Published in final edited form as:
AIDS Behav. 2010 February ; 14(1): 162–172. doi:10.1007/s10461-008-9423-z.

Scaling sexual behavior or "sexual risk propensity" among men at risk for HIV in Kisumu, Kenya

Mattson CL[1], **Campbell D**[1], **Karabatsos G**[2], **Agot K**[3], **Ndinya-Achola JO**[4], **Moses S**[5], and **Bailey RC**[1]

[1]School of Public Health, University of Illinois at Chicago, Chicago, IL, USA

[2]College of Education, University of Illinois at Chicago, Chicago, IL, USA

[3]UNIM Project, Kisumu, Kenya

[4]Department of Medical Microbiology, University of Nairobi, Kenya

[5]Department of Medical Microbiology, University of Manitoba, Canada

Abstract

We present a scale to measure sexual risk behavior or "sexual risk propensity" to evaluate risk compensation among men engaged in a randomized clinical trial of male circumcision. This statistical approach can be used to represent each respondent's level of sexual risk behavior as the sum of his responses on multiple dichotomous and rating scale (i.e. ordinal) items. This summary "score" can be used to summarize information on many sexual behaviors or to evaluate changes in sexual behavior with respect to an intervention. Our 18 item scale demonstrated very good reliability (Cronbach's alpha of 0.87) and produced a logical, unidimensional continuum to represent sexual risk behavior. We found no evidence of differential item function at different time points (except for reporting a concurrent partners when comparing 6 and 12 month follow-up visits) or with respect to the language with which the instrument was administered. Further, we established criterion validity by demonstrating a statistically significant association between the risk scale and the incident STI at the 6 month follow-up visit and HIV at the 12 month follow-up visits. This method has broad applicability to evaluate sexual risk behavior in the context of other HIV and sexually transmitted infection (STI) prevention interventions (e.g. microbicide or vaccine trials), or in response to treatment provision (e.g., anti-retroviral therapy).

Keywords

non-parametric item response theory; male circumcision; risk compensation; HIV/AIDS; Africa

INTRODUCTION

Recently, three randomized controlled trials (RCTs) have demonstrated the efficacy of male circumcision (MC) as a means of sharply reducing HIV incidence (Auvert, 2005; Bailey et al., 2007; Gray et al., 2007). These results show that the risk of HIV infection can be reduced by approximately 60% by the introduction of a medically appropriate program of voluntary circumcision. However, many public health professionals and others are concerned that men who have been circumcised will believe they are "inoculated" against

Correspondence and request for reprints: Christine L. Mattson Division of Epidemiology and Biostatistics School of Public Health University of Illinois Chicago 1603 W. Taylor St. Chicago, IL 60612 Tel: (312) 355-0440 Fax: (312) 996-0064 cmatts1@uic.edu or christine.mattson@gmail.com.

HIV and engage in a higher level of risky behavior than they would if they remained uncircumcised (World Health Organization, 2005; World Health Organization (WHO), the United Nations Population Fund (UNFPA), the United Nations Children's Fund (UNICEF), the World Bank, & the UNAIDS Secretariat, 2006). Such risk compensation or behavioral disinhibition could reduce the biologically based protective effect of MC (Bailey et al., 2001; Weiss et al., 2000).

Data from the RCTs of MC have not shown consistent or compelling evidence of risk compensation; however, minimal increases in some risky behaviors have been noted. Further, the evaluation of risk compensation in the RCTs focused on a limited number of variables. In the Orange Farm study, five variables were evaluated: at least one sexual contact without a condom, being married or living as married, more than 1 non-spousal partner, at least one sexual partnership with only one sexual contact, and more than 5 sexual contacts (Auvert, 2005). The Ugandan trial evaluated four behaviors among men reporting sexual activity in the previous 6 or 12 month reference period: condom use (defined as none, inconsistent, or consistent), number of sexual partners (0, 1, 2, 3+), any non-marital partners, alcohol use with sexual intercourse (none versus any), and transactional sex (exchanging money or gifts for sex)(Gray et al., 2007). Finally, the Kenyan trial evaluated the following 5 variables: unprotected intercourse with any partner in the previous 6 months, whether the last time the man had sexual relations, it was with a casual partner, sexual abstinence in the last 6 months, consistent condom use in the previous 6 months, 2 or more partners in the previous 6 months (Bailey et al., 2007). While these variables are obviously important indicators of risk behavior, they are by no means the only ones that might be investigated. A more extensive evaluation of sexual behaviors could provide important insights into the occurrence of risk compensation associated with male circumcision.

Within the field, there is no agreed upon method of evaluating risk compensation or behavioral disinhibition. Although some researchers use biologic markers of sexual activity (pregnancy or an incident sexually transmitted infection (STI)) to characterize sexual risk behavior, most primarily rely on self-reported information (Aral and Peterman, 2002; Pequegnat et al., 2000; Schroder et al., 2003b). Considerable research has evaluated optimal methods to collect information about sexual risk behavior (Catania, 2005; Schroder et al., 2003b); however, less has been done to develop comprehensive methods to make use of it. Typically, researchers evaluate sexual behavior in terms of a few key outcome variables such as total number of partners or number of unprotected encounters (Schroder et al., 2003a). However, given the complex transmission dynamics of HIV and STIs, (Catania, 2005) it is useful to describe changes in multiple sexual behaviors, not just a select few. Moreover, from a statistical point of view, analyzing a set of correlated behaviors as though they were independent outcomes is not sound statistical practice. Thus, some method of scaling the behaviors has a great deal to recommend it, both conceptually and statistically. Multiple related variables lend themselves to scaling procedures, which create a single summary variable that represents the degree to which an attribute is present or absent along a continuum.

