Published in final edited form as:

Psychol Assess. 2016 February ; 28(2): 194–204. doi:10.1037/pas0000175.

Evaluation of Internal Validity Using Modern Test Theory: Application to Word Association

Yusuke Shono, **Susan L. Ames**, and **Alan W. Stacy**

School of Community and Global Health, Claremont Graduate University

Abstract

Word association tests (WATs) have been widely used to examine associative/semantic memory structures and shown to be relevant to behavior and its underpinnings. Despite successful applications of WATs in diverse research areas, few studies have examined psychometric properties of these tests or other open-ended cognitive tests of common use. Modern test theory models, such as item response theory (IRT) models, are well suited to evaluate interpretations of this class of test. In this evaluation, unidimensional IRT models were fitted to the data on the WAT designed to capture associative memory relevant to an important applied issue: casual sex in a sample of 1138 adult drug offenders. Using association instructions, participants were instructed to generate the first behavior or action that came to mind in response to cues (e.g., “hotel/motel”) that might elicit casual sex-related responses. Results indicate a multitude of evidence for the internal validity of WAT score interpretations. All WAT items measured a single latent trait of casual sex-related associative memory, strongly related to the latent trait, and were invariant across gender, ethnicity, age groups, and sex partner profiles. The WAT was highly informative at average-to-high levels of the latent trait and also associated with risky sex behavior, demonstrating the usefulness of this class of test. The study illustrates the utility of the assessments in this at-risk population as well as the benefits of application of the modern test theory models in the evaluation of internal validity of open-ended cognitive test score interpretation.

Keywords

word association test; associative memory; at risk populations; item response theory; validity

Recent years have seen burgeoning research efforts to determine the interplay between health behaviors and preexisting associative memory structures, as revealed through tests that have been documented to assess associative memory (Stacy & Wiers, 2010). Though diverse models of memory are available to help explain this interplay, traditional associative network models (e.g., Anderson, 1983; Wickelgren, 1976) are perhaps the most well known across research areas. Associative network models posit that concepts are represented in an associative memory network as nodes, which are interconnected among one another through

Correspondence concerning this manuscript should be addressed to: Yusuke Shono, School of Community and Global Health, Claremont Graduate University, 675 W. Foothill Blvd., Suite 310, Claremont, CA 91711, USA; yusuke.shono@cgu.edu.
Yusuke Shono, School of Community and Global Health, Claremont Graduate University.
Susan L. Ames, School of Community and Global Health, Claremont Graduate University.
Alan W. Stacy, School of Community and Global Health, Claremont Graduate University

links with varying levels of associative strength. Processing a concept (e.g., a visual word "rain") *spontaneously* activates other related concepts, and this memory activation pattern, or the associative memory structure, is assumed to reflect one's repetitive experiences in everyday life. Thus, the concepts that tend to co-occur in the same context (e.g., "rain" and "umbrella") are strongly connected to each other in memory (Nelson, McKinney, Gee, & Janczura, 1998).

Word Association Tests

Word association tests (WATs), which employ either free association or controlled association instructions, are among the most widely used procedures that capture the pattern and strength of association among concepts in long-term memory (Nelson, McEvoy, & Schreiber, 2004; Steyvers, Shiffrin, & Nelson, 2005). A typical WAT, using free association instructions, involves a visual or auditory presentation of a cue (e.g., "rain") to which participants are instructed to produce the first word or phrase that comes to mind (e.g., "umbrella"). Instructions are indirect in the sense that they focus on top of mind, spontaneous responses, rather than recollection of a previous event, judgments or inferences about concept relations or subjective probability. In WATs designed to elicit a target response related to a particular behavior, indirect instructions also imply that the behavior is not an explicit focus of the assessment. First responses generated by participants in WATs using indirect instructions are thought to represent the strongest associates for a given cue (Nelson, McEvoy, & Dennis, 2000) though variants of the test have shown that subsequent responses can be quite informative (Szalay & Deese, 1978). In cognitive research, WATs have continually been used to infer implicit processes in experimental paradigms that assess associative responses (e.g., Levy, Stark, & Squire, 2004; Shimamura & Squire, 1984). These assessments also have provided an effective index of association strength in memory used in association norms, which have been found to be among the strongest predictors of cognitive effects across diverse procedures such as illusory memory (Roediger, Watson, McDermott, & Gallo, 2001), semantic priming (Hutchison, Balota, Cortese, & Watson, 2008), and extralist cued-recall (Nelson, McKinney, Gee, & Janczura, 1998). No other open or closed ended assessment has shown this degree of widespread utility in revealing associations in memory at both individual and normative levels.

A variety of studies have demonstrated that WATs are also relevant to behavior, its underpinnings, and applied research. As some examples, WATs have been used effectively in research on cross-cultural comparisons (Szalay, Strohl, Fu, & Lao, 1994), repression-sensitization traits (Galbraith & Lieberman, 1972), depression (Alison & Burgess, 2003; Watkins, Vache, Verney, & Mathews, 1996), attitude toward food (Rozin, Kurzer, & Cohen, 2002), and other health related topics, such as alcohol, tobacco, and other drug use (Ames, et al., 2007; Kelly, Masterman, & Marlatt, 2005; Kelly, Haynes, & Marlatt, 2008) and HIV risk (Ames, Grenard, & Stacy, 2013). WATs have been found to be good predictors of various health behaviors studied to date, even when more traditional predictors have been controlled for in the analysis (see Stacy & Wiers, 2010, for review). Thus, WATs are frequently used to predict behavior and to infer underlying associations in memory; cognitive research also has shown strong value as just outlined. When associations in memory are a focus of theory, prediction models, or intervention, then WATs provide one of the few assessment options with previous empirical support across domains.

Though WATs have been successfully applied in basic and applied research, few studies have reported comprehensive psychometric properties of these tests. The only exception to our knowledge is a recent study on alcohol- and marijuana-related WATs (Shono, Grenard, Ames, & Stacy, 2014). The results are in line with a recent meta-analytic study showing that the WATs are, among other tests of associative memory, some of the most useful predictors of substance use behavior (Rooke, Hine, & Thorsteinsson, 2008). In other research areas across cognitive, social, and behavioral domains, including work on HIV risk, the dearth of prior comprehensive psychometric investigations of WATs is clear. With HIV risk in particular, there is a more general lack of application of thoroughgoing attempts to understand the full psychometric profile of tests of cognitive mediators of HIV risk behavior. The current study attempts to fill this void by illustrating comprehensive psychometric modeling from modern test theory and internal validity perspectives to evaluate the interpretations of the WAT as a test of associative memory structure. Although traditional classical test theory (CTT) approaches have provided useful information about cognitive processes and behavior in this behavior domain (Czopp, Monteith, Zimmerman, & Lynam, 2004; Galbraith, 1968), several advantages of modern test theory over CTT have long been noted in the literature (e.g., Hambleton & Jones, 1993; Reise, Ainsworth, & Haviland, 2005). Thus, the present study utilizes an item response theory (IRT) framework to fully understand the psychometric characteristics and evaluate the internal validity of casual sex-related WAT score interpretation.

