



Item response theory and computer adaptive testing of the sexual knowledge scale of the sexual knowledge and attitude test in a college sample

Martin Komarc, Aya Shigeto & Lawrence M. Scheier

To cite this article: Martin Komarc, Aya Shigeto & Lawrence M. Scheier (22 Mar 2024): Item response theory and computer adaptive testing of the sexual knowledge scale of the sexual knowledge and attitude test in a college sample, *Psychology & Sexuality*, DOI: [10.1080/19419899.2024.2332630](https://doi.org/10.1080/19419899.2024.2332630)

To link to this article: <https://doi.org/10.1080/19419899.2024.2332630>



Published online: 22 Mar 2024.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

RESEARCH ARTICLE



Item response theory and computer adaptive testing of the sexual knowledge scale of the sexual knowledge and attitude test in a college sample

Martin Komarc^a, Aya Shigeto ^b and Lawrence M. Scheier^{c,d}

^aFaculty of Physical Education and Sport, Department of Kinanthropology and Humanities, Charles University in Prague, Prague, Czech Republic; ^bDepartment of Psychology and Neuroscience, Nova Southeastern University, Fort Lauderdale, FL, USA; ^cLARS Research Institute, Inc, Sun City, AZ, USA; ^dDepartment of Public Health Education, Prevention Strategies, Greensboro, NC, USA

ABSTRACT

Our study addresses the limited availability of well-validated scales for assessing general sexual knowledge among college populations. To fill this gap, we examined the psychometric properties of a 41-item sexual knowledge scale derived from the Sexual Knowledge and Attitudes Test – Adolescents (SKAT-A) in a sample of young adults aged 18–25 ($N = 1,291$). We employed classical test theory (CTT) procedures, followed by item response theory (IRT) and computerized adaptive testing (CAT), to refine the SKAT-A knowledge scale. Both CTT and IRT analyses identified six items for removal due to poor discrimination and difficulty parameters. A confirmatory factor analysis supported a well-defined unidimensional latent trait of sexual knowledge. The results of CAT simulations using dynamic item administration demonstrated the scale's measurement precision, with moderate test reliability and relatively low standard errors. On average, there was a 54.3% reduction in the number of items administered without compromising scale reliability. This study concludes that the SKAT-A efficiently assesses a unidimensional trait of sexual knowledge in college-attending young adults. It highlights that only a subset of items from the full test bank is necessary to achieve this, providing a practical and reliable tool for assessing sexual knowledge in this population.



ARTICLE HISTORY

Received 20 February 2023
Accepted 4 March 2024

KEYWORDS

Sexual knowledge; college students; classical test theory; item response theory; computerised adaptive testing

Numerous sex education programmes targeting adolescents and young adults are built on the premise that providing relevant factual information will dissuade youth from engaging in risky sexual behaviour, such as unprotected sex, sex under the influence of alcohol and other substances, and having multiple sexual partners, all of which increase the risk of sexually transmitted infections (STIs) and unplanned pregnancy (e.g. D. B. Kirby et al., 2007; Lightfoot et al., 2015; Wang et al., 2006). The significance of these programmes underscores the increasing prevalence of STI cases within the 15–24 age group (Centers for Disease Control and Prevention [CDC], 2021). According to the National Sex Education Standards (Future of Sex Education Initiative, 2020), key content areas for comprehensive sex education programmes include healthy relationship dynamics, reproductive anatomy and physiology, sexual orientation, gender identity, the modes of transmission and prevention of STI/HIV, and interpersonal and sexual violence. Among these content areas, sexual knowledge stands out as a central focus in many programmes, including abstinence-only/plus programmes (e.g. Arnold et al., 2000; Bennett & Assefi, 2005; Kohler et al., 2008; Lindberg & Maddow-Zimet, 2012),

CONTACT Aya Shigeto  as1959@nova.edu  Department of Psychology and Neuroscience, College of Psychology, Nova Southeastern University, 3301 College Avenue, Fort Lauderdale, FL 33314, USA

comprehensive sex education programmes (Goldfarb & Lieberman, 2021), STI/HIV risk reduction programmes (e.g. D. Kirby, 2007; Walter & Vaughan, 1993), and HIV peer education (e.g. Borgia et al., 2005; Maticka-Tyndale & Barnett, 2010; T. Wong et al., 2019).

Despite the heralded importance of sexual knowledge in sex education (Allen, 2001; Schaalma et al., 2004), there is a paucity of psychometrically refined instruments that assess general sexual knowledge. This puts research on sex education as a whole as well as many individual programmes at a disadvantage. Specifically, a reliable sexual knowledge scale can provide a means to gauge 'how much' young people really know about sexuality, which in turn can undergird epidemiological efforts to establish the extent and accuracy of sexual knowledge in various populations. Studies done in the United States (US; Guzzo & Hayford, 2018; Opt & Loffredo, 2004), China (Davis et al., 1998; Huang et al., 2005; Lyu et al., 2020; Yip et al., 2013; Zhao et al., 2019), and other countries (Butts et al., 2017; Fennie & Laas, 2014; Kumar et al., 2017; Singh et al., 2005; L. P. Wong, 2012; Yoo et al., 2005) are good examples of such epidemiological efforts for adolescent and young adult populations. Notwithstanding these efforts, most of these studies developed instruments specific to their needs and did not provide sufficient evidence of scale reliability. This leaves a gap in the literature because an essential part of psychological assessment is empirically validating that an instrument is internally consistent and accurately measures what it is designed to measure (Nunnally & Bernstein, 1994). In addition, a reliable instrument to measure sexual knowledge can provide a barometer of whether individuals exposed to sex education programmes acquire greater knowledge, which is a useful metric to validate programme effectiveness (e.g. Arnold et al., 2000; Borawski et al., 2005; Carey & Schroder, 2002; Coyle et al., 2021; Singh et al., 2005; Wang et al., 2006).

In the current study, we ascertain the reliability and item performance of a sexual knowledge scale that is part of a more comprehensive instrument – the Sexual Knowledge and Attitudes Test – Adolescents (SKAT-A). The 41-item scale has undergone extensive psychometric analysis using classical test theory (CTT) methods to provide estimates of scale score reliability (Fullard & Scheier, 2011; Fullard et al., 1998; Lief et al., 1990; Motedayen et al., 2019), although it is yet to be tested with young adults. Unlike many other scales that are specific to STI/HIV/AIDS, contraceptive use, or other areas of sexuality (e.g. Carey & Schroder, 2002; Condelli, 2011; Jaworski & Carey, 2007; Kelly et al., 1989; Kutner et al., 2022; McCabe & Cummins, 1996), the SKAT-A knowledge scale assesses 'general' sexual knowledge. The scale captures the total amount of knowledge covering a wide range of sex-related topics, including pregnancy, abortion, condoms or contraception, orgasm, masturbation, STIs, sexual orientation, and sexual violence. To establish the psychometric properties of the SKAT-A knowledge scale with young adult populations, we used both CTT and item response theory (IRT) methods with the latter offering several advantages to traditional CTT methods. Building off the premise of IRT, we also conducted computerised adaptive testing (CAT) in an effort to develop a streamlined version of the knowledge scale that retains its high levels of scale score reliability but with fewer items.

Sexual knowledge scales

Efforts to establish psychometric properties of the SKAT-A knowledge scale can allow us to address two major limitations in research on sexual knowledge: the paucity of scales that assess general sexual knowledge and the lack of psychometrically refined sexual knowledge scales.

A number of domain-specific sexual knowledge scales that have been developed in the US are designed to assess knowledge of STIs and HIV/AIDS. These scales typically cover topics such as modes of transmission, behaviours that increase the risk of transmission, prevention, and treatment. Carey and Schroder (2002) developed an abbreviated 18-item version of the original 45-item HIV Knowledge Questionnaire (Carey et al., 1997) with low-income adults by testing for internal consistency (α 's = .75 ~ .89), validity (r 's = .93 ~ .97), stability over time (r 's = .76 ~ .94), and sensitivity to change following an HIV-risk reduction intervention. Jaworski and Carey (2007) performed iterative testing procedures to reduce 85 items and, from this pool of items, developed a 27-item STD-

Knowledge Questionnaire with US college students. The authors reported good internal consistency ($\alpha = .86$) and test-retest reliability ($r = .88$). Kelly et al. (1989) created a 40-item test of AIDS risk behaviour knowledge (high risk sexual practices, risk reduction approaches, and misconceptions regarding HIV/AIDS), which they normed with a sample of white and black college students and gay men recruited from establishments serving alcohol. The scale demonstrated adequate KR-20 reliability (.74) and test-retest reliability ($r = .84$).