Scale development has received little attention in the field of epidemiology. Although common in psychology, education, and sociology, scale development in public health has been primarily limited to studies of quality-of-life or functional status. (Adams et al., 2005; Noerholm et al., 2004) In reproductive health, instruments have been created to evaluate sensation seeking and sexual compulsivity (Kalichman et al., 1994; Kalichman and Rompa, 2001; Weinhardt et al., 2002), AIDS related stigma (Kalichman et al., 2005) and psycho-social correlates of HIV risk behavior (Basen-Engquist et al., 1999), but few attempts have been made to characterize sexual behavior in terms of a “risk scale.”(Campostrini, 1993; Darke, 1991; Gerbert, 1998; McClelland, 2002; Stigum, 1997; Susser, 1998) One

population-based risk index was developed according to a Delphi method (Campostrini, 1993) and another using a "multivariate ordinal risk method" (Susser, 1998) that relied on *a priori* knowledge about risk behavior and actual HIV seroconversion status. A third risk index was developed to evaluate STI-specific risk potential based on the concept of the basic reproductive ratio (R_0), which is a parameter used to model disease spread in populations by estimating the number of secondary cases produced by an average infected person (Stigum, 1997). These indices may be useful in estimating risk behavior at the population level, but they are not applicable to assessment of individual risk or behavior change.

Two studies based on individual level data utilized classical test theory (CTT) or "true score" approaches (Darke, 1991; Gerbert, 1998). CTT methodologies have dominated the field of scale development until relatively recently (Nunnally, 1994). CTT approaches are gradually being supplanted by methods based on Item Response Theory (IRT), which, despite its introduction nearly 50 years ago (Lord, 1980; Rasch, 1960) is only now gaining widespread acceptance. We identified one HIV risk behavior scale that was developed according to the Rasch model, which can be viewed as a particular type of IRT model. The scale was created to evaluate sexual behavior and injection drug use practices among female jail detainees in the USA (McClelland, 2002). However, the instrument was not ideal because it contained some items not appropriate for all study populations (e.g. injection drug use) and lacked others (e.g. concurrent partnerships).

In order to ultimately conduct an extensive, comprehensive evaluation of risk compensation associated with male circumcision, the purpose of this study was: 1) to develop a scale to describe sexual behavior or "sexual risk propensity" and to test it with an IRT model, 2) to test the scales' ability to perform at multiple time points and in different languages, and 3) to demonstrate criterion validity (with STI and HIV as criteria). The successful development of such a scale could provide a useful template methodology for evaluating behavior change in response to a variety of reproductive health interventions (e.g. vaccine and microbicide trials, behavioral interventions, etc.).

Although CTT has served as the basis of scale construction for many decades and is well known to almost all social scientists, scaling methods based on CTT have a number of disadvantages: 1) they assume homoscedasticity, meaning that the error of measurement is the same at the high and low end of the scales as it is in the middle; 2) they are not well equipped to handle missing data; 3) traditionally, they require all variables in the scale to be in the same form (e.g. all dichotomous or all rating scale), although this is no longer the case given modern methods of factor analysis; 4) CTT relies almost entirely on correlational data, but correlations are sharply attenuated when dealing with relatively rare behaviors.

Despite their relative recency and unfamiliarity to many researchers, IRT methods are advantageous because: 1) they incorporate both variation in persons on the trait being measured and variations in item difficulty, 2) they allow missing data (individuals can fail to answer items and still have a summary score computed), item applicability can be assessed through time or according to subject characteristics (e.g. males versus females), and analyses can include a mixture of items scores (e.g., a mixture of dichotomous items and Likert-scale items). Essentially, IRT provides a way to rigorously evaluate the theoretical justification for each item by testing a set of assumptions and model fit.

Our use of IRT is a bit unusual in that we are dealing with self reported behaviors rather than response to attitude items. However, there is no inherent reason why IRT methods can't be used for this purpose. For example, Raudenbush and his colleagues have used this approach to scale self-reported criminal behaviors (Johnson, 2006; Raudenbush, 2003).

METHODS

To evaluate whether circumcision alters men's sexual risk behavior, we recruited men who were enrolled in the Kenyan RCT of MC between March 2004 and September 2005 to participate in this study. Respondents included men aged 18-24 years, who were residents of Kisumu District, who had experienced sex within the last 12 months, and were uncircumcised at baseline. All respondents provided written informed consent in their language of choice: English, DhoLuo or Kiswahili, which are the three most common languages spoken in Kisumu. The sexual history instrument and consent documents were developed in English and were translated into DhoLuo and Kiswahili independently by two indigenous speakers and discrepancies were resolved with assistance from a third individual. Interviews were conducted by men fluent in all three languages. The research protocol was approved by the Kenyatta National Hospital Ethics and Research committee, the University of Illinois at Chicago's Institutional Review Board #3, and the University of Manitoba Biomedical Research Ethics Board.

Data Collection and Management

To obtain comprehensive sexual histories, we adapted the well-validated Timeline Followback (TLFB) approach (Carey et al., 2001) to collect information about every sexual relationship in the last 6 months for up to 12 partners. The following variables were obtained for each partner: age, gender, type of partner, beginning and end dates of the relationship, length of time knowing the partner prior to sex, approximate number of sexual encounters, sexual practices (vaginal, oral, anal), exchange money or gifts for sex, condom use (ever used a condom with the partner, used at the first encounter, at last encounter, and at every encounter), and respondents' perception of their partners' behaviors (e.g. if partners had other partners concurrently, engaged in transactional sex, or were thought to be HIV positive). Men were identified as having a concurrent partnership if the start and end dates of any two partners overlapped by at least one month. We did not attempt to include all known risk factors for HIV. Items were selected for consideration in the scale based on whether previous epidemiologic research had demonstrated that they were risk factors for HIV or STIs in Kisumu (Mattson, et al., 2007) and/or it was postulated that the behavior might be affected by circumcision.

In order to summarize information across 0-12 partnerships, we created count variables to describe the frequency in which a given behavior was reported. Counts that did not yield much variability, or behaviors that were not frequently reported such as the number of partners identified as commercial sex workers in the last 6 months, were re-coded into dichotomous items: any partner identified as a commercial sex worker versus no partner identified as a sex worker. In contrast, behaviors that demonstrated considerable variability (defined by having at least 15% of the sample in each of three categories for at least two visits) were broken down into three categories. For example, total number of sexual partners in the last 6 months was re-coded as: 0 partners, 1 partner, or 2 or more partners.