Validity Theory

Though various theoretical perspectives on validity have been developed (e.g., Borsboom, Mellenbergh, & van Heerden, 2004; Cronbach & Meehl, 1995; Embretson, 1983; Loevinger, 1957; McArdle & Prescott, 1992; Messick, 1989), the present study adopted a dual account of validity (Grimm & Widaman, 2012; McArdle & Prescott, 1992), in which validity is distinguished between internal and external aspects[1]. Internal validity, which is a main focus of the current investigation, is concerned with item-level psychometric questions, such as item content, the relationship among items, the relationship of items to a latent construct, and the difficulty of items (Grimm & Widaman, 2012). External validity, on the other hand, addresses the relation of test scores of interest and those of other measures, seeking criterion-related evidence, convergent and divergent evidence, evidence of change in construct, and positive and negative consequences of test score interpretations (Grimm & Widaman, 2012). The present study evaluates the various components of the internal validity of WAT score interpretation, using IRT and related methods.

Item Response Theory

IRT is particularly well suited for evaluation of validity of test score interpretation at the item-level (Grimm & Widaman, 2012). IRT provides a series of nonlinear relationships

[1] Note that the concepts of internal and external validity in this context are applied to a property of *test score interpretations*. Though similar, they are not isomorphic with the internal and external validity of research design (Campbell and Stanley, 1963).

between observed item responses and a latent trait of interest, with an assumption that one's item response is influenced by the underlying latent trait (Embretson & Reise, 2000). In the present study, we assumed that a latent trait represented casual sex-related associative memory. A higher level of the latent casual sex-related associative memory can be defined as an associative memory network wherein a greater number of concepts are interconnected with relative strength to a casual sex-related concept. Following the associative memory account of health behavior (Stacy, 1997; Stacy, Ames, Ullman, Zogg, & Leigh, 2006), we further assumed that participants' levels of the latent casual sex-related associative memory would lead to variations in their item response patterns on the WAT. Hence, those who possess higher levels of the latent casual sex associative memory would be likely to exhibit different item response patterns in the WAT (e.g., more endorsement of casual sex-related responses) as compared to those with lower levels.

The present study employed a unidimensional two-parameter logistic model (2PLM; Birnbaum, 1968) because the WAT was designed to assess a single latent factor, and all WAT items were dichotomous responses (i.e., 1 = related to casual sex, 0 = not related to casual sex). In the 2PLM, the probability that participants endorse a casual sex-related response to a given WAT item is expressed as:

$$\mathrm{P}(x_i{=}1|\theta){=}\frac{1}{1{+}\exp[-Da_i(\theta-b_i)]},$$

where x_i is the observed response for an item *i*, θ is a latent attribute of interest (e.g., casual sex-related associative memory), *a* is an item discrimination parameter, *b* is an item difficulty parameter, and *D* is a scaling factor of 1.7. The item discrimination parameter describes how closely an item is related to θ, and thus a higher *a* is desirable. The item difficulty parameter defines the level of θ at which approximately half of the participants endorse an item. It indicates that the higher the *b* value a WAT item has, the higher level of the latent associative memory would be required for a sex-related response to be endorsed. Located on the same continuum as θ, the difficulty parameter and θ share the same metrics, which usually has a mean of 0 and a standard deviation of 1. In IRT, a participant's level of a latent attribute is estimated by using his or her response pattern for a set of items, along with item parameter estimates (Embretson & Reise, 2000). More specifically, IRT scoring takes into account not only how many items participants endorse, but also which items they endorse. For example, two participants who endorsed five sex-related responses on the WAT may have different estimates of the latent trait scores since, for example, items with higher discrimination parameter values contribute more to the latent trait score (Edwards, 2009). In contrast, all items are weighted equally to a total score in CTT.

IRT also permits estimations of reliability with great flexibility. Reliability in IRT can be obtained at *any point* on a latent trait continuum at both item and test levels. In CTT, on the other hand, a reliability index (e.g., coefficient α) is a fixed estimate for a set of test items. The present study reports detailed information on the precision of both WAT items and WAT as a test in relation to the underlying latent casual sex-related associative memory.

Moreover, IRT-based tests of differential item functioning (DIF) provides some advantages over CTT-based tests, including interpretation (e.g., sample-independent characteristics) and evaluation (e.g., use of graphical representation) of DIF (Edwards & Edelen, 2009). DIF is said to occur when an item parameter estimate is not equivalent across subgroups (e.g., male and female) of a study sample after adjusting for overall differences between subgroups on a latent trait. Detection of DIF items is essential since DIF items could be a potential threat to validity of test score interpretation (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005). The IRT-based DIF test, with use of the linking procedures, provides another attractive methodological advantage. IRT can analyze simultaneously items from multiple studies or items from different tests designed to measure the same construct by aggregating raw data from multiple studies or multiple tests (Curran & Hussong, 2009). In the present study, we pooled raw data from two independent studies on cognitive processes in HIV risk behaviors. Pooling increases the power for evaluating DIF across groups and can make subgroup comparisons possible. Subgroups in the present study included the study samples (Study 1 vs. Study 2), gender (male vs. female), ethnicities (Hispanic vs. Non-Hispanic White), age groups (emerging adult vs. early adult vs. middle adult), and sex partner profiles (main partner only vs. casual or exchange partner only), demonstrating the flexibility and utility of IRT evaluation of DIF. The present work illustrates that these comprehensive procedures are quite useful for investigation of internal validity even when the items originate in an open-ended, top of mind form.

Current Study

The samples selected in the present study are highly relevant to the focus on casual sex in the WAT and the illustration of IRT in cognitive domains of risk behavior. The samples were drawn from a population of adult drug offenders known to have high levels of casual sex, unprotected sex, and drug use (Leigh, Ames, & Stacy, 2008). Assessment in this population can be challenging, but the protocol has been developed and refined over time to make rigorous psychometric evaluation feasible. Comprehensive evaluation is important so that advances can be made in assessments, relevant to future work on both theory and evidence-based interventions. The present study illustrates the utility of the assessments in this at-risk population as well as the benefits of application of the modern test theory models in the evaluation of evidence of internal validity.

Following Grimm and Widaman (2012), we examined various components of internal validity, including underlying dimensionality, detailed item properties, reliability at item and test levels across the latent trait continuum, DIF across various subgroups and latent trait scores across groups. Moreover, we also examined one component of external validity: criterion-related evidence through the association between casual sex-related associative memory and HIV risk behavior. As the past studies have shown the predictive utility of the WAT on HIV risk behaviors (e.g., Stacy, Ames, Ullman, Zogg, & Leigh, 2006), it was expected that higher levels of casual sex-related associative memory would be associated with more risky HIV behavior. Though a main focus of the current investigation was the evaluation of the internal validity of the WAT score interpretation, examination of the criterion-related validity would be informative from both psychometric and substantive perspectives.

Method

Participants

Data for the current study were drawn from two studies on cognitive processes in drug abuse and HIV risk behaviors. Participants in both studies were adult drug offenders who were enrolled in various drug diversion programs in Southern California. Study 1, conducted during the period between 2009 and 2010, included 485 participants (female = 131). Participants in Study 2, which was part of a large-scale longitudinal study being conducted between 2011 and 2012, were 666 adult drug offenders (female = 186). Of 666, 13 participants were dropped from the analyses due to an excess number of missing trials on the word association test (WAT; see the measures section for more detail). Thus, the total number for the current study was 1138 participants (female = 315), ranging in age from 18 to 55 (*M* = 31.44, *SD* = 10.81).