In addition to STI/HIV/AIDS-related knowledge scales, there are a few other domain-specific scales that have been developed in the US, such as the Sexual Knowledge, Experience, Feelings, and Needs Scale (McCabe & Cummins, 1996) with respect to sexuality and disability, the Herpes Knowledge Scale (K. E. M. Bruce & Bullins, 1989; K. Bruce & McLaughlin, 1986) for general knowledge about genital herpes, the Inventory of Anal Sex Knowledge (Kutner et al., 2022) for general knowledge about anal sex, and the Contraceptive Utilities, Intention, and Knowledge Scale (Condelli, 2011), which assesses women's general knowledge of conception and contraception as well as knowledge about the primary contraception that a respondent is currently using. Mackin and Perkhounkova (2019), guided by the National Sexuality Education Standards, developed the Test of Adolescent Sexual Knowledge (TASK), which assesses knowledge in different domains of sexuality, such as anatomy and physiology, health relationships, STI/HIV, personal safety, and puberty. The authors pilot tested the original 86-item version of the scale with a relatively small sample of youth ($n = 132$, mean age 14). Following modifications based on item difficulty reported by the participants, Cronbach's alpha for a modified 82-item version was .93.

Several other studies have developed sexual knowledge scales for application in international settings outside the US. For example, Yoo et al. (2005) developed a 19-item HIV Knowledge scale for South Korean adolescents. The items were translated from a Chinese AIDS knowledge assessment with known scale score reliability ($\alpha = .76$; Davis et al., 1998). Two subscales assessed myths (e.g. HIV transmission by shaking hands) and facts (e.g. HIV transmission from mother to baby), although no information was provided for scale score reliability or factor analysis results. Wang et al. (2006) developed a 38-item knowledge scale to assess intervention effects for a community-based comprehensive sexual education programme targeting adolescents and young adults ages 15 to 24. Knowledge items were relevant to the course content and included sexual physiology, contraceptive methods, and HIV/STD transmission and prevention. Unfortunately, no psychometric information was provided attesting to scale score reliability. Contrasting with the scales with no or limited psychometric information, a notable exception arises from the study conducted by Sanz-Martos et al. (2019). They performed extensive psychometric analyses on a 15-item scale assessing sexuality and contraceptive methods knowledge among Spanish young adults between 18 and 25 years of age. The scale showed good scale score reliability ($\alpha = .73$) and temporal stability ($r = .81$). Using Horn's parallel analysis (i.e. contrasting an empirical quantification of eigenvalues to a simulated model), the authors also demonstrated that the scale comprises a single dimension.

Taken together, this brief review of the literature reveals the scarcity of psychometrically refined scales that measure general sexual knowledge among adolescents and young adults. In addition, many scales are developed specifically for the content of a particular sex education programme. As a result, they have very limited utility outside of their specific application. Moreover, in many cases, the authors failed to report any psychometric properties of the scales (e.g. DeGroot et al., 2014; Lal et al., 2000; Wang et al., 2006; Yoo et al., 2005). Even when they did, they often relied on factor loadings or item-to-total correlations to establish reliability. These statistics, which have been referred to as 'omnibus statistics' (Santor & Ramsay, 1998), can determine scale coherence and its dimensionality (McCrae et al., 2011), but not the relative efficiency of individual items in assessing sexual knowledge.

CTT versus IRT methods

Classical test theory (CTT) has been a mainstay of psychometric scale development for a considerable amount of time (Embretson, 1996; Lord & Novick, 1968). Historically, it has been used to better

understand scale properties including reliability and scale coherence. However, CTT has several drawbacks. First, with CTT, there is a heavy influence of the sample on item characteristics (i.e. difficulty, discrimination) and the interpretation of the test norm. Depending on sample characteristics, the test norm changes, and the interpretation of it also changes, rendering cross-sample comparisons difficult. Second, CTT does not allow researchers to separate the respondent's ability from item performance characteristics. A respondent's underlying ability to perform well on the test (or 'trait') cannot be separated from the item difficulty (Hays et al., 2000). Third, related to the second drawback, CTT requires all of the test items to be administered. Because each item is considered equally effective in measuring the underlying trait, the total or aggregate score is what is informative in the CTT framework. Therefore, the same total test score received by two individuals implies the same trait level for these two individuals despite different patterns of item endorsement.

Application of IRT is one way to address the aforementioned drawbacks of CTT and, at the same time, to conduct a more extensive analysis of item properties. There are numerous examples of where IRT methods have been used to demonstrate the quality of an item and how well it assesses an underlying latent trait (i.e. an underlying propensity to have a certain state of mind; sometimes referred to as 'ability' in education). IRT has been applied to various scales, including those assessing alcohol use disorder (Gelhorn et al., 2008), cannabis use disorder (Compton et al., 2009), depression (K. R. Evans et al., 2004), health-related quality of life (Cook et al., 2007), quality of medical education (De Champlain, 2010), as well as sexual functioning (Sills et al., 2005). One of the many strengths of IRT is that it can be used to examine the performance of individual items and establish their efficiency in determining a respondent's latent trait along a continuum (Embretson & Reise, 2000), including sexual knowledge in the current study. This is because the IRT approach uses a common score scale, which is a mean of the trait set to zero (i.e. a standard normal score), rendering the interpretation of scores identical between different samples or within the same respondent over time.

Two unique features of IRT that improve upon CTT are the item and test information functions. The item information function tells us how well each proficiency or ability level is being estimated by a particular item. The amount of information at a given proficiency level is the inverse of its sampling variance. Therefore, the larger the amount of information provided by the item, the greater the precision of the measurement. In other words, if the amount of information at a particular proficiency level of sexual knowledge is small, the knowledge at that level cannot be estimated as precisely as other proficiency levels where the amount of information provided by an item is large. Given that each of the items *independently* (rather than conjointly) contributes to the measurement in IRT, the test information function equals the sum of all item information functions within the scale (or test), providing an indication of an individual's proficiency level of sexual knowledge. Another unique feature of IRT modelling is that a measurement error (due to the lack of precision in identifying a respondent's latent trait) is conditionally dependent on a latent trait level of the respondent (Lord, 1952). Moreover, unlike in CTT, item characteristics in IRT are not affected by sample characteristics (De Champlain, 2010), and likewise, individual latent trait estimates are not affected by particular items used to estimate them. This item/latent trait invariance property in combination with the IRT's focus on the items rather than the test as a whole (Lord, 1953) enables researchers to rank individuals on the same continuum of a certain underlying latent trait, even if the individuals have received different sets of items from a larger pool designed to measure the latent trait of interest. Using this IRT modelling approach, a test developer can create a reliable test customised to each individual. Customising a test to the individual's trait level, which is referred to as 'adaptive testing', cannot be easily accomplished within the CTT framework, but it is a natural extension of using IRT models (e.g. Embretson, 1996).

Computerized adaptive testing (CAT)

When working with a set of items that accurately measure a latent trait, another important consideration arises: How many items need to be administered to obtain an accurate

measurement of the latent trait? Many test banks are quite large and time-consuming especially if they are administered in concert with other tests. The ability to use only a subset of these items would help to reduce the response burden. Moreover, with a large test bank, many items often fail to effectively discriminate between respondents who are high in knowledge from those with lower levels of knowledge. In addition, respondents can be given items that contribute very little to discerning whether they possess the latent trait in question. These challenges can be addressed by computerised adaptive testing (CAT). In a standard (non-adaptive) testing format, all respondents begin by responding to a particular test item and then continue in an orderly sequence until they have responded to all of the test items. CAT, on the other hand, administers items in a dynamic and 'adaptive' fashion corresponding to the respondent's proficiency in sexual knowledge. In other words, the test items are administered successively based on what is known about the respondent's performance from their answers to the previous items that have been administered. After each response, the respondent's latent ability level is updated, and the next item is selected accordingly so that the respondent receives only items that comfort to their ability level, thereby avoiding items that are either too difficult or too easy. The CAT algorithm repeats this process until a prespecified termination criterion is reached.

In the current study, both IRT and CAT are utilised, representing an advance from a CTT method with the goal of optimising the scale reliability and reducing the response burden (Gershon, 2005; Wainer et al., 2000). IRT provides a means for test developers interested in assessing sexual knowledge to have a better understanding of each item's performance. IRT also makes it possible for us to rank individuals on a latent trait continuum, even though they respond to a different set of items. As a mode of test administration, CAT will enable test developers to find the optimal number of items, while avoiding fatigue among respondents at the same time.