Interviewers entered data into SPSS Version 10.0. ("SPSS Base 10.0 for Windows User's Guide," 1999) Approximately 30% of files were double entered to evaluate the accuracy of the entry procedures, which proved to have an error rate of less than 1%. Data management and descriptive analyses were performed in SAS Version 8.2 ("SAS OnlineDoc, Version 8.2," 2000).

Laboratory and Diagnostic Procedures

An STI was defined by a laboratory-based diagnosis of gonorrhea, chlamydial infection or trichomoniasis, using the following diagnostic criteria: *Neisseria gonorrhoeae* (Ng) and

Chlamydia trachomatis (CT): by polymerase chain reaction (PCR) assay (AMPLICOR® CT/NG Test, Roche Diagnostics, Montreal Canada) and *Trichomonas vaginalis* (TV): by culture (InPouch™ TV test, Biomed Diagnostics, Oregon, United States). Also, since men were treated for STIs at baseline, subsequent infections were considered incident. Men were tested for HIV-1 using two parallel rapid tests: Determine HIV 1/2 (Abbott Diagnostic Division, Hoofddorp, Netherlands) and the recombinant antigen test Unigold Recombigen HIV Test (Trinity Biotech, Wicklow, Ireland). Men with discordant rapid tests results underwent further testing by double ELISA (Detect HIV 1/2, Adaltis Inc, Montreal, Canada, and Recombigen HIV 1/2, Trinity Biotech, Wicklow, Ireland). Because all men tested negative for HIV at baseline, HIV positive test results at 6 and 12 month visits were considered incident.

Scale Development

Based on each person's scores on items of a test, Item Response Theory (IRT) enables users to construct a scale that can measure the latent trait of each test respondent (e.g., sexual risk propensity), while accounting for any differences among the test items. In this framework, θ is defined as the level of the latent trait that the person possesses. In our study, based on interview data of sexual risk behaviors, we assume the latent trait θ to define a level of sexual risk.

An IRT model assumes that the probability of a person endorsing a given item depends on that person's level on the latent variable and the item. This dependence is captured by the Item-Step Response Function (ISRF). For any test item scored on m+1 ordered levels (k=0,1,...,m), this function is a probability function defined by:

$$\Pr[\,Item\ Response \geq k|\theta]$$

In words, the ISRF denotes the probability that a person with latent trait level θ attains a score of at least k on the test item. All unidimensional IRT models for latent trait measurement are characterized by at least three assumptions (Holland, 1986; Junker, 2001), which are as follows:

1. *Continuity:* The latent trait θ can be represented by a unidimensional numerical continuum;
2. *Local Independence*: Given any specific level of the latent trait θ, the responses to all the items of the test are independent.
3. *Monotonicity of the ISRF*: As θ increases, the ISRF Pr[*Item Response* ≥ $k|\theta$] does not decrease (for all item categories k = 0,1,2,...).

These three assumptions alone define the Monotone Homogeneity Model (MHM). Data which meet these assumptions will show monotonic IRSFs and produce a unidimensional scale in IRT terms. In fact, a set of test items consistent with a factor analysis model with multiple factors and all factor loadings positive, is also consistent with the monotone homogeneity model (Holland, 1986).

Provided that a set of test items satisfy the three assumptions, a person's score can be estimated by his total test score, that is, his/her total score over all the items of the test (Van der Ark, 2005). Likewise, a person's performance on a test can be summarized by his/her proportion score on the test:

$$\text{Person Proportion Score}=\frac{\text{person's total score on the test}-\text{minimum possible score on the test}}{\text{maximum possible score on the test}-\text{minimum possible score on the test}}$$

Given a sample of persons from a population of persons, item difficulty can be estimated by 1 minus the proportion of persons in the sample that endorses the item. This proportion, called the item proportion score, is defined by:

$$\text{Item Proportion Score}=\frac{\text{total item score}-\text{minimum possible total item score}}{\text{maximum possible total item score}-\text{minimum possible total item score}}$$

Usually, person location and item difficulty are reported on the logit scale. The logit person score is defined by:

$$\text{Logit Person Score}=\log\left[\frac{Person\ Proportion\ Score}{1-Person\ Proportion\ Score}\right]$$

and item difficulty on the logit scale is defined by

$$\text{Logit Item Difficulty}=\log\left[\frac{1-Item\ Proportion\ Score}{Item\ Proportion\ Score}\right]$$

In many IRT models, the ISRF Pr[*Item Response* ≥ *k*|θ] is defined by the logistic (distribution) function, or by some other parametric function such as the standard-normal distribution function. To convey this idea directly, consider the simple case where all the test items are scored in two categories (*k*=0,1), so that for *k*=1, the ISRF is defined by

$$\Pr\left[Item\ Response\geq k|\theta\right]=\Pr\left[Item\ Response=1|\theta\right]$$

Then for such items, the "two-parameter logistic" IRT model assumes the ISRF to be a logistic (distribution) function, defined by:

$$\Pr\left[Item\ Response=1|\theta\right]=\frac{\exp\left(Item\ Slope^{*}\left(\theta-Item\ Difficulty\right)\right)}{1+\exp\ Item\ Slope^{*}\left(\theta-Item\ Difficulty\right)\right)}$$

where "Item Difficulty" denotes the parameter of the difficulty of the item," Item Slope" is the slope of the item's ISRF. In the Rasch model, the slope of all the items are assumed to equal 1, while in the "two-parameter logistic" IRT model, the items are allowed to have different slopes. Figure 1 provides examples of three ISRFs under the two-parameter IRT and Rasch models. Notice in this figure that ISRF of item 2 is higher than the slope of the ISRFs of items 1 and 3.