Measures

Word association test (WAT)—The casual sex-related WAT is an indirect memory test, designed to measure preexisting structures of associative memory relevant to casual sex concepts. Twenty-four target cue words or phrases were selected based on a pilot study where 107 participants produced the first word that came to mind when they thought of 46 potential cues that might elicit sex-related responses. Half of the 24 target cues consisted of a single word or phrase related to affective outcomes of sexual behaviors (e.g., "excitement"), locations (e.g., "hotel/motel") or situations (e.g., "hanging out"), and the other half were compound cues consisting of a combination of location/situation and affective outcome (e.g., "At a motel, feels pleasant", "Friday night, having fun"), or situation and two affective outcomes, some of which could be related to drug use (e.g., "Friday night, pleasant feelings, getting high"). On each trial, a cue appeared in the center of a computer display and participants were asked to type the first behavior or action that came to mind as quickly as they could. Participants typed their responses in a text box presented right below the cue in the center of the display, and the next trial began after they typed their response or after 21 seconds had elapsed since the onset of the cue, whichever came first. In addition to the 24 target cues, there were 10 filler cues not related to either sex or drug use. After participants completed all WAT items, responses were scored using a validated self-coding procedure (Frigon & Krank, 2009; Krank, Schoenfeld, & Frigon, 2010). In this procedure, participants were provided each WAT cue and their own typed response, in conjunction with a list of behavior-related categories. They were asked to select which category from the list most closely matched each of their responses. The categories included casual sex, safe sex, sex with main partner, alcohol, marijuana, cocaine, methamphetamine, other drugs, exercise, food, sleeping, playing sports, and ’none of the above.’ For the sake of the present analysis, responses selected for the casual-sex category were coded 1 and other responses were coded 0. Recent studies[2] have shown that self-coded WAT scores can be more accurate indicators

[2] The current study had four raters independently code WAT responses in terms of casual sex concepts in a random sample of 150 participants selected from Study 2. Inter-rater agreement among the four raters, as measured by Light’s kappa (1971), was excellent (kappa = .92). Unit weight total WAT scores were created based on the scores coded by each rater. These four independent total scores were strongly related to the total WAT scores calculated from self-coding (*r*s ranged from .51 to .53). The results were comparable to those reported in previous studies (e.g., Krank et al., 2010).

of memory association and were more predictive of behavior than rater-coded scores since participants likely know their intended meaning of their responses, some of which could be ambiguous to raters (Krank et al., 2010). The predictive utility of WAT self-coded scores has been demonstrated in studies on risk behavior and WAT (e.g., Ames et al., 2013).

HIV risk behaviors—*Multiple sex partners* were assessed with three items using a 7-point scale ranging from 0 (none) to 6 (six or more; coefficient alpha= .87). Participants were instructed to indicate the number of partners that they had in the past 4 months for casual sex, a one-night stand, and sex on a first meeting. *Sex partner profiles* were assessed with three items that asked participants to estimate separately the number of main, casual, and exchange sex partners they had during the past year. Based on responses to these items, participants were categorized into three groups: those who had sex with only main partners, those who had sex with only casual or exchange partners, or others.

Procedures

Participants in both Study 1 and 2 were recruited from drug diversion programs. They were provided with the consent information verbally and were informed that their participation in the study was completely voluntary and confidential and that their responses were fully protected by a Certificate of Confidentiality from the National Institutes of Health. Those who agreed to participate were invited to a mobile computer laboratory that was set up in a room provided by a drug diversion program facility. A trained, experienced data collector from Claremont Graduate University led each data collection session and provided introductory instructions on how to complete the computerized assessment. Further instructions were provided on the computer. The computer-based assessments were administered in small groups, with a maximum size of 15. Participants received $15 in exchange for their participation at the end of the assessment.

Data Analysis Plan

IRT assumptions: Unidimensionality and local independence—First, we conducted categorical confirmatory factor analysis (CCFA), using Mplus 6.11 (Muthen & Muthen, 2011), to evaluate whether the casual sex WAT satisfied the unidimensionality assumption. A one-factor CCFA model was fit to the aggregated data such that all items would load onto a single latent construct of associative memory relevant to casual sex. Weighted least-squares with mean and variance adjustment (WLSMV; Muthen, du Toit, & Spisic, 1997) was used as an estimator to accommodate binary WAT variables (Flora & Curran, 2004). Model fit was evaluated according to the guidelines suggested by Hu and Bentler (1999). In relation to the unidimensionality assumption, we also assessed local dependence (LD) in all item pairs, using two criteria: (a) modification indices (MI) of residual covariances from the CCFA model described above and (b) the local dependence X^2(LD X^2; Chen & Thissen, 1997), obtained from an item calibration with a two-parameter logistic model (2PLM). Item pairs with LD X^2 values of 10 or greater and/or MI values of 20 or greater were flagged as possible instances of LD, and similarity of item contents was examined to ensure the conceptual plausibility of these instances. Presence of potential LD is observed when there are residual correlations among two or more items after adjusting for a single latent trait (Thissen & Steinberg, 2009). LD brings negative influences to various aspects of IRT modeling, including biased estimates of item parameters, latent trait scores, and latent trait dimensions (Embretson & Reise, 2000).

Item calibration—Following the evaluation of the unidimensionality of the casual sex WAT items, we estimated both one-parameter (1PLM) and two parameter (2PLM) logistic models, with marginal maximum likelihood estimation using flexMIRT (Cai, 2012) to evaluate item properties. Item difficulty parameters were estimated freely for all items in both 1PLM and 2PLM. In contrast, discrimination parameters were estimated freely in 2PLM whereas they were estimated with equivalence constraints across all items in 1PLM. A likelihood ratio test was used to compare the fit of the two models. The parameter estimates obtained from this step were then used in the subsequent DIF analyses.

DIF and latent trait means—A series of DIF analyses were conducted to examine if the initial item parameter estimates were invariant across different subgroups of the study sample after controlling for the levels of a latent trait. More specifically, we sought to examine if each WAT item functioned equivalently in terms of the *a* and b parameters across studies (Study 1 vs. Study 2), genders (male vs. female), ethnicities (Hispanic vs. Non-Hispanic White), age groups (emerging adult vs. early adult vs. middle adult), and sex partner profiles (main partner only vs. casual or exchange partners only). For a comparison of each subgroup, we first conducted two-step Wald tests (Langer, 2008; Woods, Cai, & Wang, 2013) to identify potential WAT items that could serve as unbiased anchor items in subsequent one-step Wald tests (Woods et al., 2013). It is necessary to find anchor items to link the groups in comparison (e.g., Study 1 vs. Study 2) so that the parameters to be estimated for each group can be placed onto a common scale (Embretson & Reise, 2000). In the first step of the two-step Wald test, the mean (*M*) and standard deviation (*SD*) of the focal group (e.g., Study 2) were freely estimated while 1) *M* and *SD* of the reference group (e.g., Study 1) were fixed to 0 and 1, respectively, and 2) item parameters were constrained to be equal between groups. In the second step, all item parameters were estimated freely for both groups while applying the *M* and *SD* obtained in step 1 for the focal group (the reference group still had the *M* and *SD* of 0 and 1, respectively). After identifying anchor items, we performed one-step Wald tests (Woods et al., 2013) to determine whether or not candidate items (i.e., the items not identified as anchors) exhibited DIF. In the one-step Wald tests, 2PLMs for each group (e.g., Study 1 and Study 2) were simultaneously fit to the data with the following specification: 1) the *M* and *SD* of the focal group were estimated freely while the *M* and *SD* of the reference group were fixed to 0 and 1, respectively, and 2) the item parameters of the candidate items were estimated freely in each group while those of the anchor items were constrained equal across groups. A Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) was used to control for a false discovery rate in both the one-step and two-step Wald tests. Following the DIF tests, latent trait mean scores were compared across studies, genders, ethnicities, age groups, and sex partner profiles by estimating multiple group 2PLMs wherein the item parameters of non-DIF items were constrained to be equal across groups and those of DIF items were free to vary.