Method

Participants

Prior to starting the study, ethical approval had been obtained for all protocols from the institutional review board (IRB) at the second author's institution (Protocol Code 2019–441). All the participants included in the study provided appropriate informed consent, which was only verbal given that written consent was waived by IRB due to the sensitive nature of the questions on the survey (i.e. sexual behaviour). The sample consisted of 1,291 college-age students attending a four-year university located in the southeastern portion of the US. Data were collected using online anonymous self-report surveys across four academic years: Fall 2019–Winter 2020 ($n = 245$), Fall 2020–Winter 2021 ($n = 97$), Fall 2021–Winter 2022 ($n = 576$), and Fall 2022 ($n = 373$). Given the pandemic, there were no data collected between the beginning of March 2020 and the end of Winter 2020. For the 2020–2021 academic year, classes were held in a hybrid mode (available both in person and online), but a majority of students attended classes exclusively online. For Fall 2021, Winter 2022, and Fall 2022, all students were required to be on campus. The average age of the sample was 18.89 years ($SD = 1.29$ years), 78% were female, <1% identified as non-binary or other gender. Almost one third (31.84%) of the sample identified as White or European American, 23.93% as Latino/a/x/Hispanic or Spanish origins, 16.96% as Black or African American, 12.24% as Asian, 1.86% as Middle Eastern or Northern African, 1.16% as Native Hawaiian or Pacific Islander, 1.01% as indigenous, 10.07% as multiple races, and .93% as other. When asked about their sexual orientation, 82.65% identified as heterosexual or straight, 10.38% as bisexual, 1.39% as lesbian, 1.16% as questioning, 0.85% as gay men, and 3.56% as other (e.g. demisexual, asexual, pansexual). More than half of the students (52.90%) reported being sexually active at the time of the study.

Item response theory (IRT)

The IRT parameters for SKAT-A items were estimated using 2-parameter (2-PL) and 3-parameter logistic models (3-PL). These two models are well suited for unidimensional scales consisting of true/false items (dichotomous) and yield the probability of a correct answer for an item as a function of the respondent's latent trait level and item properties. In the current study, a correct response (i.e. making a correct judgement about truthfulness of a statement) was coded as 1, and both an incorrect response and 'not sure' were coded as zero.

In the case of the 2-PL, the probability of a correct answer to an item is defined as a logit transformation of the linear equation $\omega = a(\theta - b)$, where a and b represent an item's discrimination (slope) and difficulty (location) parameters, respectively. The logit transformation brings the outcome ω on a probability scale, resulting in a typical s-shaped ogive curve (referred to as Item Characteristic Curve or ICC) that relates the probability of an item being endorsed plotted on the vertical y-axis against values of the underlying trait (θ) plotted on the horizontal x-axis. Individuals who have low levels of the trait (i.e. sexual knowledge) and consequently are less likely to respond correctly to an item fall to the left or negative side of the trait scale. On the other hand, individuals with higher levels of the trait are more likely to respond correctly to the item and fall to the right or positive side of the trait scale. Parameter a indicates how well an item discriminates between different θ levels and typically ranges from ~ 0.5 to ~ 3 . Higher a parameter leads to a faster increase in the probability of a correct response with increasing θ values (McDonald, 1985). In other words, for items with high discrimination, even small differences in θ values will lead to large differences in the probability of getting an item correct. Items with low discrimination, on the other hand, are problematic as they are not as informative regarding the underlying latent trait, and they can be eliminated or refined to better distinguish the respondent's ability.

The difficulty parameter b is expressed in the same metric as the respondent's latent trait parameter θ and is defined as the θ value at which there is a 50% probability to answer the item correctly. Parameter b typically ranges from ~ -3 to ~ 3 where higher values indicate more difficult items. If a respondent's θ parameter is higher than the item difficulty, the respondent has more than 50% probability that they will respond correctly to that item. The opposite is true when a trait level of the respondent is below the item difficulty. It is conceivable in various testing situations that the respondent can get a difficult item correct even if their θ level is very low, which is considered 'guessing' behaviour in an IRT framework. Guessing behaviour can be modelled by adding the guessing parameter denoted as c to the 2-PL model. The addition of a third parameter c results in the 3-PL model where c dictates the lower asymptote of the item's ICC (Birnbaum, 1968). The higher the guessing parameter for a particular item, the higher the probability of a correct answer for that item even for respondents with infinitely low ability levels.

In the current study, the fit of both 2-PL and 3-PL models was assessed using a log-likelihood difference test, given that the models are nested (i.e. the guessing parameter c is fixed to 0 in the 2-PL model, whereas it is freely estimated in the 3-PL model). Further, we evaluated the fit of the individual items using $S-X^2$ statistic (Orlando & Thissen, 2003) and compared test information functions between the two models (Chalmers, 2012). We conducted the IRT analyses using both the mirt package in R (Chalmers, 2012) and Mplus statistical software (B. O. Muthén & Muthén, 1998–2017). Because Mplus uses a scaling factor (~ 1.7), its results will be close to the R model findings. Any notable differences are due to the weighted least squares that Mplus uses to estimate model parameters (for computational speed), whereas R and other programmes use maximum likelihood estimation in their solution. Precise derivation of the equation for analysis of dichotomous variables in an IRT framework can be found in B. Muthén (1978) and Mplus Technical Report (Asparouhov & Muthén, 2020).

Computerized adaptive testing (CAT)

The CAT simulation was conducted using 'catlrt' package (Nydick, 2014) in R software. For each respondent in the dataset, all responses to the SKAT-A items that fitted the IRT model were used, and the items were selected and evaluated as if they were administered adaptively with a series of integrated decisions that address how initial and interim ability estimates will be calculated (1 - Start), how items will be selected and administered based on those estimates (2 - Continue), when the testing procedure will be terminated, and how the final ability estimate will be derived (3 - Stop) (van der Linden & Pashley, 2010).

Start. Within the starting phase of a CAT simulation, the initial ability level (θ) for each respondent was set to a logit of 0 (corresponding to the average value of the latent trait). If there is no information about the particular respondent available, the population mean (i.e. zero logit) is the most reasonable choice (Thissen & Mislevy, 2000). Subsequently, the three most informative items for the arbitrarily chosen initial level of θ were selected, and one of these items was randomly administered first. The observed respondent's response to the first administered item was then used to update θ using the expected a posteriori (EAP) estimation with a standard normal prior distribution.

Continue. For the updated θ estimate after the first administered item, the next item is selected from the pool of knowledge items using the unweighted Fisher information selection method. In other words, the item with the highest information function at the provisional θ point estimate was administered next. Based on the observed response to that item, the new θ estimate is calculated, again using the EAP estimator with normal prior, and another item is selected for the updated latent trait estimate.

Stop. The process of selecting an optimal item and updating the interim latent trait estimate based on the response to a selected item is repeated until a prespecified criterion is met. Given that the termination criterion employed was the required measurement precision, the simulated CAT administration continued until the standard error (SE) of the θ estimate for a particular respondent dropped below (a) $SE = .47$ and (b) $SE = .55$. These two SE values were selected because they represent CTT-based score reliability of (a) .78 and (b) .70, which, in turn, represent (a) the reliability estimate for the full SKAT-A (including the select items that fit the IRT model where Cronbach's $\alpha = .78$) and (b) the recommended minimal level of reliability for screening purposes, respectively.¹ As a result, each respondent may differ in the number of administered items in order to reach the prespecified measurement precision. If the measurement precision stopping rule was not satisfied with additionally administered items, the testing algorithm was terminated after the full test bank of SKAT-A items was administered. The performance of the CATs was evaluated with respect to (a) the number of administered items required to reach the termination criteria and (b) the association of CAT-estimated latent trait values (θ^{CAT}) with latent trait estimates based on the full SKAT-A ($\theta^{\text{SKAT-A}}$).

Results

Classic test theory (CTT)

Table 1 shows the results of the preliminary CTT analyses. This includes the mean (i.e. percentage of respondents answering the item correctly), standard deviation, correlation of each item with the total score, and Cronbach's coefficient alpha as a measure of scale score reliability. A low mean indicates that few respondents were able to answer the question correctly (or make a correct judgement about truthfulness of the statement). A low correlation of the item to the total score indicates that the item does a poor job of assessing the underlying trait. As shown, six items Sk7 (true): 'It is rare for a teenage boy to have a sexual encounter with another boy', SK11 (false): 'Many people dream at night about having sex with someone of the same sex', SK27 (true): 'More than half of all teenagers in America lose their virginity [have sex] by age 15', SK33 (true): 'When teenagers have sex [intercourse] for the first time, the majority of them use rubbers [condoms]', SK37 (true):

Table 1. Descriptive item information based on classical test theory (CTT).