While the Rasch model and the two-parameter model are consistent with the three assumptions, each of these models makes the strong assumption that, in the population of persons, the ISRF can be described exactly by a parametric, S-shaped logistic distribution function (see Figure 1). Unfortunately, it is difficult to argue that for any given test item, its "true" ISRF in a person population is *exactly* S-shaped. Moreover, S-shaped curves are not even necessary for monotonicity; in particular, there exist ISRFs that which are not S-shaped

but are consistent with the 3 basic assumptions of a unidimensional latent trait model. Thus, in any of these three models, a given item, with a monotonic ISRF, may be incorrectly judged as not contributing to the measurement of a latent trait, simply because the item has an ISRF that is not S-shaped.

Such difficulties are avoided through the use of a more flexible, nonparametric IRT model (NIRT) (Mokken, 1971). Such a model only assumes that the ISRF is monotonic, and thus does not make any additional assumptions about it's shape. For each item and each of the item categories, nonparametric IRT analysis involves estimating the true ISRF in the population using the kernel regression method (Ramsay, 1991), and then comparing that estimate against the ISRF Pr[*Item Response* ≥ $k|\theta$] estimated under the constraint that it is monotonic. Then to determine whether a given item contributes to the measurement of the latent trait (e.g. sexual risk propensity), the estimated true ISRF is compared against the monotonic ISRF through a statistical test (Azzalini, 1989). Moreover, given the estimated true ISRFs, it is possible to perform statistical tests of invariant item ordering.

Using the kernel approach to IRT, each point of a true ISRF is estimated by taking a local average of the (dichotomized) item responses with respect to a neighborhood of the value of the proportion total test score and the size of the neighborhood is determined by a bandwidth parameter. The bandwidth controls the tradeoff between sampling variation and bias, where low values of the bandwidth produce functions with large variance and small bias and high values of the bandwidth yield small variance, but large bias. As suggested elsewhere, a useful, automatic choice of bandwidth parameter is defined by Bandwidth = $1.1^{n-.2}$, where n denotes the number of persons in the sample of data (Douglas J, 2001). A recent and accessible discussion of non-parametric IRT methods applied to health care data can be found in Sijtsma et al. (Sijtsma, 2008). Our analyses were performed using program written for the R statistical package (Karabatsos, 2006), and a copy of the program can be obtained through e-mail correspondence with the third author.

RESULTS

Between March 2004 and September 2005, 1319 of the 1780 eligible participants in the parent RCT of MC enrolled in the study, yielding an overall response rate of 74%. For a complete description of the sample, see Table I. Men who enrolled were slightly more likely to have been randomized to the control (53%) versus circumcision (47%) arm ($p < 0.001$) of the main study, were younger (46% vs. 41% $p = 0.03$), more educated (58% vs. 52% completed secondary school, $p = 0.03$), and more likely to be unemployed (67% vs. 60%, $p = 0.02$) than those who did not enroll. There were no statistically significant differences between the median number of lifetime sex partners (Wilcoxon Two Sample Z Test = 0.01, $p = 0.95$), number of sex partners in the last 6 months ($\chi^2 = 0.53$, $p = 0.77$), or occurrence of a prevalent sexually transmitted infection ($\chi^2 = 0.17$, $p = 0.68$) at baseline between the men who enrolled and those who did not.

Testing the Scale with the IRT Model

The “sexual risk propensity” instrument contained 18 items with item difficulty at the baseline visit ranging on the logit scale from - 0.42 for the most frequently reported behavior (total number of partners) to +4.22 logit for the least reported behavior (unprotected sex with a sex worker). Item difficulty was calculated as the natural log ((maximum score in item over all respondents-total score in item over respondents/ (total score in item over all respondents)). The mean person score was -1.31 (95% CI -4.62-0.83) and mean item score was 1.52 (95% CI -0.5-3.6) See figures 2 and 3 for a graphical depiction of the distribution of scores, which indicate an approximate normal distribution for items and a slightly skewed distribution for people. The reliability or internal consistency of the instrument at baseline,

estimated by Cronbach's alpha via the bootstrap approach (Efron, 1993), was 0.87 (95% CI 0.86-0.87). Table II provides a complete description of the items and their difficulty in descending order.

Monotonicity was tested using an approach developed by Azzalini, et al, (1989) involving a statistical comparison of the fit of a regression curve under shape constraints (e.g. monotonicity) against the true regression curve estimated under the kernel approach. No violations of monotonicity were present in the instrument when utilizing data from the baseline visit. Because we intend to use the scale to evaluate behavior change through time, we also assessed the scales' properties at the 6 and 12 month follow-up visits. At the 6 month follow-up visit we identified no violations of monotonicity. At the 12 month follow-up visit we identified one violation of monotonicity for the variable that measured unprotected sex with more than 1 regular partner (see figure 4). We eliminated this item from the scale and re-evaluated its properties. Because the scale did not significantly change without the item, and because the item is important from a conceptual stand point, we chose to retain it.

Testing the Scales Ability to Perform at Different Time Points and in Different Languages

Differential Item Functioning Analysis (DIF) occurs when an item has differing IRSF's across time or groups. In order to test for this, we first evaluated the stability of the ISRFs at different time points. We compared the estimated ISRF of each item at baseline and 6 months, baseline and 12 months and at 6 months and 12 months in order to identify any statistically significant differences in item function, as defined by curves where the 95% confidence intervals (estimated by the bootstrap approach, (Efron, 1993) did not overlap. We identified one instance of differential item function with the item "had a concurrent partner in the last 6 months" when comparing the item at the 6 and 12 month follow-up visits (See figure 5). Although the curves do not overlap, they maintain a similar direction, and given the conceptual relevance of the item and since it performs consistently at the other time points, we chose to retain it.

Because participants were given a choice to conduct the interviews in English, DhoLuo or Kiswahili, we also analyzed differential item function according to the language of the interview. The majority of men, 60%, chose to be interviewed in English, 40% chose DhoLuo and 1% chose Kiswahili. When we evaluated potential differences in items administered in English versus DhoLuo/Kiswahili at the baseline visit, we found no statistically significant differences in item function according to the language of the interview.

