	8.88	9.44	7.87	8.72	14.43	9.68	5.99
	4	a	1.43	1.46	1.43	1.44	1.5	1.4
		-0.14	0.17	0.23	0.14	1.06	0.15	0.36
b_1		0.31	0.18	0.5	0.31	0.31	0.34	0.14
		-0.07	0.09	0.13	0.08	0.47	0.08	0.17
	b_2	1.76	1.6	2	1.76	1.66	1.82	1.52
		-0.16	0.19	0.31	0.16	0.99	0.19	0.31
	b_3	15.77	14.81	15.44	15.53	20.02	11.89	19.16

Item	Parameter	Parameter Estimate (SE)						
		Total Sample (N=3,872)	Male (N=1,757)	Female (N=2,115)	Black (N=3,387)	“Other” (N =487)	Age 19 (N=3,483)	Age 20 (N=388)
	a	0.77	0.93	0.61	0.76	1.05	0.76	0.82
		-0.07	0.1	0.12	0.08	0.46	0.08	0.2
	b_1	1.12	0.24	2.56	1.11	1.27	1.18	0.78
		-0.17	0.11	0.59	0.16	0.53	0.18	0.31
	b_2	3.31	2.32	5.21	3.37	2.25	3.37	2.97
		-0.54	0.36	1.67	0.51	1.09	0.55	1.03
5	b_3	15.62	17.17	17.66	14.81	18.29	14.91	20.07

* Large SE given extreme parameter estimates.

Table 3
Maximum Item Information Estimates and Locations for Total Sample and Groups

Item	Total Sample	Maximum Information Estimate					Theta Values ^a with Highest Information
		Race: Black	Race: Other	Young	Older	Males	Females
1	0.65	0.627	1.174	0.634 *	0.653	0.657	0.846
2	1.965 *	1.646	3.798	1.551	2.458	1.398	1.988
3	0.663	0.666	0.986 **	0.59	1.036 *	0.618	0.712
4	0.581	0.59	0.644	0.558	0.703	0.607	0.579
5	0.175	0.17	0.322	0.171	0.196	0.248 ***	0.108

^aTheta values increase in steps of 0.2.

* -1.4,0.0

** 1.8,3.0

*** 0.2,1.6

Table 4

Mean and percentile scores for 5-Item Risk Scale by age, race and gender.

5-Item Risk Scale													
		N	Mean	Standard Deviation	Percentile 90	Percentile 93	Percentile 95	Percentile 99					
Age	19 years	Race	Black	Gender	Male	1472	3.12	3.28	8.00	9.00	10.00	13.00	
					Female	1745	2.08	2.54	5.00	6.00	7.00	11.00	
			Other	Gender	Male	204	.89	2.27	4.00	5.00	6.00	9.00	
					Female	264	.41	1.24	1.00	3.00	3.00	7.00	
20 years	Race	Black	Gender	Male	74	4.58	3.60	9.00	10.00	12.00	15.00		
				Female	96	3.48	3.59	9.00	10.00	11.00	15.00		
			Other *	Gender	Male	7	3.86	3.98	11.00	11.00	11.00	11.00	
				Female	10	5.10	4.23	11.00	14.00	14.00	14.00		
Total Sample						3872	2.4	3.0	7.00	8.00	9.00	12.00	

* Due to skewness in distribution and small cell n, percentile scores equivalent to the "Black Older Group" are recommended for use to differentiate high risk individuals.