Item	Content	<i>M</i>	<i>SD</i>	Corr ¹	Alpha ²
SK01	Orgasm-M	.27	.44	.22	.75
SK02	Orgasm-F	.64	.48	.39	.74
SK03	Orgasm-F	.70	.46	.37	.74
SK04	Masturbation	.96	.19	.38	.75
SK05	Sexual Performance	.72	.45	.33	.74
SK06	Sexual Activity	.33	.47	.19	.75
SK07	Sexual Orientation	.11	.32	-.01	.75
SK08	Orgasm-F	.72	.45	.42	.74
SK09	Sexual Violence	.58	.49	.30	.74
SK10	Masturbation	.86	.34	.49	.74
SK11	Sexual Orientation	.20	.40	.04	.75
SK12	Sexual Orientation	.86	.35	.29	.74
SK13	Sex Education	.19	.40	.13	.75
SK14	Sexual Violence	.50	.50	.20	.75
SK15	Contraception	.89	.31	.27	.75
SK16	Sexual Violence	.89	.31	.40	.74
SK17	Masturbation	.74	.44	.44	.74
SK18	Pregnancy	.53	.50	.33	.74
SK19	Contraception	.68	.47	.44	.74
SK20	STIs	.88	.32	.28	.75
SK21	Sexual Violence	.84	.37	.28	.74
SK22	Masturbation	.96	.19	.34	.75
SK23	Sexual Performance	.46	.50	.32	.74
SK24	Orgasm-F	.35	.48	.39	.74
SK25	Pregnancy	.60	.49	.38	.74
SK26	Orgasm-F	.68	.47	.45	.74
SK27	Virginity	.42	.49	.09	.75
SK28	Pregnancy	.38	.49	.16	.75
SK29	Pregnancy	.91	.28	.49	.74
SK30	Orgasm-F	.44	.50	.35	.74
SK31	STIs	.48	.50	.11	.75
SK32	Contraception	.83	.38	.27	.75
SK33	Contraception	.48	.50	.09	.75
SK34	Sexual Orientation	.13	.33	.16	.75
SK35	Abortion	.78	.41	.32	.74
SK36	Sexual Deviance	.29	.46	.16	.75
SK37	Sexual Performance	.21	.41	.08	.75
SK38	Sexual Orientation	.73	.44	.36	.74
SK39	Pregnancy	.33	.47	.09	.75
SK40	Abortion	.28	.45	.16	.75
SK41	Sex Education	.57	.50	.21	.75

Note: *M* = Male, *F* = Female, *SK* = Sexual knowledge. ¹Correlation of each item with the total score corrected for item overlap. Correlations in bold indicate inordinately low CTT-based item discrimination.

²Cronbach's alpha if an item is removed from the scale. Alphas in bold indicate an increase in scale reliability after removing an item.

'Men in their 30s have less interest in having sex compared to their interest when they were teenagers', and SK39 (false): 'The majority of girls who drop out of high school, drop out because they are pregnant' had very low mean scores, correspondingly low item-to-total correlations, and Cronbach's alphas increased in magnitude with removal of the item (using .75 as the critical threshold).

Dimensionality of the SKAT-A knowledge items

An assumption of the IRT procedure is that the trait being assessed is unidimensional. We tested this assumption using confirmatory factor analysis (CFA) in the Mplus statistical software (B. O. Muthén & Muthén, 1998–2017). In light of the dichotomous nature of the items (i.e. 0 = incorrect response or 'not sure', 1 = correct response), we used a Weighted Least Squares with Mean and Variance (WLSMV) adjustment, which involves a probit regression with a matrix of tetrachoric correlations with the

weighted least square parameter estimates from the diagonal of the weight matrix. By all indications, the CFA model with all 41 items produced an adequate fit, $\chi^2(779) = 1845.55, p < .0001$, comparative fit index (CFI) = .82, root-mean-square error of approximation (RMSEA) = 0.03 (90% CI [.031, .034]), standardised root mean residual (SRMR) = .08. With the exception of the SRMR, all of these model fit indices are well within the benchmark values indicating a reasonable fit (Hu & Bentler, 1999). The same six items identified in the CTT analyses were subsequently removed with a corresponding change in model fit, $\chi^2(785) = 1816.66, p < .0001$, CFI = .83, RMSEA = 0.03 (90% CI [.030, .034]), SRMR = 0.08. The DIFFTEST option available in the Mplus software yielded $\Delta\chi^2(6) = 15.35, p = .018$ between the model specifying all 41 items and the more restricted model with 35 items (constraining 6 factor loadings to zero). This suggests a better fit for the restricted model with 35 items and that the observed associations among the 35 items can be effectively accounted for by a single dimension assessing sexual knowledge (Hambelton et al., 1991).²

Item response theory (IRT)

With the CTT and CFA results in hand, we then tested the measurement precision of the 41 SKAT-A knowledge items using IRT. IRT replicated the findings of the CTT models and reinforced the six problematic items that were identified in the CTT analyses. Two items (SK7 and SK11) had negative discrimination parameters (denoted as 'a'; typical range from ~ 0.5 to ~ 3). For these items, the probability of a correct answer decreases with an increasing level of knowledge proficiency. Four additional items (SK27, SK33, SK37, and SK39) had very low discrimination parameters (.03, .00, .02, and .09, respectively). Moreover, four of the six items with negative or very low discrimination also had difficulty parameters that far exceeded the typical range of difficulty in practical testing situation (i.e. all six items had difficulty parameters denoted as 'b'; typical range from ~ -3 to ~ 3 that exceeded the acceptable benchmarks: $-8.34, -19.93, 6.20, 25.80, 35.47,$ and 4.76 , respectively). Negative scores indicate that the items are too easy, whereas large positive scores indicate that the items are too hard, even for respondents who have high levels of sexual knowledge. The six items with unacceptable parameter estimates were excluded in subsequent analyses, and the 2- and 3-PL models were fitted again using the remaining 35 items. Table 2 shows the results of the IRT analysis with the 35 SKAT-A knowledge items for both the 2-PL and 3-PL models.

A global nested comparison revealed that introducing 'guessing' parameters in the 3-PL model did not appreciably improve the model fit over the 2-PL model, $\chi^2(35) = 19.1, p = 0.99$. On the other hand, although many items displayed guessing parameters close to 0, there were several items for which guessing behaviour had a nontrivial likelihood – SK21 (false): 'When a child is raped or molested, it is usually done by a stranger', SK29 (false): 'A woman can only get pregnant if she has an orgasm during sex', SK31 (true): 'You can get a sexually transmitted disease through kissing a person who has a sexually transmitted disease', and SK40 (false): 'Most teenage girls who become pregnant will have an abortion'. Test information functions presented in Figure 1 indicated that by taking the guessing parameters into account in the 3-PL model, the SKAT-A provides more efficient testing for individuals with above-average knowledge proficiency, while it still provides similar testing efficiency for below-average proficiency when compared to the test information function based on the 2-PL model. Therefore, we consider the 3-PL model as a superior model for psychometric description of SKAT-A items.

Computerized adaptive testing (CAT)

Table 3 shows the results of post-hoc CAT simulation based on the observed responses by the respondents in the current study. The stopping rules included test reliability of .78 (i.e. standard error [SE] = .47), which is the CTT actual reliability of the instrument, as well as .70 ($SE = .55$), which is considered the minimum desirable level of scale score reliability. On average, only 19 ($SD = 9.2$) out of 35 SKAT-A items were administered in order to reach the measurement precision stopping rule of

Table 2. Item response theory (IRT) Model parameters.

Item	2 PL model			3 PL model			
	a	b	S-X ^{2*}	a	b	c	S-X ^{2*}
SK01	0.29	2.20	0.33	1.54	1.79	0.17	0.58
SK02	0.59	-0.73	0.40	1.07	-0.46	0.11	0.40
SK03	0.58	-1.06	0.33	1.04	-0.80	0.11	0.29
SK04	1.23	-2.30	0.73	1.95	-2.39	0.13	0.57
SK05	0.48	-1.32	0.43	0.81	-1.05	0.12	0.40
SK06	0.32	1.47	0.33	0.64	1.68	0.08	0.29
SK08	0.76	-0.95	0.79	1.38	-0.79	0.08	0.77
SK09	0.40	-0.55	0.82	0.68	-0.27	0.09	0.84
SK10	1.09	-1.49	0.84	2.11	-1.27	0.15	0.82
SK12	0.52	-2.31	0.98	0.89	-2.12	0.12	0.96
SK13	0.20	4.43	0.79	1.40	2.43	0.14	0.88
SK14	0.20	-0.04	0.40	0.40	1.01	0.17	0.40
SK15	0.42	-3.20	0.98	0.70	-2.03	0.14	0.96
SK16	0.77	-2.05	0.79	1.36	-1.91	0.12	0.70
SK17	0.81	-1.00	0.13	1.40	-0.88	0.07	0.14
SK18	0.50	-0.15	0.98	1.09	0.26	0.15	0.96
SK19	0.78	-0.76	0.33	1.75	-0.56	0.09	0.29
SK20	0.43	-3.01	0.98	0.74	-2.74	0.15	0.96
SK21	0.48	-2.31	0.79	0.88	-1.86	0.18	0.63
SK22	0.89	-2.65	0.33	1.45	-2.73	0.13	0.29
SK23	0.49	0.24	0.57	1.27	0.63	0.16	0.63
SK24	0.69	0.66	0.98	1.53	0.72	0.06	0.96
SK25	0.60	-0.50	0.79	1.27	-0.08	0.17	0.84
SK26	0.74	-0.80	0.61	1.37	-0.58	0.10	0.57
SK28	0.16	2.01	0.98	0.33	3.15	0.15	0.96
SK29	1.20	-1.76	0.36	2.39	-1.50	0.21	0.44
SK30	0.57	0.31	0.36	1.36	0.58	0.13	0.43
SK31	0.12	0.331	0.36	0.30	2.44	0.23	0.40
SK32	0.31	-3.21	0.33	0.49	-2.97	0.15	0.63
SK34	0.17	7.00	0.33	0.84	3.86	0.08	0.29
SK35	0.47	-1.83	0.84	0.82	-1.55	0.12	0.82
SK36	0.19	2.88	0.33	0.44	3.24	0.12	0.29
SK38	0.58	-1.21	0.33	1.05	-0.90	0.14	0.29
SK40	0.19	3.11	0.13	1.84	2.14	0.23	0.29
SK41	0.30	-0.63	0.46	0.50	-0.15	0.11	0.57