Criterion Validity

To establish the scale's criterion validity we evaluated the association between men's scores on the risk scale and the presence of an incident STI or incident HIV infection. In order to ensure that the reported behavior preceded STI or HIV acquisition, we included incident infections that were diagnosed within 3 months prior to the interview or 1 month afterwards at the 6 or 12 month follow-up visits. At the 6 month follow-up visit, 44 men were diagnosed with an incident STI: 22 had gonorrhea, 18 had chlamydial infection, 1 had trichomoniasis and 3 men were co-infected with gonorrhea and chlamydia. At the 12 month follow-up visit, 24 men were diagnosed with an incident STI: 6 had gonorrhea and 18 had chlamydial infection. Exact logistic regression analyses showed that men diagnosed with an incident STI had higher mean logit values (indicative of higher risk behavior) than uninfected men, but the difference was only statistically significant at the 6 month follow-up visit (OR 1.3, 95% CI 1.05-1.61), not at the 12 month follow-up visit (OR = 1.13, 95% CI 0.85-1.52).

At the 6 and 12 month follow-up visits, 8 and 4 men were respectively diagnosed with incident HIV infection. Similar to the STI analyses, men who seroconverted throughout the study had higher risk scores than those who did not. The non-parametric Savage Two-Sample Test with one-sided probability was borderline at 6 months ($p= 0.07$) and statistically significant at 12 months ($p = 0.01$).

Finally, to compare the sexual risk propensity scores' ability to predict incident STIs to commonly used single behavioral variables (e.g. total number of sexual partners in the last 6 months), we conducted random effects logistic regression analyses. Three models were run where incident STI at the 6 month follow up visit was the dependent variable and the independent variables were the risk score, total number of sexual partners in the last 6 months, and total number of partners where a condom was not always used in the last 6 months The risk score yielded an OR of 1.26 (95% CI 1.05 - 1.52), total number of sexual partners in the last 6 months had an OR of 1.68 (95 % CI 1.01-2.69) and total number of partners where a condom was not always used had an OR of 2.16 (95 % CI 1.17-3.70). Thus, all three variables were statistically significant predictors. Unfortunately, there is no way to perform a statistical test to identify the "best model" since likelihood ratio statistics cannot be compared when models are not nested. Also, because the risk score is on a different scale than the other two measures, the odds ratios cannot be directly compared. However, a notable advantage of the risk propensity score is that it contains more information and thus yields tighter confidence intervals around the estimate. Note that the range of the CI for the risk scale is on the order of .5 while it is more than five times as large for the other two variables. As a result, the risk score provides greater statistical power to detect smaller changes in risk behavior and requires smaller study samples than dichotomous or ordinal variables.

DISCUSSION

The results of this study indicate that it is possible to use non-parametric item response theory to create a robust scale measuring sexual behavior or "sexual risk propensity". The 18 item scale demonstrated very good reliability (Cronbach's alpha of 0.87) and produced a logical, unidimensional continuum to represent sexual risk behavior. We found only one violation of monotonicity at the 12 month follow-up visit. Because the properties of the scale were not altered by removing the item, it was retained on conceptual grounds. We found no evidence of differential item function at different time points (except for reporting a concurrent partner at the 6 and 12 month follow-up visits) or with respect to the language with which the instrument was administered. Further, we established criterion validity by demonstrating a statistically significant association between the risk scale and the biologic outcome of an incident STI at the 6 month follow-up visit and HIV at the 12 month follow-up visit. Men repeatedly received risk reduction counseling throughout the course of the parent study, so it is not surprising that risk scores decreased at the 6 and 12 month follow-up visits or that the number of men diagnosed with an incident STI also declined from the 6 month to the 12 month follow-up visit. In fact, a longitudinal random effects regression analysis indicated a statistically significant decline in risk scores from the 6 to 12 month follow-up visit (OR = 0.55, 95% CI 0.34-0.89) . The lack of a significant association between risk propensity scores and incident STIs at the 12 month follow-up and HIV at the 6 month is likely due to the very small sample sizes.

This study has limitations. We chose to use non-parametric IRT to evaluate the properties of the sexual risk propensity score. This method may be unfamiliar to many epidemiologists or other public health researchers and may seem too complicated to use. However, given the frequent need in HIV/AIDS research to evaluate multiple, related behavioral outcomes, and the lack of an existing gold standard to do so, the minor inconvenience of learning a new

methodology is far outweighed by the advantages of NIRT. Next, we collected behavioral data using the Timeline Followback approach, which is most often criticized for being too time consuming (Wennberg, 1998). However, we did not find this to be the case. In fact, on average, the interview was completed in 9 minutes (range 1-85, s.d. = 7.0) and on average, 1.52 partners were discussed (range 1-12, s.d. = 1.4). Finally, we would not argue that this scale or even these specific items are appropriate in all times and all places. On the contrary, we would expect to see variation in relevant behaviors across cultures. Thus, while we think that we have established that this 18 item scale is appropriate for this particular application, other researchers may well wish to establish their own measurement methods. We think that the approach taken here, non-parametric IRT, is worth emulating and we hope that as other researchers begin to do so we will arrive at a core set of items which show consistent usefulness in all or most situations in conjunction with other items which represent local cultural norms and practices.

CONCLUSION

We have demonstrated the utility of IRT methods to describe sexual risk behavior or “sexual risk propensity” in the context of a randomized controlled trial of male circumcision to reduce HIV incidence in Kisumu, Kenya. The method has broad applicability and could be implemented to summarize and evaluate sexual behavior in the context of other HIV or STI prevention studies (e.g. behavioral interventions, and vaccine or microbicide trials) or to assess behavior may change after HIV positive individuals initiate antiretroviral therapy. Given the lack of methods available to comprehensively analyze sexual behavior data, this method may fill a pervasive gap in our ability to assess sexual risk behavior.

Acknowledgments

We thank all of the participants, without whom this work would not have been possible. We are grateful to Evans Otieno, Nicholas Ouma, Bob Ogollah, Kevine Kamollah, and the entire UNIM Project staff for their assistance in data collection and recruitment efforts and to Dr. Donald Hedeker, Dr. Ronald Hershow, and Nelli Westercamp for their helpful comments on the manuscript.