Final item calibration and IRT scoring—Alternative sets of WAT items were evaluated by estimating a 2PLM. Latent trait scores were calculated, using the *expected a posteriori* (EAP) method.

Associations with criterion measures—We examined the association between the estimated latent trait scores on the WAT and the multiple sex partner scale scores. Using a latent variable modeling (LVM) approach, a point and interval estimate of the correlation coefficient was obtained using Mplus 6.11 (Muthen & Muthen, 2011).

Results

CTT-based Summary Statistics

The second and third columns of Table 1 display, respectively, corrected total-item correlations as measured by point-biserial correlation coefficients (r_{pb}) and the mean item endorsement (p-value). The mean item endorsement related to casual sex varied from 6% to 31% across 24 items[3], which is common especially when word cues have many potential associates (Nelson, McEvoy, & Schreiber, 2004). All items showed moderate-to-strong item-total correlations (r_{pb} range [.37, .62]), indicating that for each item, those who endorsed greater numbers of casual sex-related responses across trials tended to endorse a casual sex-related response to the item. Total WAT scores, observed score-based indices of the strength of memory associations related to casual sex, were computed for each participant by summing the number of casual sex-related responses across 24 items. Higher scores indicate a stronger degree of memory association. The distribution of the total scores was skewed positively, as is typical with this type of open-ended item format (range [0, 23]; M = 3.33, SD = 4.46; median = 1.00; skewness = 1.71).

IRT Analysis

IRT assumptions—The fit of the CCFA model was rejectable, X^2 (df = 252) = 735.79, p < .001, likely due to the large sample size; however, the practical fit indices indicated a good model fit, with Tucker-Lewis Index (TLI) = .954, Comparative fit index (CFI) = .958, and RMSEA = .041, 90% CI [.038, .045]. Factor loadings of 24 WAT items ranged from .59 to . 83 and they were all significant (p < .01). Although these results may be seen as a support of the unidimensionality assumption, subsequent tests of the local independence (LI) assumption suggested that several items should be examined for potential local dependence (LD). Specifically, relatively large values of MI (> 20.0) and/or LD X^2 (> 10.0) were obtained, respectively, from the CCFA and IRT (2PLM) for the following item pairs[4]: (a) "At a motel, feels good" and "At a motel, feels pleasant" (MI = 67.8, LD X^2 = 15.3), (b) "Friday night, feeling good, using drugs" and "Friday night, pleasant feelings, getting high" (MI = 107.5, LD X^2 =25.1), (c) "Friday night, feeling good, drinking" and "Friday night,

[3] A cut-off criterion for an extremely low proportion of item endorsement was set at 2% (e.g., Curran, et al., 2008), which would likely cause convergence issues in subsequent IRT parameter estimations. Although not investigated in this article, other WAT responses with more than minimal endorsement rates across items were alcohol use (6% to 60%), cannabis use (6% to 38%), and methamphetamine use (3% to 33%).

[4] Besides the five item pairs, two other item pairs had MI > 20.0, but LD X^2 < 10.0. However, inspection of the item contents revealed less lexical similarity between the items (“At a bar, feels pleasant” vs. “Friday night, feeling good, drinking”; “feeling relaxed” vs. “having fun”). Thus, these items were retained.

having fun, getting drunk" (MI = 24.41, LD X^2 = 2.8), (d) “Friday night, having fun, getting drunk” and “Friday night, feeling good, using drugs” (MI = 28.66, LD X^2 = 9.1) and (e) "Friend’s House/Apartment" and "With Friends" (MI = 23.87, LD X^2 = 3.0). Following examination of the item contents, we set aside from the analyses the items, "At a motel, feels good," "Friday night, feeling good, drinking," "Friday night, feeling good, using drugs," and "with friends."

With 20 WAT items, CCFA results showed a good model fit, TLI = .978, CFI = .980, RMSEA = .029, 90% CI [.024, .033] though the exact fit of the model was rejectable, X^2 (*df* = 170) = 328.14, *p* < .01. No item pairs exceeded the cut-off scores for MI and LD X^2, except that the two item pairs that had MI values over 20.0 in the initial CCFA but were not excluded from the current model again showed MI values greater than 20.0 (26.8 and 22.5).

Initial item calibration—After confirming the unidimensionality of the casual sex WAT, we estimated 1PLM and 2PLM using 20 WAT items to evaluate the item properties. Both models revealed a good fit with RMSEA = 0.03 and no indication of item misfit as judged by item fit indices (*S-*X^2; Orlando & Thissen, 2000) after controlling for the false discovery rate using the Benjamini-Hochberg procedures (see Table 1). In terms of the item parameter estimates, the discrimination parameter ranged from 1.33 to 3.02 in 2PLM whereas it was 2.00 across the items in 1PLM (see Table 1). The ranges of the difficulty parameters were estimated from 0.83 to 2.11 in 2PLM and from 0.72 to 2.07 in 1PLM. A likelihood ratio test revealed that 2PLM provided a significantly better fit than 1PLM, G^2 (*df* = 19) = 81.85, *p* < .001.

DIF and latent trait means—DIF tests were conducted to examine if the WAT items functioned equivalently across subgroups, including study samples (Study 1 vs. Study 2), gender (male vs. female), ethnicity (Hispanic vs. Non-Hispanic White), age groups (emerging adults vs. early adults vs. middle adults), and sex partner profiles (main partner only vs. casual/exchange partners). Table 2 presents the results of the one-step Wald test (Cai, Thissen, & du Toit, 2011) for the candidate items identified from the two-step Wald test. Adjusting for the false discovery rate with the Benjamini-Hochberg procedures, the overall Wald test revealed that significant DIF was detected only for the item "With friends, feels pleasant" between males and females, X^2 (2) = 7.3, *p* = .003. The subsequent individual tests showed that the discrimination parameter of this item differed significantly between genders, X^2 (1) = 7.2, *p* = .007, indicating that, after matching male and female participants on the latent trait, the item was more strongly related to the latent trait for males (*a* = 2.37) than for females (*a* = 0.93). Moreover, the difficulty parameter of this item was estimated with an extremely large standard error value for females (*b* = 3.57, *SE* = 1.14) though no evidence of item misfit for the item was found, *S-*X^2(2) = 5.3, *p* = 0.07. Thus, this item was set aside from the final calibration of the casual sex-related WAT (see below).