Note: 2-PL = 2-parameter logistic, 3-PL = 3-parameter logistic, a = discrimination, b = difficulty, c = guessing. * = *p* value for S-X² item goodness-of-fit fit statistic (Orlando & Thissen, 2003).

$SE = .47$ (reliability = .78). There were almost 9% of participants for which the pre-specified stopping rule was not satisfied and therefore all the 35 of the sexual knowledge items had to be administered. On the other hand, a little more than half (54.5%) of participants had to answer less than half of the items in the pool (i.e. fewer than 17 items), representing a considerable decrease in the response burden for most of the sample. The average number of administered items decreased further when the CAT stopping rule was set to $SE = 0.55$ (reliability = 0.70). This relatively low but widely accepted level of measurement precision led to 9.79 ($SD = 5.3$) administered items on average. Few participants (1%) had to answer all of the items in the pool for this stopping rule. Latent ability estimates based on CAT (θ^{CAT}) were strongly associated with latent trait estimates based on the full SKAT-A (θ^{SKAT-A}) for both stopping rules ($r = 0.98$, $SE = .47$; $r = 0.93$, $SE = .55$, respectively).

Discussion

This study demonstrated the use of Item Response Theory (IRT) modelling and Computerized Adaptive Testing (CAT) simulation to establish the psychometric properties of the 41-item sexual knowledge scale of the Sexual Knowledge and Attitudes Test – Adolescents (SKAT-A). To our knowledge, most of the psychometric work that has been conducted with scales assessing sexual knowledge has relied on Classical Test Theory (CTT) methods, focusing almost exclusively on

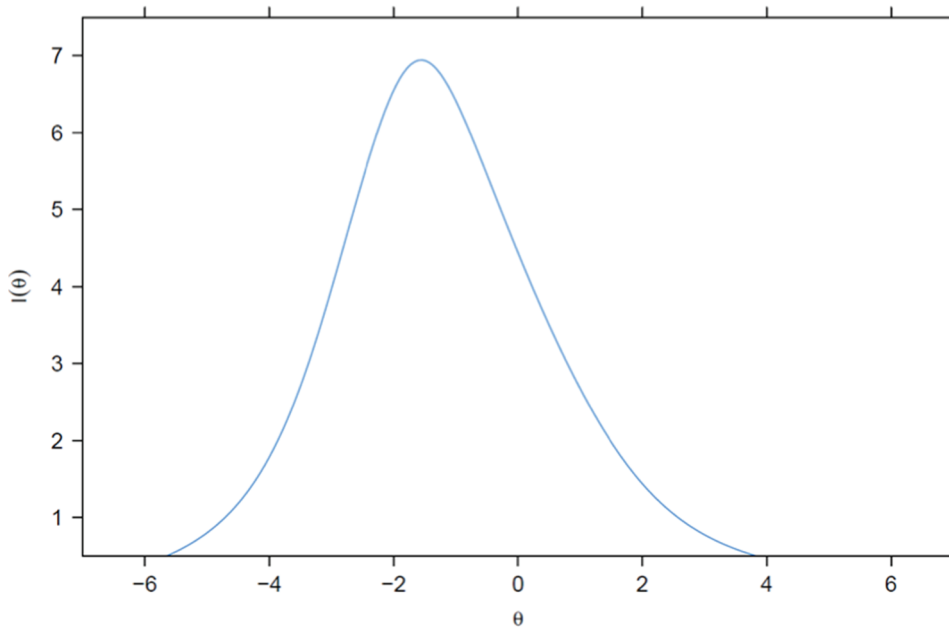
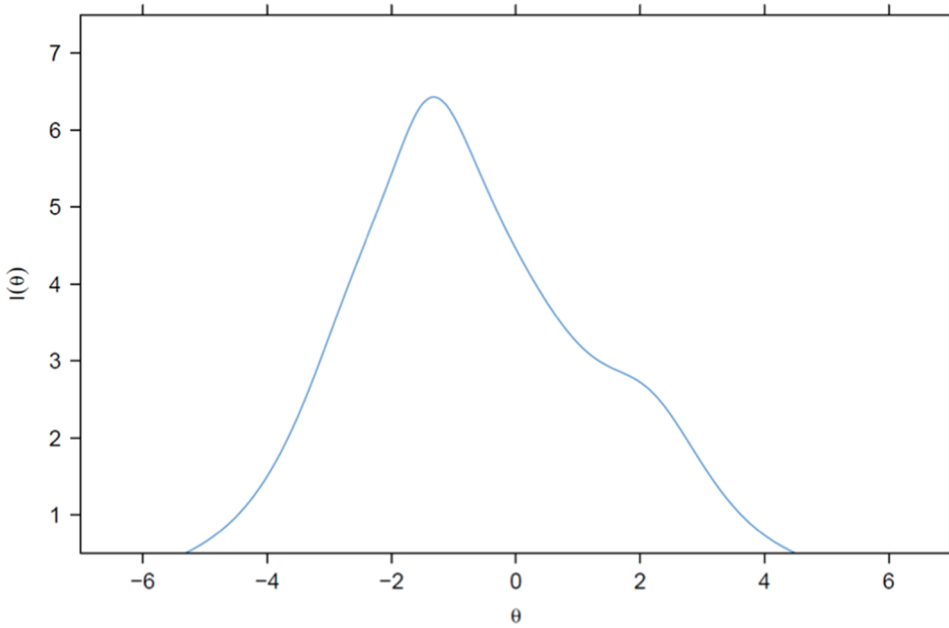
Test Information for 2-Parameter Logistic Model*Test Information for 3-Parameter Logistic Model*

Figure 1. Test information for 2-Parameter Logistic Model. Test information for 3-Parameter Logistic Model.

Table 3. Results of the CAT analyses.

Statistics	Reliability ¹	
	.78	.70
Mean of administered items	19.16	9.79
Standard deviation	9.2	5.3
Minimum	7	4
Maximum	35	35
% of participants who received all available items	8.5%	1.0%
% of participants who received less than a half of items	54.5%	94.4%
Correlation between proficiency based on CAT and proficiency based on all available items	0.98	0.93

Note: ¹Both numbers are the prespecified stopping criteria for the CAT simulation.

establishing the scale's dimensionality and reliability. IRT offers several improvements over CTT because it provides a means to achieve better scale construction by detailing the performance of an item with respect to an underlying trait, eliminating items that fail to discriminate the underlying trait well, and further eliminating redundant items. The information provided by IRT models extends beyond what CTT provides and shows the likelihood of endorsing an item at a given level of sexual knowledge, and how well the item functions to place an individual at some point on the latent trait continuum.

Confirmatory Factor Analysis (CFA) provided evidence that the SKAT-A assesses a underlying unidimensional trait of sexual knowledge. The one-factor model fit well, and alternative model specifications did not improve the fit. Models of this nature can always be tightened through the addition of correlated residuals or removing poorly fitting items. Finding a perfect fit was not the intended goal in the current study, and there is evidence from simulation studies that added residual correlations may not be stable with relatively small samples (MacCallum, 1986). In addition, the lack of a perfect fit in the current study is not problematic as it is well known that IRT models are robust against moderate violations of unidimensionality (Hambelton & Cook, 1983).

Results from both CTT (i.e. means and alpha) and IRT (i.e. discrimination and difficulty parameters) indicated that six of the knowledge items could be removed, producing a more streamlined 35-item scale (Nandakumar, 1991). Furthermore, the IRT models indicated a better fitting model for the 3-PL model, suggesting that greater reliability (i.e. ability or proficiency estimation) can be achieved when controlling for guessing. Many of the items eliminated in the IRT procedure reflect nebulous questions. For instance, one item that was eliminated asked whether the sex drive of men wanes after age 30 ('men in their 30s tend to have less interest in having sex compared to their interest when they were teenagers'), which would be difficult for women to answer accurately, let alone college-age men who have not reached this age and may not have experienced a diminution of their sex drive to speak of. The other items that were eliminated also suggest that knowledge items need to address factual knowledge rather than experience-dependent, subjective knowledge (e.g. abortion, same-sex encounters, age of sexual debut), which can vary considerably across individuals and sociocultural contexts.