REFERENCES

Adams R, Rosier M, Campbell D, Ruffin R. Assessment of an asthma quality of life scale using item-response theory. Respirology 2005;10(5):587–593. [PubMed: 16268911]

Aral SO, Peterman TA. A stratified approach to untangling the behavioral/biomedical outcomes conundrum. Sexually Transmitted Diseases 2002;29(9):530–532. [comment]. [PubMed: 12218844]

Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R, Puren A. Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 trial. PLoS Medicine 2005;2(11):e298. [PubMed: 16231970]

Azzalini AWB, Hardle W. On the use of nonparametric regression for model checking. Biometrika 1989;76:1–11.

Bailey R, Moses S, Parker CB, Agot K, Maclean I, Krieger JN, Williams CFM, Campbell RT, Ndinya-Achola JO. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. The Lancet 2007;369(9562):643–656.

Bailey RC, Plummer FA, Moses S. Male circumcision and HIV prevention: current knowledge and future research directions. Lancet Infectious Disease 2001;2001(1):223–231. [Review] [69 refs].

Basen-Engquist K, Masse LC, Coyle K, Kirby D, Parcel GS, Banspach S, et al. Validity of scales measuring the psychosocial determinants of HIV/STD-related risk behavior in adolescents. Health Education Research 1999;14(1):25–38. [PubMed: 10537945]

Campostrini S, McQueen DV. Sexual behavior and exposure to HIV infection: estimates from a general population risk index. American Journal of Public Health 1993;83:1139–1143. [PubMed: 8342723]

Carey MP, Carey KB, Maisto SA, Gordon CM, Weinhardt LS. Assessing sexual risk behaviour with the Timeline Followback (TLFB) approach: continued development and psychometric evaluation with psychiatric outpatients. International Journal of STD & AIDS 2001;12(6):365–375. [PubMed: 11368817]

Catania JA, Osmond D, Neilands TB, Canchola J, Gregorich S, Shiboski S. Commentary on Schroder et al (2003a, 2003b). Annals of Behavioral Medicine 2005;29(2):86–95. [PubMed: 15823781]

Darke S, Hall W, Heather N, Ward J, Wodak A. The reliability and validity of a scale to measure HIV risk-taking behavior among intravenous drug users. AIDS 1991;5(2):181–185. [PubMed: 2031690]

Douglas, J, a. C.; A. Nonparametric ICC estimation to assess fit of parametric models. Applied Psychological Measurement 2001;25:234–243.

Efron, BT.; RJ. An Introduction to the Bootstrap. Chapman and Hall; 1993.

Gerbert B, Bronstone A, McPhee S, Pantilat S, Allerton M. Development and testing of an HIV-risk screening instrument for use in health care settings. American Journal of Preventive Medicine 1998;15(2):103–113. [PubMed: 9713665]

Gray RH, Kigozi G, Serwadda D, Makumbi F, Watya S, Nalugoda F, et al. Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. The Lancet 2007;369(9562):657–666.

Holland P, Rosenbaum PR. Conditional association and unidimensionality in monotone latent variable models. The Annals of Statistics 1986;14(4):1523–1543.

Johnson S, Raudenbush SW. A repeated measures, multilevel Rasch model with application to self-reported criminal behavior. Methodological issues in Aging Reserach 2006;5:131–164.

Junker B, Sijtsma K. Nonparametric item response theory in action: an overview of a special issue. Applied Psychological Measurement 2001;25(3):211–220.

Kalichman SC, Johnson JR, Adair V, Rompa D, Multhauf K, Kelly JA. Sexual sensation seeking: scale development and predicting AIDS-risk behavior among homosexually active men. Journal of Personality Assessment 1994;62(3):385–397. [PubMed: 8027907]

Kalichman SC, Rompa D. Sexual sensation seeking and Sexual Compulsivity Scales: reliability, validity, and predicting HIV risk behavior. Journal of Personality Assessment 1995;65(3):586–601. [PubMed: 8609589]

Kalichman SC, Rompa D. The Sexual Compulsivity Scale: further development and use with HIV-positive persons. Journal of Personality Assessment 2001;76(3):379–395. [PubMed: 11499453]

Kalichman SC, Simbayi LC, Jooste S, Toefy Y, Cain D, Cherry C, et al. Development of a brief scale to measure AIDS-related stigma in South Africa. AIDS & Behavior 2005;9(2):135–143. [PubMed: 15933833]

Karabatsos, G. An R program for nonparametric Item Response Theory (Version (Copyright 2007)). Chicago: 2006.

Lord, FM. Applications of item response theory to practical testing problems. New Jersey: 1980.

Mattson CL, Bailey RC, Agot K, Ndinya-Achola JO, Moses S. A nested case-control study of sexual practices and risk factors for prevalent HIV-1 infection among young men in Kisumu, Kenya. Sexually Transmitted Diseases 2007;34(10):731–736. [PubMed: 17495591]

McClelland G, Teplin LA, Abram KM, Jacobs N. HIV and AIDS risk behaviors among female fail detainees: implications for public health policy. American Journal of Public Health 2002;92(5):818–825. [PubMed: 11988453]

Mokken, RJ. A theory and procedure of scale analysis. de Gruyter; Berlin: 1971.

Noerholm V, Groenvold M, Watt T, Bjorner JB, Rasmussen NA, Bech P. Quality of life in the Danish general population--normative data and validity of WHOQOL-BREF using Rasch and item response theory models. Quality of Life Research 2004;13(2):531–540. [PubMed: 15085925]

Nunnally, J.; Bernstein, IH. Psychometric Theory. 3rd ed.. McGraw Hill; 1994.

Pequegnat W, Fishbein M, Celentano D, Ehrhardt A, Garnett G, Holtgrave D, et al. NIMH/APPC workgroup on behavioral and biological outcomes in HIV/STD prevention studies: a position statement. Sexually Transmitted Diseases 2000;27(3):127–132. [PubMed: 10726643]

Ramsay J. Kernel smoothing approaches to nonparametric item resopnse curve estimation. Psychometrika 1991;56(4):611–630.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Danmarks Paedogogiske Institute; Copenhagen: 1960.