Table 3 reports estimates of latent trait means and observed means[5] by subgroups. In each of the subgroups, the latent mean was fixed to 0 for the reference groups (Study 1, male, Hispanic, early adulthood, and main sex partner only). In a multiple group 2PLM that compared males and females, item parameters for all but one item that was found to show DIF were constrained to be equal across groups. Results showed that the latent trait scores were lower for females than males ($Z = -7.50$, $p < .01$). For the comparison of the three age groups (emerging, early, and middle adults), middle adults, as compared with early adults, scored significantly lower on the latent casual sex associative memory ($Z = -2.00$, $p < .01$). No difference was found between emerging adults and early adults ($Z = -.50$, *ns*). For the comparison of sex partner profiles, the latent trait mean scores were higher for those who had sex with either casual or exchange partners than those who had sex with only main partners in the past year ($Z = 4.50$, $p < .01$). The latent trait means did not differ between Study 1 and Study 2 ($Z = -.50$, *ns*) and between Hispanic and Non-Hispanic Whites ($Z = 0.86$, *ns*).

Final item calibration and IRT scoring—After excluding one item with DIF across genders, we estimated another 2PLM with a total of 19 WAT items. The model fit the data well (RMSEA = .003) and all items had good item fit. The last four columns of Table 1 display the item parameters and item fit indices for the 19 WAT items. All the items were strongly related to the latent casual sex associative memory ($a \geq 1.35$), indicating that the WAT items differentiated effectively among the participants with different trait levels.

Table 4 presents the item and test information functions (IIF and TIF, respectively) as a function of various levels of the underlying latent associative memory relevant to casual sex. Almost all the WAT items were most informative at the moderate-to-high levels of the latent trait ($0.5 < \theta < 2.0$). TIF and its corresponding reliability estimates presented at the bottom two lines of Table 4 show that the casual sex WAT with the 19 items was most reliable at the range of θ between 0 and 2.0.

Criterion-related Validity

As expected, the latent associative memory relevant to casual sex was correlated with risky sex behaviors. The adult drug offenders who had higher levels of casual sex associative memory tended to have greater numbers of casual sex partners ($r = .35$, 95% CI [.30, .41]).

Discussion

As a test of preexisting associative memory structures, WATs have been quite useful in previous research across several domains of cognitive research, as well as research on culture, traits, health behavior, and clinically-relevant symptoms. The present study evaluated the internal validity and criterion-related validity of interpretations of WAT scores designed to assess preexisting associative memory structures relevant to casual sex concepts,

[5] Similar patterns of results were obtained between the latent trait and observed scores largely because the current WAT did not include DIF items across subgroups. Note, however, that the distribution of the latent trait scores appeared closer to a normal distribution, relative to that of the observed scores (e.g., skewness was .54 and 1.6, for latent trait and observed scores, respectively). Thus, the latent trait scores were used in the subsequent analysis of criterion-related validity (see texts).

in a sample of adult drug offenders. Applying IRT and related psychometric approaches, we obtained evidence for both internal validity and criterion-related validity.

Validity Evidence

The results from the CCFA and IRT analysis confirmed that all the WAT items measured a single latent factor of casual sex-related associative memory, and no item pairs exhibited local dependency in the revised 19-item WAT. An examination of the estimated item parameters of the two-parameter logistic model (2PLM) revealed that all these items were strongly related to the underlying latent casual sex-related associative memory, and moderate-to-high levels of the latent trait were likely to be required for adult drug offenders to endorse casual sex-related responses. These results were consistent with those of a previous study, which demonstrated unidimensionality and good item properties of WATs for alcohol and marijuana use among at-risk adolescents (Shono et al., 2014). Moreover, the casual sex-related WAT exhibited evidence for content validity. As noted earlier, the WAT items represented different content domains relevant to casual sex concepts, including affective outcomes, situations and locations. These items have been shown to elicit sex-related WAT responses among those who tended to engage in risky sex behaviors (Ames et al., 2013; Grenard, Ames, & Stacy, 2013).

The present investigation also revealed that all the items in the revised WAT functioned equivalently across study groups, gender, ethnicity, age groups, and sex partner profiles, after controlling for the overall differences on the latent associative memory. The findings of the invariant WAT items not only provided additional internal validity evidence, but also indicated that potential adverse impacts on WAT score interpretations (e.g., test score bias for members of a certain subgroup) were likely minimized. Though not a main focus of the current investigation, the latter can be considered as evidence of external validity (Grimm & Widaman, 2012) or a consequential aspect of validity (Messick, 1989). Further support for external validity was observed by the finding that the estimated latent trait scores on the casual sex-related WAT were correlated with the scores on the casual sex partners scale. The result provided further evidence that sex-related associative memory, as measured through WAT, is related to HIV risk behavior (Ames et al., 2013, Stacy, Newcomb, & Ames, 2000).

Lastly, we note that the WAT in the current investigation was administered under *indirect* test instructions. In response to a WAT cue (e.g., "hotel/motel"), participants generated the very first behavior or action that came to mind without given any reference to a target behavior. Hence, observed responses presumably reflected the underlying cognitive processes of spontaneous retrieval of memory associations. Drawing on these substantive grounds, coupled with the obtained psychometric evidence discussed above, we argue that the casual sex-related WAT is a plausible test for casual sex-related associative memory.

Reliability

As noted in the introduction to this article, one of the advantages of IRT over CTT is that reliability can be estimated at any point on the latent trait continuum. The present study demonstrated that all the WAT items were most reliable at moderate-to-high levels of the latent casual sex-related associative memory (see Table 1). Similarly, at the test level, the casual sex WAT was most accurate at these levels of the latent trait, providing total information of over 10, which is equivalent to a traditional reliability estimate of .90. The WAT also had good reliability of .80 at around the mean level of the latent trait. In contrast, little information was provided on the other end of the latent trait continuum ($\theta < -1.0$). Thus, it can be concluded that the casual sex-related WAT revealed good-to-excellent reliabilities in the range between the mean and high level of the latent casual sex-related associative memory. We argue that the observed effectiveness of the WAT in the limited range of the latent trait ($\theta > 0$) does not undermine the internal validity of the WAT score interpretations. In the IRT-based psychometric literature, it is not uncommon to observe such a narrow coverage of a latent trait especially when a latent trait of interest is a *unipolar* or *"quasi-trait"* construct such as depression and anxiety (Reise & Waller, 2009). According to Reise and Waller (2009), because the lower end of a latent unipolar trait indicates an absence of the trait (e.g., no depression, no anxiety), a scale that is accurate at the lower end of the unipolar latent trait is not as informative as the one that is accurate in the central or higher region of the continuum. This appears to be the case for the latent casual sex-related associative memory. Furthermore, the WAT instructions were indirect, did not focus on any particular target behavior, and had the goal of activating one's memory associations *without* requiring awareness of what was tested (i.e., sex-related concepts). Thus, it may be difficult, if not impossible, to find a set of cues that would elicit casual sex-related responses *very easily* for individuals with lower levels of casual sex-related associative memory without an explicit reference to casual sex behavior. Therefore, we argue that the casual sex-related WAT demonstrated a good range of coverage of the latent trait.