The CAT simulation demonstrated that the 35 SKAT-A items provide a realistic test bank from which a select number of items can be drawn to efficiently assess a unidimensional and broadly defined latent trait tapping sexual knowledge. On average, only 19 of the 35 items had to be administered to achieve a precise score estimation (reflecting a reliable estimation of sexual knowledge) given the prespecified stopping rule. This number was further reduced to 9.79 if the reliability of the scale was set to .70. Importantly, the CAT administration also showed an abbreviated version (i.e. fewer than 35 items) captured the essence of the full 35-item version, as demonstrated by the high correlation between the two versions. This finding underscores that CAT offers various cost efficiencies and testing advantages by retaining the items that most efficiently assess the underlying proficiency in sexual knowledge. This aligns with the goals of CAT to tailor the sequence and administration of items from a test bank based on the

respondent's ability level. This approach minimises the need for extensive pilot testing of new items and ensures comprehensive content coverage while maintaining the existing highly relevant and reliable items.

Study limitations and future directions

There are several limitations associated with this study worth noting. First, the SKAT-A does not reflect several contemporary themes in sex education courses including, but not limited to, gender identity and expression, sexual violence, and the intersection of sexuality and online media (i.e. online dating, sexually explicit online content). Although no scale can remain timeless, updating the SKAT-A knowledge scale to incorporate these different content areas may be a critical next step in the evolution of scale development.

Second, the diversity of the sample could come into question given that it consisted of predominantly female young adults attending college. Women may possess more accurate and wider knowledge regarding their own reproductive and sexual physiology and anatomy and other sexuality-related topics such as STIs, contraception, and female orgasm. There is indeed some evidence for greater sexual knowledge among women than men, although it seems to slightly vary by topic and culture (Lou et al., 2012; Lyu et al., 2020; Rahman et al., 2011; Synovitz et al., 2002). This gender difference in sexual knowledge can be attributed to differences in sex communication: Female adolescents and young adults, compared to their male counterparts, engage in more in-depth conversations about sex with parents and friends (R. Evans et al., 2019; Lefkowitz & Espinosa-Hernandez, 2007; Trinh & Ward, 2016). Additional studies are warranted that employ differential item functioning analyses to test for gender differences in item thresholds. It is worth considering, however, that items may perform differently for different genders because of their respective different physiological and sexual awareness and experience.

Third, the sample was obtained from a single university in the southeastern portion of the US. This is not likely to represent all young adults in college and certainly provides limited information for non-college bound young adults. However, one of the unique features of IRT is that it is not sample specific, thus overcoming some of the sampling frame issues that plague other psychometric procedures, although additional studies are still warranted to determine the external validity of these findings.

Fourth, although not necessarily a limitation in the current study, we did not establish the temporal stability of knowledge over time, relying only on a snapshot of what young adults know about various sexual topics. There are several factors that may influence test-retest reliability. One of them is the fact that individuals can learn from their previous mistakes, which can affect the stability of their knowledge. For example, after initial test administration, an individual can independently seek information, which may include engaging discussions with friends that fill in the gap regarding their factual knowledge of sexuality-related topics. This behaviour would undermine the temporal stability of knowledge. Given empirical evidence suggesting that young adults increase their sexual knowledge during the college years (Franklin & Dotger, 2011), it is important for future studies to examine the extent of temporal stability of sexual knowledge and what can contribute to an increase (if any) over time (e.g. sexual experiences, sexual communication with friends).

Finally, in the current study, we set out only to establish the utility of IRT and CAT procedures applied to the 41 SKAT-A knowledge items. However, as the field becomes more advanced, and sexual knowledge scales increasingly find their way into programme evaluation, real scores relevant to levels of proficiency in sexual knowledge will become essential tools. The next step to further refine the SKAT-A knowledge scale is to develop a procedure that helps us to transform proficiency or 'trait' scores to a more meaningful format for actual testing purposes (e.g. T-scores = 50 + [logit x 10]).

Conclusion

The current study filled the gap in the sexual knowledge literature by utilising IRT and CAT to psychometrically refine the SKAT-A knowledge scale with young adults. Given the ability to assess a broad, general proficiency in sexual knowledge rather than knowledge in a specific domain of sexuality (e.g. STI/HIV, contraception), the SKAT-A knowledge scale can be used to evaluate sex education programmes targeting adolescents and young adults that cover a wide range of sex-related topics. However, more work is still needed in order to determine whether 'narrowness of content' (Loevinger, 1954) can affect what we know about the psychometric properties of the existing sexual knowledge instruments and whether domain-specific or domain-general instruments exhibit any differences in predictive power for later sexual behaviour. The need for comprehensive sex education targeting adolescents and young adults is greater than ever given the heightened prevalence of STIs within these age groups (CDC, 2021). Further applications of IRT and CAT methods to various sexual knowledge instruments can ensure scale refinement, reduce respondent burden, and ultimately contribute to effective evaluations of sex education programmes.

Notes

1. Although no hard and fast rule exists with respect to what constitutes adequate reliability, it is generally accepted that over .75 is reasonable to establish the reliability of a scale. There are a number of citations in the field of psychometrics that suggest this number (e.g. Nunnally & Bernstein, 1994).
2. Tests of a bifactor model and other alternative model configurations did not improve on the basic finding of essential unidimensionality. In addition, an exploratory factor analysis with one to five factor solutions using a Geomin (oblique) rotation with the full set of 41 items showed a scree plot with one eigenvalue (representing the first factor) considerably larger than the others (8.36, 2.72, 2.18, 1.78, 1.56, respectively for one to five factor solutions). The same outcome was obtained with the reduced set of 35 items (8.32, 2.14, 1.75, 1.56, 1.44).

Data availability statement

The data that support the findings of this study are available from the corresponding author, AS, upon reasonable request.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Aya Shigeto  <http://orcid.org/0000-0002-3127-3162>

References

- Allen, L. (2001). Closing sex education's knowledge/practice gap: The reconceptualisation of young people's sexual knowledge. *Sex Education, 1*(2), 109–122. <https://doi.org/10.1080/14681810120052542>
- Arnold, E. M., Smith, T. E., Harrison, D. F., & Springer, D. W. (2000). Adolescents' knowledge and beliefs about pregnancy: The impact of "ENABL". *Adolescence, 35*(139), 485–498.
- Asparouhov, T., & Muthén, B. O. (2020). *IRT in Mplus, Version 4*. <https://www.statmodel.com/download/MplusIRT.pdf>
- Bennett, S. E., & Assefi, N. P. (2005). School-based teenage pregnancy prevention programs: A systematic review of randomized controlled trials. *Journal of Adolescent Health, 36*(1), 72–81. <https://doi.org/10.1016/j.jadohealth.2003.11.097>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392–479). Addison-Wesley.