Raudenbush S, Johnson C, Sampson RJ. A multivariate, multilevel Rasch model with application to self-reported criminal behavior. Sociological Methodology 2003;33:169–211.

SAS OnlineDoc, Version 8.2. SAS Institute Inc.; Cary, NC: 2000. (2000)

Schroder KE, Carey MP, Vanable PA. Methodological challenges in research on sexual risk behavior: I. Item content, scaling, and data analytical options. Annals of Behavioral Medicine 2003a;26(2): 76–103. [see comment]. [PubMed: 14534027]

Schroder KE, Carey MP, Vanable PA. Methodological challenges in research on sexual risk behavior: II. Accuracy of self-reports. Annals of Behavioral Medicine 2003b;26(2):104–123. [see comment]. [PubMed: 14534028]

Sijtsma K, Emons WHM, Bouwmeester SB, Nycliček I, Roorda LD. Nonparametric IRT analysis of quality of life scales and its application to the World Health Organization quality of life scale (WHOQOQ-Bref). Quality of Life Research 2008;17:275–290. [PubMed: 18246447]

SPSS Base 10.0 for Windows User's Guide. SPSS Inc.; Chicago: 1999. (1999)

Stigum H, Magnus P. A risk index for sexually transmitted diseases. Sexually Transmitted Diseases 1997;24:102–108. [PubMed: 9111756]

Susser E, Desvarieux M, Wittkowski KM. Reporting sexual risk behavior for HIV: a practical risk index and a method for improving risk indices. American Journal of Public Health 1998;88(4): 671–674. [PubMed: 9551017]

Van der Ark LA. Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. Psychometrika 2005;70(283-304)

Weinhardt LS, Otto-Salaj LL, Brondino MJ, Norberg MM, Kalichman SC. Sex-related alcohol expectancies predict sexual risk behavior among severely and persistently mentally ill adults. Psychology of Addictive Behaviors 2002;16(1):64–67. [PubMed: 11934088]

Weiss H, Quigley M, Hayes R. Male circumcision and risk of HIV infection in sub-Saharan Africa: a systematic review and meta-analysis. AIDS 2000;14:2261–2370.

Wennberg P, Bohman M. The Timeline Follow Back Technique: Psychometric Properties of a 28-day Timeline for Measuring Alcohol Consumption. German Journal Psychiatry 1998;2:62–68.

World Health Organization. UNAIDS statement on South African trial findings regarding male circumcision and HIV Statement developed by the World Health Organization (WHO), the United Nations Population Fund (UNFPA,) the United Nations Children's Fund (UNICEF) and the UNAIDS Secretariat, 26 July 2005. 2005 [15 March, 2006]. http://www.who.int/mediacentre/news/releases/2005/pr32/en/.

World Health Organization (WHO). the United Nations Population Fund (UNFPA). the United Nations Children's Fund (UNICEF). the World Bank, & the UNAIDS Secretariat. Statement on Kenyan and Ugandan trial findings regarding male circumcision and HIV. Statement developed by the World Health Organization (WHO), the United Nations Population Fund (UNFPA), the United Nations Children's Fund (UNICEF), the World Bank and the UNAIDS Secretariat. 2006 [December 13, 2006]. http://www.who.int/mediacentre/news/statements/2006/s18/en/index.html

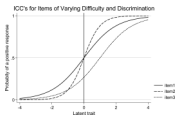


Figure 1.

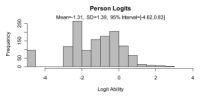


Figure 2.
Person Logit Distributions at the Baseline Visit

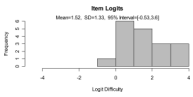


Figure 3.
Item Logit Distributions at the Baseline Visit

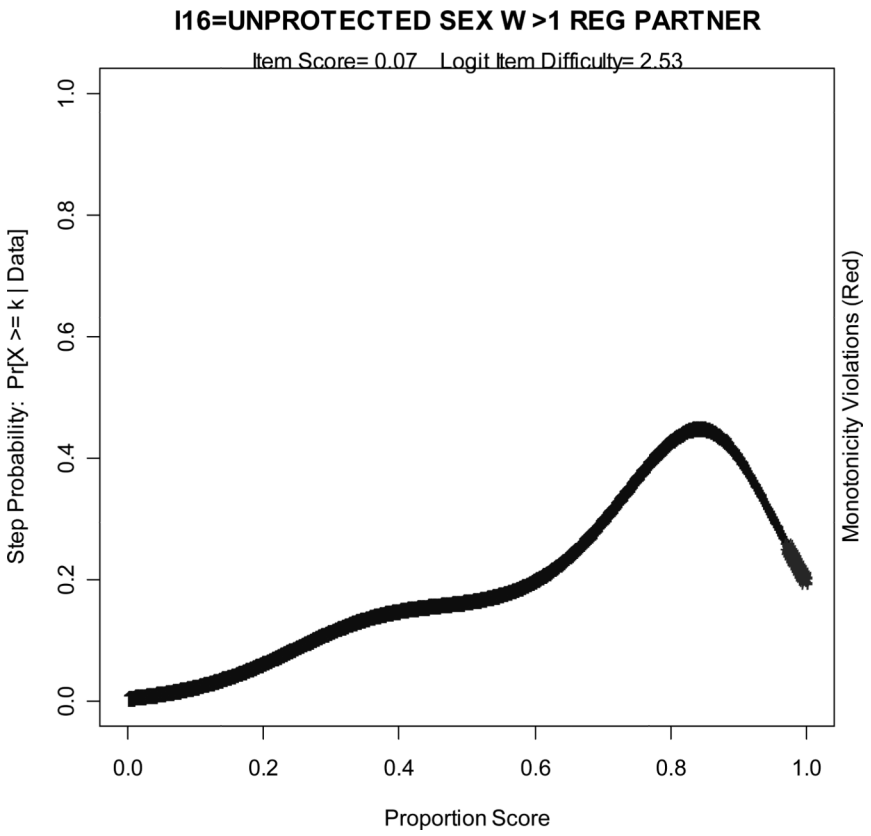
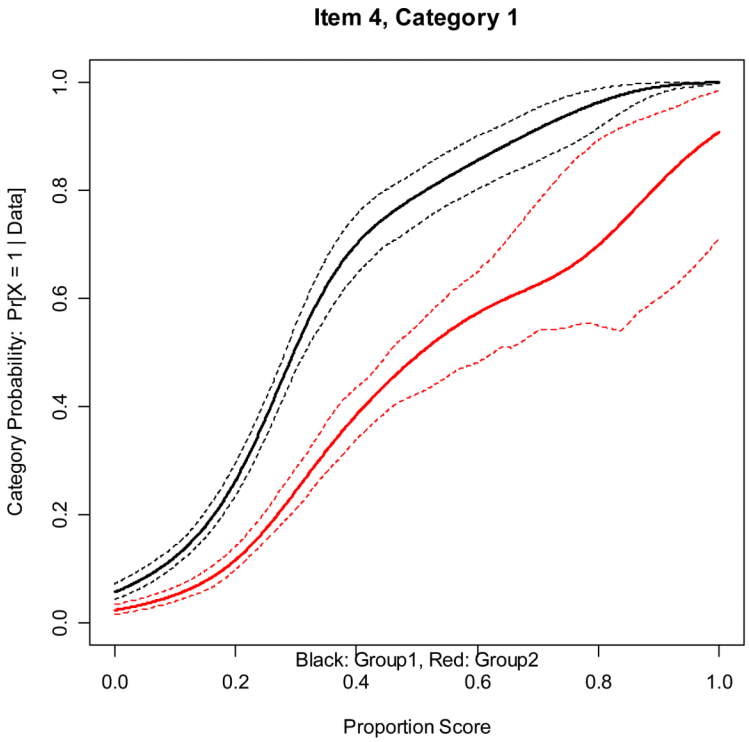