Comparisons of Latent Trait Means

The present study also compared the estimates of the latent trait scores across subgroups. In a novel finding, middle adult drug offenders had a lower latent mean than both emerging and early adult drug offenders, and the latter two groups did not differ in the latent means. These results may, in part, be explained by developmental changes in the quality and status of relationships (Augustus-Horvath & Tylka, 2011) that might have an impact on the associative memory structures relevant to the casual sex concepts. Since older adults are more likely than younger adults to engage in stable relationships, they might less often generate casual sex-related responses to given WAT cues. Another possible explanation is that age differences in the latent casual sex-related associative memory more or less reflect developmental changes in memory systems (Ofen & Shing, 2013). Future research should examine the extent to which these factors potentially related to developmental changes contribute to performance on WAT. The results also showed that male drug offenders had a higher latent trait score than their female counterparts. This finding corroborated that of Stacy et al. (2000), in which males responded with more sex-related responses than females did on a version of the WAT consisting of cues with ambiguous meanings (i.e., homographs) that could be associated with sexual behaviors (e.g., “climax”). Finally, the mean latent trait scores did not differ between drug offenders who were in Study 1 and those who were in Study 2. Although these two independent studies were administered a year apart, the results indicated the equivalence of the two pooled samples. Equivalence over time and sample suggests that history and selection effects (Cook and Campbell, 1979) were likely not operational over this period in the WAT despite the different samples and time periods. Overall, the pattern of findings across times and groups suggests that test score interpretations are generalizable across times and drug offender samples.

Limitations and Conclusions

The present study has several caveats. First, the WAT items were limited to phrasal cues that represented three types of contents (i.e., affective outcome, situations, locations) and their combinations. However, prior studies have also shown the effectiveness of the WAT using other types of cues, including homographs (e.g., "climax"; Stacy et al., 2000) and letter cues (e.g., "t____"; Stacy et al., 2006). Second, the results are generalizable only to adult drug offenders. There are other populations at risk for HIV, including youth and men who have sex with men (Centers for Disease Control and Prevention, 2012). Third, criterion-related evidence was obtained as a form of concurrent validity between WAT and multiple sex partners. Thus, it is not possible to infer the possible causal relationship between them. Last, additional evidence for external validity, such as convergent and discriminant validity and change validity could not be examined. Future research should address these limitations to provide further evidence for the validity of WAT score interpretation.

In summary, the current study provided a multitude of evidence for different components of internal validity as well as criterion-related validity of interpretations of the casual sex-related WAT scores. The WAT items were strongly related to the latent casual sex-related associative memory, and the WAT was reliable, especially at the mean-to-high levels of the latent trait. Consistent with prior research, WAT was related to behavior. These findings contribute to research in several domains in addition to work on cognitive processes related to HIV risk. The findings illustrate that comprehensive psychometric evaluations using IRT and related methods have great utility in work on word association or other assessments using open-ended (fill-in-response) assessment procedures. Open-ended assessment methods are applicable across many substantive areas, such as cognitive, social, and health psychology, as well as cross-cultural and clinical research. The current study also demonstrated that the assessment methods we employed can reveal reliable and useful data in an at-risk drug offender population in the field, providing support for the appropriateness of use of this type of population and setting for evaluation of theory (e.g., Stacy et al., 2006) as well as interventions (e.g., Nydegger, Keeler, Hood, Siegel, & Stacy, 2013). Lastly, the psychometric modeling procedures illustrated in the current study provide researchers with useful tools to evaluate various components of internal validity of interest.

Acknowledgments

This research was supported by grants from the National Institute on Drug Abuse (DA024659, DA023368, DA033871). We wish to thank Amy Custer for her work on this project, and to Steven P. Reise for helpful comments on an earlier draft of the manuscript.

References

Alison J, Burgess C. Effects of chronic non-clinical depression on the use of positive and negative words in language contexts. Brain and Cognition. 2003; 53(2):125–128. [PubMed: 14607131]

Ames SL, Grenard JL, Stacy AW. Dual process interaction model of HIV-risk behaviors among drug offenders. AIDS and Behavior. 2013; 17(3):1–12. [PubMed: 23054037]

Ames SL, Grenard JL, Thush C, Sussman S, Wiers RW, Stacy AW. Comparison of indirect assessments of association as predictors of marijuana use among at-risk adolescents. Experimental and Clinical Psychopharmacology. 2007; 15(2):204–218. [PubMed: 17469944]

Anderson, JR. The architecture of cognition. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.; 1983.

Augustus-Horvath CL, Tylka TL. The acceptance model of intuitive eating: A comparison of women in emerging adulthood, early adulthood, and middle adulthood. Journal of Counseling Psychology. 2011; 58(1):110–125. [PubMed: 21244144]

Borsboom D, Mellenbergh GJ, van Heerden J. The Concept of Validity. Psychological Review. 2004; 111(4):1061–1071. [PubMed: 15482073]

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological). 1995; 57(1):289–300.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In: Lord, FM.; Novick, MR., editors. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968. p. 395-479.

Cai, L. flexMIRT: A numerical engine for multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group; 2012.

Cai, L.; Thissen, D.; du Toit, SHC. IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.; 2011.

Campbell, DT.; Stanley, JC. Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally; 1963.

Centers for Disease Control and Prevention. HIV surveillance in adolescents and young adults. 2012. Retrieved from http://www.cdc.gov/hiv/library/slideSets/index.html

Chen W, Thissen D. Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics. 1997; 22(3):265–289.

Cook, TD.; Campbell, DT. Quasi-experimentation: Design & analysis issues for field setting. Boston, MA: Houghton Mifflin; 1979.

Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin. 1955; 52(4): 281–302. [PubMed: 13245896]

Curran PJ, Hussong AM. Integrative data analysis: The simultaneous analysis of multiple data sets. Psychological Methods. 2009; 14(2):81–100. [PubMed: 19485623]

Czopp AM, Monteith MJ, Zimmerman RS, Lynam DR. Implicit Attitudes as Potential Protection From Risky Sex: Predicting Condom Use with the IAT. Basic and Applied Social Psychology. 2004; 26(2–3):227–236.

Edwards MC. An introduction to item response theory using the need for cognition scale. Social and Personality Psychology Compass. 2009; 3(4):507–529.

Edwards, MC.; Edelen, MO. Special topics in item response theory. In: Millsap, R.; Maydeu-Olivares, A., editors. The Sage handbook of quantitative methods in psychology. New York, NY: Sage; 2009. p. 178-198.

Embretson S. Construct validity: Construct representation versus nomothetic span. Psychological Bulletin. 1983; 93(1):179–197.

Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah, NJ: Erlbaum; 2000.

Flora DB, Curran PJ. An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis With Ordinal Data. Psychological Methods. 2004; 9(4):466–491. [PubMed: 15598100]

Frigon AP, Krank MD. Self-coded indirect memory associations in a brief school-based intervention for substance use suspensions. Psychology of Addictive Behaviors. 2009; 23(4):736–742. [PubMed: 20025382]

Galbraith GG. Reliability of free associative sexual responses. Journal of Consulting and Clinical Psychology. 1968; 32(5):622. [PubMed: 5743322]

Galbraith GG, Lieberman H. Associative responses to double entendre words as a function of repression-sensitization and sexual stimulation. Journal of Consulting and Clinical Psychology. 1972; 39(2):322–327. [PubMed: 5075881]

Grenard JL, Ames SL, Stacy AW. Deliberative and spontaneous cognitive processes associated with HIV risk behavior. Journal of Behavioral Medicine. 2013; 36(1):1–13. [PubMed: 22108762]

Grimm, KJ.; Widaman, KF. Construct validity. In APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics. Washington DC: American Psychological Association; 2012. p. 621-642.

Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice. 1993; 12(3):38–47.

Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling. 1999; 6(1):1–55.

Hutchison KA, Balota DA, Cortese MJ, Watson JM. Predicting semantic priming at the item level. The Quarterly Journal of Experimental Psychology. 2008; 61(7):1030–1066.

Kelly AB, Haynes MA, Marlatt GA. The impact of adolescent tobacco-related associative memory on smoking trajectory: An application of negative binomial regression to highly skewed. Addictive Behaviors. 2008; 33(5):640–650. [PubMed: 18222050]

Kelly AB, Masterman PW, Marlatt GA. Alcohol-related associative strength and drinking behaviours: Concurrent and prospective relationships. Drug and Alcohol Review. 2005; 24(6):489–498. [PubMed: 16361205]

Krank MD, Schoenfeld T, Frigon AP. Self-coded indirect memory associations and alcohol and marijuana use in college students. Behavior Research Methods. 2010; 42(3):733–738. [PubMed: 20805595]

Kristjansson E, Aylesworth R, McDowell I, Zumbo BD. A Comparison of Four Methods for Detecting Differential Item Functioning in Ordered Response Items. Educational and Psychological Measurement. 2005; 65(6):935–953.

Langer, M. A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation. (Unpublished doctoral dissertation). the University of North Carolina at Chapel Hill; 2008.

Leigh BC, Ames SL, Stacy AW. Alcohol, drugs, and condom use among drug offenders: An event-based analysis. Drug and Alcohol Dependence. 2008; 93(1–2):38–42. [PubMed: 17928167]

Levy DA, Stark CEL, Squire LR. Intact conceptual priming in the absence of declarative memory. Psychological Science. 2004; 15(10):680–686. [PubMed: 15447639]

Light RJ. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin. 1971; 76(5):365–377.

Loevinger J. Objective tests as instruments of psychological theory. Psychological Reports. 1957; 3:635–694.

McArdle JJ, Prescott CA. Age-based construct validation using structural equation modeling. Experimental Aging Research. 1992; 18(3–4):87–115. [PubMed: 1459166]

Messick, S. Validity. In: Linn, RL., editor. Educational measurement. New York, NY: Macmillan; 1989. p. 13-103.

Muthen, B.; Muthen, L. MPlus (Version 6.11) [Computer software]. Los Angeles, CA: Muthen & Muthen; 2011.

Muthen, B.; du Toit, SHC.; Spisic, D. Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes (Unpublished technical report). 1997.

Nelson DL, McEvoy CL, Dennis S. What is free association and what does it measure? Memory & Cognition. 2000; 28(6):887–899. [PubMed: 11105515]

Nelson DL, McEvoy CL, Schreiber TA. The University of South Florida free association, rhyme, and word fragment norms. Behavior Research Methods, Instruments & Computers. 2004; 36(3):402–407.

Nelson DL, McKinney VM, Gee NR, Janczura GA. Interpreting the influence of implicitly activated memories on recall and recognition. Psychological Review. 1998; 105(2):299–324. [PubMed: 9577240]

Nydegger LA, Keeler AR, Hood C, Siegel JT, Stacy AW. Effects of a one-hour intervention on condom implementation intentions among drug users in Southern California. AIDS Care. 2013; 25(12):1586–1591. [PubMed: 23656365]

Ofen N, Shing YL. From perception to memory: Changes in memory systems across the lifespan. Neuroscience and Biobehavioral Reviews. 2013; 37(9, Part B):2258–2267. [PubMed: 23623983]

Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement. 2000; 24(1):50–64.

Reise SP, Ainsworth AT, Haviland MG. Item response theory: Fundamentals, applications, and promise in psychological research. Current Directions in Psychological Science. 2005; 14(2):95–101.

Reise SP, Waller NG. Item response theory and clinical measurement. Annual Review of Clinical Psychology. 2009; 5:27–48.

Roediger HL III, Watson JM, McDermott KB, Gallo DA. Factors that determine false recall: A multiple regression analysis. Psychonomic Bulletin & Review. 2001; 8(3):385–407. [PubMed: 11700893]

Rooke SE, Hine DW, Thorsteinsson EB. Implicit cognition and substance use: A meta-analysis. Addictive Behaviors. 2008; 33(10):1314–1328. [PubMed: 18640788]

Rozin P, Kurzer N, Cohen AB. Free associations to 'food': The effects of gender, generation and culture. Journal of Research in Personality. 2002; 36(5):419–441.

Shapiro SS. Word association norms: Stability of response and chains of association. Psychonomic Science. 1966; 4(6):233–234.

Shimamura AP, Squire LR. Paired-associate learning and priming effects in amnesia: A neuropsychological study. Journal of Experimental Psychology: General. 1984; 113(4):556–570. [PubMed: 6240522]

Shono Y, Grenard JL, Ames SL, Stacy AW. Application of Item Response Theory to Tests of Substance-Related Associative Memory Psychology of Addictive Behaviors. 2014; 28(3):852–862. [PubMed: 25134051]

Stacy AW. Memory activation and expectancy as prospective predictors of alcohol and marijuana use. Journal of Abnormal Psychology. 1997; 106(1):61–73. [PubMed: 9103718]

Stacy AW, Ames SL, Ullman JB, Zogg JB, Leigh BC. Spontaneous cognition and HIV risk behavior. Psychology of Addictive Behaviors. 2006; 20(2):196–206. [PubMed: 16784366]

Stacy AW, Newcomb MD, Ames SL. Implicit cognition and HIV risk behavior. Journal of Behavioral Medicine. 2000; 23(5):475–499. [PubMed: 11039159]

Stacy AW, Wiers RW. Implicit cognition and addiction: A tool for explaining paradoxical behavior. Annual Review of Clinical Psychology. 2010; 6:551–575.

Steyvers, M.; Shiffrin, RM.; Nelson, DL. Experimental cognitive psychology and its applications. Washington DC: American Psychological Association; 2005. Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory; p. 237-249.

Szalay, LB.; Deese, J. Subjective meaning and culture : An assessment through word associations. New Jersey, NJ: Erlbaum Associates; 1978.

Szalay, LB.; Strohl, JB.; Fu, L.; Lao, P. American and Chinese perceptions and belief systems: A People's Republic of China–Taiwanese comparison. New York, NY: Plenum Press; 1994.

Thissen, D.; Steinberg, L. Item response theory. In: Millsap, R.; Maydeu-Olivares, A., editors. The Sage handbook of quantitative methods in psychology. Sage; 2009. p. 148-177.

Watkins PC, Vache K, Verney SP, Mathews A. Unconscious mood-congruent memory bias in depression. Journal of Abnormal Psychology. 1996; 105(1):34–41. [PubMed: 8666709]

Wickelgren WA. Network strength theory of storage and retrieval dynamics. Psychological Review. 1976; 83(6):466–478.

Woods CM, Cai L, Wang M. The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. Educational and Psychological Measurement. 2013; 73(3):532–547.