- Borawski, E. A., Trapl, E. S., Lovegreen, L. D., Golabrianchi, N., & Block, T. (2005). Effectiveness of abstinence-only intervention in middle school teens. *American Journal of Health Behavior, 29*(5), 423–434. <https://doi.org/10.5993/AJHB.29.5.5>
- Borgia, B., Marinacci, C., Schifano, B., & Perucci, C. A. (2005). Is peer education the best approach for HIV prevention in school? Findings from a randomized controlled trial. *Journal of Adolescent Health, 36*(6), 508–516. <https://doi.org/10.1016/j.jadohealth.2004.03.005>
- Bruce, K., & McLaughlin, J. (1986). The development of scales to assess knowledge and attitudes about genital herpes. *The Journal of Sex Research, 22*(1), 73–84. <https://doi.org/10.1080/00224498609551290>
- Bruce, K. E. M., & Bullins, C. G. (1989). Students' attitudes and knowledge about genital herpes. *Journal of Sex Education and Therapy, 15*(4), 257–270. <https://doi.org/10.1080/01614576.1989.11074968>
- Butts, S. A., Kayukwa, A., Langlie, J., Rodriguez, V. J., Alcaide, M. L., Chitalu, N., Weiss, S. M., & Jones, D. L. (2017). HIV knowledge and risk among Zambian adolescent and younger adolescent girls: Challenges and solutions. *Sex Education, 18*(1), 1–13. <https://doi.org/10.1080/14681811.2017.1370368>
- Carey, M. P., Morrison-Beedy, D., & Johnson, B. T. (1997). The HIV-Knowledge questionnaire: Development and evaluation of a reliable, valid, and practical self-administered questionnaire. *AIDS and Behavior, 1*(1), 61–74. <https://doi.org/10.1023/A:1026218005943>
- Carey, M. P., & Schroder, K. E. E. (2002). Development and psychometric evaluation of the brief HIV knowledge questionnaire. *AIDS Education and Prevention, 14*(2), 172–182. <https://doi.org/10.1521/aeap.14.2.172.23902>
- Centers for Disease Control and Prevention. (2021). *Sexually Transmitted Infections Prevalence, Incidence, and Cost Estimates in the United States*. <https://www.cdc.gov/std/statistics/prevalence-2020-at-a-glance.htm>
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Compton, W. M., Saha, T. D., Conway, K. P., & Grant, B. F. (2009). The role of cannabis use within a dimensional approach to cannabis use disorders. *Drug and Alcohol Dependence, 100*(3), 221–227. <https://doi.org/10.1016/j.drugalcdep.2008.10.009>
- Condelli, L. (2011). Contraceptive Utilities, intention, and knowledge scale. In T. D. Fisher, C. M. Davis, W. L. Yarber, & S. L. Davis (Eds.), *Handbook of sexuality-related measures* (3rd ed., pp. 180–185). Routledge.
- Cook, K. F., Teal, C. R., Bjorner, J. B., Cella, D., Chang, C. H., Crane, P. K., Gibbons, L. E., Hays, R. D., McHorney, C. A., Ocepke-Welikson, K., Raczek, A. E., Teresi, J. A., & Reeve, B. B. (2007). IRT health outcomes data analysis project: An overview and summary. *Quality of Life Research, 16*(Suppl 1), 121–132. <https://doi.org/10.1007/s11136-007-9177-5>
- Coyle, K., Anderson, P., Laris, B. A., Barrett, M., Unti, T., & Baumler, E. (2021). A group randomized trial evaluating high school FLASH, a comprehensive sexual health curriculum. *Journal of Adolescent Health, 68*, 686–695. <https://doi.org/10.1016/j.jadohealth.2020.12.005>
- Davis, C., Noel, M. B., Chan, S.-F.-F., & Wing, L. S. (1998). Knowledge, attitudes and behaviours related to HIV and AIDS among Chinese adolescents in Hong Kong. *Journal of Adolescence, 21*(6), 657–665. <https://doi.org/10.1006/jado.1998.0186>
- De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*(1), 109–117. <https://doi.org/10.1111/j.1365-2923.2009.03425.x>
- DeGroot, S., Vogelaers, D., Liefhooghe, G., Vermeir, P., & Vandijck, D. M. (2014). Sexual experience and HIV-related knowledge among Belgian university students: A questionnaire study. *BMC Research Notes, 7*, 299. <https://doi.org/10.1186/1756-0500-7-299>
- Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341–349. <https://doi.org/10.1037/1040-3590.8.4.341>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.
- Evans, K. R., Sills, T., DeBrot, D., Gelwicks, S., Englehardt, N., & Santor, D. (2004). An item response analysis of the Hamilton Depression Rating Scale using shared data from two pharmaceutical companies. *Journal of Psychiatric Research, 38*(3), 275–284. <https://doi.org/10.1016/j.jpsychores.2003.11.003>
- Evans, R., Widman, L., Kamke, K., & Stewart, J. L. (2019). Gender differences in parents' communication with their adolescent children about sexual risk and sex-positive topics. *Journal of Sex Research, 57*(2), 177–188. <https://doi.org/10.1080/00224499.2019.1661345>
- Fennie, T., & Laas, A. (2014). HIV/AIDS-related knowledge, attitudes and risky sexual behaviour among a sample of South African university students. *Gender and Behaviour, 12*(1), 6035–6044.
- Franklin, R. M., & Dotger, S. (2011). Sex education knowledge differences between freshmen and senior college undergraduates. *College Student Journal, 45*(1), 199–213.
- Fullard, W., Johnston, D. A., & Lief, H. I. (1998). The sexual knowledge and attitude test for adolescents. In C. M. Davis, W. L. Yarber, R. Bauserman, G. Schreer, & S. L. Davis (Eds.), *Handbook of sexuality related measures* (pp. 33–35). Sage Publications.
- Fullard, W., & Scheier, L. (2011). The sexual knowledge and attitude test for adolescents. In T. D. Fisher, C. M. Davis, W. L. Yarber, & S. L. Davis (Eds.), *Handbook of sexuality-related measures* (3rd ed., pp. 16–18). Taylor & Francis.
- Future of Sex Education Initiative. (2020). *National sex education standards: Core content and skills, K-12* (2nd ed.). <https://www.advocatesforyouth.org/wp-content/uploads/2021/11/NSES-2020-web-updated2.pdf>

- Gelhorn, H., Hartman, C., Sakai, J., Stallings, M., Young, S., Rhee, S. H., Corley, R., Hewitt, J., Hopfer, C., & Crowley, T. (2008). Toward DSM-V: An item response theory analysis of the diagnostic process for DSM-IV alcohol abuse and dependence in adolescents. *The Journal of the American Academy of Child & Adolescent Psychiatry*, 47(11), 1329–1339. <https://doi.org/10.1097/CHI.0b013e318184ff2e>
- Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement*, 6(1), 109–127.
- Goldfarb, E. S., & Lieberman, L. D. (2021). Three decades of research: The case for comprehensive sex education. *Journal of Adolescent Health*, 68, 13–27. <https://doi.org/10.1016/j.jadohealth.2020.07.036>
- Guzzo, K. B., & Hayford, S. R. (2018). Adolescent reproductive and contraceptive knowledge and attitudes and adult contraceptive behavior. *Maternal and Child Health*, 22(1), 32–40. <https://doi.org/10.1007/s10995-017-2351-7>
- Hambelton, R. K., & Cook, L. L. (1983). The robustness of item response theory models and effects of length and sample size on precision of ability estimation. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31–49). Academic Press.
- Hambelton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, 38(9 Suppl), 1128–1142. <https://doi.org/10.1097/00005650-200009002-00007>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Huang, J., Bova, C., Fennie, K. P., Rogers, A., & Williams, A. B. (2005). Knowledge, attitudes, behaviors, and perceptions of risk related to HIV/AIDS among Chinese university students in Hunan, China. *AIDS Patient Care and STDs*, 19(11), 769–777. <https://doi.org/10.1089/apc.2005.19.769>
- Jaworski, B. C., & Carey, M. P. (2007). Development and psychometric evaluation of a self-administered questionnaire to measure knowledge of sexually transmitted diseases. *AIDS and Behavior*, 11(4), 557–574. <https://doi.org/10.1007/s10461-006-9168-5>
- Kelly, J. A., St Lawrence, J. S., Hood, H. V., & Brasfield, T. L. (1989). An objective test of AIDS risk behavior knowledge: Scale development, validation and norms. *Journal of Behavior Therapy and Experimental Psychiatry*, 20(3), 227–234. [https://doi.org/10.1016/0005-7916\(89\)90027-X](https://doi.org/10.1016/0005-7916(89)90027-X)
- Kirby, D. (2007). Abstinence, sex, and STD/HIV education programs for teens: Their impact on sexual behavior, pregnancy, and sexually transmitted disease. *Annual Review of Sex Research*, 18(1), 143–177. <https://doi.org/10.1080/10532528.2007.10559850>
- Kirby, D. B., Laris, B. A., & Rollieri, L. A. (2007). Sex and HIV education programs: Their impact on sexual behaviors of young people throughout the world. *Journal of Adolescent Health*, 40, 206–217. <https://doi.org/10.1016/j.jadohealth.2006.11.143>
- Kohler, P. K., Manhart, L. E., & Lafferty, W. E. (2008). Abstinence-only and comprehensive sex education and the initiation of sexual activity and teen pregnancy. *Journal of Adolescent Health*, 42(4), 344–351. <https://doi.org/10.1016/j.jadohealth.2007.08.026>
- Kumar, R., Goyal, A., Singh, P., Bhardwaj, A., Mittal, A., & Yadav, S. S. (2017). Knowledge attitude and perception of sex education among school going adolescents in Ambala District, Haryana, India: A cross-sectional study. *Journal of Clinical and Diagnostic Research*, 11(3), LC01–LC04. <https://doi.org/10.7860/JCDR/2017/19290.9338>
- Kutner, B. A., Perry, N. S., Stout, C., Pala, A. N., Paredes, C. D., & Nelson, K. M. (2022). The inventory of anal sex knowledge (iASK): A new measure of sexual health knowledge among adolescent sexual minority males. *The Journal of Sexual Medicine*, 19(3), 521–528. <https://doi.org/10.1016/j.jsxm.2021.12.011>
- Lal, S. S., Vasan, R. S., Sarma, P. S., & Thapnkappan, K. R. (2000). Knowledge and attitude of college students in Kerala towards HIV/AIDS, sexually transmitted diseases and sexuality. *The National Medical Journal of India*, 13(5), 231–236.
- Lefkowitz, E. S., & Espinosa-Hernandez, G. (2007). Sex-related communication with mothers and close friends during the transition to university. *Journal of Sex Research*, 44(1), 17–27. <https://doi.org/10.1080/00224490709336789>
- Lief, H. I., Fullard, W., & Devlin, S. J. (1990). A new measure of adolescent sexuality: SKAT-A. *Journal of Sex Education and Therapy*, 16(2), 79–91. <https://doi.org/10.1080/01614576.1990.11074980>
- Lightfoot, A. F., Taboada, A., Taggart, T., Tran, T., & Burtaine, A. (2015). “I learned to be okay with talking about sex and safety”: Assessing the efficacy of a theatre-based HIV prevention approach for adolescents in North Carolina. *Sex Education*, 15(4), 348–363. <https://doi.org/10.1080/14681811.2015.1025947>
- Lindberg, L. D., & Maddow-Zimet, I. (2012). Consequences of sex education on teen and young adult sexual behaviors and outcomes. *Journal of Adolescent Health*, 51(4), 332–338. <https://doi.org/10.1016/j.jadohealth.2011.12.028>
- Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, 51(5), 493–504. <https://doi.org/10.1037/h0058543>
- Lord, F. M. (1952). *A theory of test scores*. Psychometric Monographs.
- Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13(4), 517–548. <https://doi.org/10.1177/001316445301300401>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