Figure 4.
A nonmonotonic curve: unprotected sex with more than 1 regular partner at the 12 month follow-up visit



First three lines from the top: In the 6-month follow-up data, the estimated ISRF of item 4 for category 1 (solid line), and the 95% confidence interval (dashed lines) estimated by the bootstrap procedure.

Bottom three lines: In the 12-month follow-up data, the estimated ISRF of item 4 for category 1 (solid line), and the 95% confidence interval (dashed lines) estimated by the bootstrap procedure.

Figure 5.
Differential item function: concurrent partnership in the last 6 months when comparing the 6 and 12 month follow up visit (where category one indicates the presence of a concurrent partnership in the last 6 months)

Table I

Baseline Comparison of Eligible Men who Enrolled in Study and Those Who Did Not (n = 1780)

Variable	Enrolled in study (n = 1319)	Did not Enroll (n = 461)	Test Statistic	p-value
Treatment Assignment				
Circumcision	622 (47)	266 (58)	15.2	< 0.001
Control	697 (53)	195 (42)		
Age				
18-20	600 (46)	185 (41)	3.6	0.03
21-24	700 (54)	266 (59)		
Education				
Primary (0-8)	455 (34)	191 (41)	7.2	0.03
Secondary (9-12)	761 (58)	236 (51)		
Post-Secondary (13 or more)	103 (8)	34 (7)		
Employment Status				
Employed	94 (7)	43 (9)	8.4	0.02
Self-Employed	338 (26)	142 (31)		
Unemployed	887 (67)	276 (60)		
Occupation				
Farm laborer/Fisherman	119 (9)	33 (7)	11.2	0.05
Professional/Managerial	9 (<1)	10 (2)		
Semi-skilled worker	78 (6)	33 (7)		
Skilled Worker	92 (7)	40 (9)		
Student	293 (22)	94 (20)		
Unskilled Worker	728 (55)	251 (54)		
Income				
2000 ksh/month or less	702 (53)	230 (50)	1.3	0.25
More than 2000 ksh/month	616 (46)	231 (50)		
Marital Status				
Married	72 (5)	30 (7)	0.68	0.40
Single	1246 (95)	431 (93)		
Lifetime Sex Partners*	Median	Median		
	4.0	4.0	0.01	0.95
	IQR 1-7	IQR: 1-7		
Number of Sex Partners last 6 months				
None	141 (11)	53 (11)	0.53	0.77
One	577 (44)	207 (45)		
More than one	598 (45)	201 (44)		
Diagnosed with a STI at baseline				
Yes	114 (9)	37 (8)	0.17	0.68
No	1205 (91)	424 (92)		

*Medians tested by the Wilcoxon Two Sample Z Test

Table II

The 18-item Sexual Behavioral Scale at the Baseline Visit

Item Description	Item Difficulty (logit scores)
Had unprotected sex with a sex worker	4.22
Always exchanged $ for sex w/ partner not identified as a sex worker	3.98
Believed a partner had HIV/AIDS at the time of the relationship	3.55
Had sex with a partner identified as a commercial sex worker	3.22
Had unprotected sex w/ partner after knowing her <=1 day	3.12
Had unprotected sex with >1 “casual” partner	2.97
Had unprotected sex with >1 “regular” partner	2.45
Ever exchanged $ for sex w/ partner not identified as a sex worker	2.18
Had sex with a partner after knowing her <= day	1.99
Had sex with a partner while she was menstruating	1.97
Believed a partner had sex with others for $ at time of relationship	1.81
Had unprotected sex with >1 partner	1.67
Believe a partner had other “regular” partners at time of relationship	1.05
Believe a partner had other “casual” partners at time of relationship	0.98
Believe a partner had any other partners at time of relationship	0.78
Had a concurrent partnership	0.76
*Total number of unprotected partners (2, 1, 0)	0.64
*Total number of sex partners (2, 1, 0)	-0.42

* Indicate the only non-dichotomous items (2+ partners, 1 partner, no partners)