- Lou, C., Cheng, Y., Gao, E., Zuo, X., Emerson, M. R., & Zabin, L. S. (2012). Media's contribution to sexual knowledge, attitudes, and behaviors for adolescents and young adults in three Asian cities. *Journal of Adolescent Health, 50*(3), S26–S36. <https://doi.org/10.1016/j.jadohealth.2011.12.009>
- Lyu, J., Shen, X., & Hasketh, T. (2020). Sexual knowledge, attitudes and behaviours among undergraduate students in China—implications for sex education. *International Journal of Environmental Research and Public Health, 17*, 6716. <https://doi.org/10.3390/ijerph17186716>
- MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100*(1), 107–120. <https://doi.org/10.1037/0033-2909.100.1.107>
- Mackin, M. L., & Perkhounkova, Y. (2019). Development of the test of adolescent sexual knowledge based on the national sexuality education standards and results of pilot testing. *American Journal of Sexuality Education, 14*(2), 212–232. <https://doi.org/10.1080/15546128.2018.1548990>
- Maticka-Tyndale, E., & Barnett, J. P. (2010). Peer-led interventions to reduce HIV risk of youth: A review. *Evaluation and Program Planning, 33*(2), 98–112. <https://doi.org/10.1016/j.evalprogplan.2009.07.001>
- McCabe, M. P., & Cummins, R. A. (1996). The sexual knowledge, experience, feelings and needs of people with mild intellectual disability. *Education and Training in Mental Retardation and Developmental Disabilities, 31*(1), 13–21.
- McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*(1), 28–50. <https://doi.org/10.1177/1088868310366253>
- McDonald, R. P. (1985). *Factor analysis and related methods*. Psychology Press. <https://doi.org/10.4324/9781315802510>
- Motedayen, M., Kalantarkousheh, S. M., Scheier, L. M., & Komarc, M. (2019). Psychometric validation of the sexual knowledge and attitudes test –adolescents (SKAT-A) in an Iranian sample. *Cogent Psychology, 6*(1), 1585505. <https://doi.org/10.1080/23311908.2019.1585505>
- Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551–560. <https://doi.org/10.1007/BF02293813>
- Muthén, B. O., & Muthén, L. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.
- Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*(2), 99–117. <https://doi.org/10.1111/j.1745-3984.1991.tb00347.x>
- Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Nydick, S. W. (2014). Catlirt: An R package for simulating IRT-based computerized adaptive tests. R package version 0.5-0. <https://CRAN.R-project.org/package=catlirt>
- Opt, S. K., & Loffredo, D. A. (2004). College students and HIV/AIDS: More insights on knowledge, testing, and sexual practices. *The Journal of Psychology, 138*(5), 389–402. <https://doi.org/10.3200/JRLP.138.5.389-403>
- Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X²: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289–298. <https://doi.org/10.1177/0146621603027004004>
- Rahman, A. A., Rahman, R. A., Ibrahim, M. I., Salleh, H., Ismail, S. B., Ali, S. H., Wan Muda, W. M., Ishak, M., & Ahmad, A. (2011). Knowledge of sexual and reproductive health among adolescents attending school in Kelantan, Malaysia. *The Southeast Asian Journal of Tropical Medicine and Public Health, 42*(3), 717–725.
- Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 19*(4), 345–359. <https://doi.org/10.1037/1040-3590.10.4.345>
- Sanz-Martos, S., López-Medina, I. M., Álvarez-García, C., & Álvarez-Nieto, C. (2019). Sexuality and contraceptive knowledge in university students: Instrument development and psychometric analysis using item response theory. *Reproductive Health, 16*, 127. <https://doi.org/10.1186/s12978-019-0791-9>
- Schaalma, H. P., Abraham, C., Gillmore, M. R., & Kok, G. (2004). Sex education as health promotion: What does it take? *Archives of Sexual Behavior, 33*(3), 259–269. <https://doi.org/10.1023/B:ASEB.0000026625.65171.1d>
- Sills, T., Wunderlich, G., Pyke, R., Segraves, R. T., Leiblum, S., & Clayton, Cotton, D., Evans K. (2005). The sexual interest and desire inventory – female (SIDI-F): Item response analysis of data from women diagnosed with hypoactive sexual desire disorder. *The Journal of Sexual Medicine, 2*(6), 801–818. <https://doi.org/10.1111/j.1743-6109.2005.00146.x>
- Singh, S., Bankole, A., & Woog, V. (2005). Evaluating the need for sex education in developing countries: Sexual behavior, knowledge of preventing sexually transmitted infections/HIV and unplanned pregnancy. *Sex Education, 5*(4), 307–331. <https://doi.org/10.1080/14681810500278089>
- Synovitz, L., Herbert, E., Kelley, R. M., & Carlson, G. (2002). Sexual knowledge of college students in a southern state: Relationship to sexuality education. *American Journal of Health Studies, 17*(4), 163–172.
- Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101–134). Lawrence Erlbaum.
- Trinh, S. L., & Ward, L. M. (2016). The nature and impact of gendered patterns of peer sexual communications among heterosexual emerging adults. *Journal of Sex Research, 53*(3), 298–308. <https://doi.org/10.1080/00224499.2015.1015715>
- van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). Springer.

- Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum. <https://doi.org/10.4324/9781410605931>
- Walter, H. J., & Vaughan, R. D. (1993). AIDS risk reduction among a multiethnic sample of urban high school students. *Journal of the American Medical Association*, 270(6), 725–730. <https://doi.org/10.1001/jama.1993.03510060071035>
- Wang, B., Meier, A., Shah, I., & Li, X. (2006). The impact of a community-based comprehensive sex education program on Chinese adolescents' sex-related knowledge and attitudes. *Journal of HIV/AIDS Prevention in Children and Youth*, 7(2), 43–64. https://doi.org/10.1300/J499v07n02_04
- Wong, L. P. (2012). An exploration of knowledge, attitudes, and behaviours of young multiethnic Muslim-majority society in Malaysia in relation to reproductive and premarital sexual practices. *BMC Public Health*, 12, 865. <https://doi.org/10.1186/1471-2458-12-865>
- Wong, T., Pharr, J. R., Bungum, T., Coughenour, C., & Lough, N. L. (2019). Effects of peer sexual health education on college campuses: A systematic review. *Health Promotion Practice*, 20(5), 652–666. <https://doi.org/10.1177/1524839918794632>
- Yip, P. S. F., Zhang, H., Lam, T.-H., Lam, K. F., Lee, A. M., Chan, J., & Fari, S. (2013). Sex knowledge, attitudes, and high-risk sexual behaviors among unmarried youth in Hong Kong. *BMC Public Health*, 13, 691. <https://doi.org/10.1186/1471-2458-13-691>
- Yoo, H., Lee, S. H., Kwon, B. E., Chung, S., & Kim, S. (2005). HIV/AIDS knowledge, attitudes, related behaviors, and sources of information among Korean adolescents. *Journal of School Health*, 75(10), 393–399. <https://doi.org/10.1111/j.1746-1561.2005.tb06643.x>
- Zhao, R., Zhang, L., & Fu, X. X. (2019). Sexual and reproductive health knowledge, attitude, and behavior among senior high school and college students in 11 provinces and municipalities of China. *Chinese Journal of Public Health*, 35(10), 1330–1338. <https://doi.org/10.11847/zgggws1124531>