Table 1

CTT Item Analysis, IRT Item Fit statistics, and Item Parameter Estimates for 1PLM and 2PLMs

		24-item WAT		20-item WAT								19-item WAT			
		CTT		1PLM				2PLM				2PLM			
	Item	r_{pb}	*p-value*	*a*	*b*	$S\text{-}X^2$	*p*	*a*	*b*	$S\text{-}X^2$	*p*	*a*	*b*	$S\text{-}X^2$	*p*
1	hotel/motel	.55	.27	2.00	0.83	8.9	.71	2.00	0.83	9.1	.70	2.00	0.83	10.1	.52
2	friend's house/apartment	.52	.10	2.00	1.70	9.7	.79	2.54	1.55	12.4	.50	2.54	1.55	12.2	.59
3	with friends	.50	.09	-	-	-	-	-	-	-	-	-	-		
4	hanging out	.53	.10	2.00	1.74	17.8	.21	2.64	1.57	15.7	.33	2.66	1.56	15.3	.29
5	his house/her house	.48	.17	2.00	1.31	12.0	.61	1.85	1.35	10.9	.70	1.87	1.35	12.5	.49
6	at a motel, feels good	.62	.24	-	-	-	-	-	-	-	-	-	-		
7	at a motel, feels pleasant	.54	.20	2.00	1.17	12.7	.55	1.99	1.17	12.6	.56	1.98	1.18	10.4	.67
8	with friends, feels pleasant	.46	.06	2.00	2.07	10.4	.80	2.26	1.96	9.4	.81	-	-		
9	at a bar, feels pleasant	.44	.09	2.00	1.81	21.2	.13	1.85	1.87	19.4	.19	1.85	1.87	21.2	.10
10	Friday night, feeling good, drinking	.60	.14	-	-	-	-	-	-	-	-	-	-		
11	Friday night, having fun, getting drunk	.53	.14	2.00	1.46	18.0	.21	2.01	1.45	18.2	.20	1.99	1.46	19.2	.12
12	Friday night, feeling good, using drugs	.49	.13	-	-	-	-	-	-	-	-	-	-		
13	Friday night, pleasant feelings, getting high	.48	.11	2.00	1.66	6.4	.95	1.79	1.74	7.2	.93	1.80	1.73	8.3	.88
14	forgetting problems	.44	.13	2.00	1.50	17.4	.23	1.77	1.58	13.5	.57	1.78	1.58	13.7	.48
15	feeling relaxed	.43	.08	2.00	1.89	12.7	.62	1.99	1.89	12.7	.63	2.00	1.88	11.8	.70
16	having Fun	.43	.11	2.00	1.60	20.5	.11	1.84	1.66	18.2	.20	1.87	1.65	19.5	.15
17	Friday night, having fun	.60	.15	2.00	1.41	30.0	.01	2.85	1.26	20.5	.08	2.86	1.26	16.5	.17
18	Friday night, excitement	.62	.18	2.00	1.27	18.8	.13	3.02	1.12	16.5	.17	3.02	1.12	21.3	.05
19	Friday night, feelings of pleasure	.60	.26	2.00	0.91	16.7	.16	2.49	0.85	16.7	.16	2.50	0.85	16.2	.13
20	Friday night, pleasing others	.38	.09	2.00	1.84	21.4	.09	1.73	1.96	17.7	.28	1.73	1.96	19.0	.21
21	new partner	.42	.31	2.00	0.72	23.9	.02	1.33	0.86	6.2	.94	1.35	0.85	11.2	.60
22	excitement	.51	.17	2.00	1.26	12.8	.54	1.99	1.26	12.6	.56	2.00	1.26	9.9	.62
23	feels good	.53	.20	2.00	1.15	17.2	.19	2.11	1.13	17.4	.18	2.15	1.12	28.1	.01
24	pleasant	.37	.07	2.00	1.93	12.6	.56	1.67	2.11	13.8	.47	1.69	2.09	14.0	.52

Note. WAT = word association test; CTT = classical test theory; r_{pb} = corrected point-biserial correlation (item-total correlation); *p-value* = proportion of participants who generated casual sex-related responses; 1PLM = one-parameter logistic model; 2PLM = two-parameter logistic model; *a* = item discrimination parameter; *b* = item difficulty parameter; $S\text{-}X^2$= item fit index.

Table 2

Results of Differential Item Functioning Tests using One-Step Wald Test.

Subgroup		Items	X^2	*df*	*p*	*a*		*b*	
Gender						M	F	M	F
	8	with friends, feels pleasant	7.3	2	**0.03**	**2.27**	**0.97**	1.83	3.57
	19	Friday night, feelings of pleasure	4.8	2	0.09	2.53	1.56	0.66	1.00
Age Group						EM	ER	EM	ER
	2	friend's house/apartment	6.3	2	0.04	2.46	2.51	1.32	1.76
	19	Friday night, feelings of pleasure	7.7	2	0.02	1.88	2.77	0.72	0.92
Sex Partners						MN	C/E	MN	C/E
	13	Friday night, pleasant feelings, getting high	5.1	2	0.08	1.68	1.18	1.99	1.87
	21	new partner	5.7	2	0.58	1.74	0.96	0.74	0.79
Ethnicity						H	NHW	H	NHW
	1	hotel/motel	9.2	2	0.07	2.17	1.66	0.68	1.12
	2	friend's house/apartment	5.7	2	0.06	2.21	3.20	1.53	1.91
	19	Friday night, feelings of pleasure	6.8	2	0.03	2.09	1.18	1.78	1.94

Notes. *p*-values are compared against the adjusted critical values obtained from the Benjamini-Hochberg procedure (not shown here) to control for the false discovery rate, which is set at .05. *p*-values and item parameter values in bold-face indicate that they are significant. *a* = discrimination parameter; *b* = difficulty parameter; M = Male; F = Female; EM = Emerging adults; ER = Early adults; MN = Main partners; C/E = Casual or exchange partners; H = Hispanic; NHW = Non-Hispanic White.

Table 3

Estimates of Latent Trait Scores and Observed Scores by Subgroups

			Latent trait scores				Observed scores	
Subgroup		***n***	***M***	***SE***	***Z***	***p***	***M***	***SD***
Study	Study 1	485	0.00	-			2.93	3.82
	Study 2	653	−0.04	0.08	−0.50	*ns*	2.63	3.55
Gender	Male	818	0.00	-			3.24	3.92
	Female	315	−0.75	0.1	−7.50	< .01	1.49	2.51
Ethnicity	Hispanic	561	0.00	-			2.60	3.41
	Non-Hispanic White	426	0.06	1.08	0.86	*ns*	2.91	3.77
Age Group	Emerging Adult	448	−0.04	1.02	−0.50	*ns*	2.72	3.60
	Early Adult	388	0.00	-			2.89	3.70
	Middle Adult	273	−0.18	1.13	−2.00	< .01	2.62	3.73
Sex Partners	Main only	307	0.00	-			2.48	3.43
	Casual/Exchange	314	0.36	1.17	4.50	<.01	3.10	3.64

Note. *Z* = Wald Z statistic.

Table 4

Item and test information as a function of a latent trait (θ) in the final casual sex-related WAT.

		Latent trait level (θ)												
	Items	**−3.0**	**−2.5**	**−2.0**	**−1.5**	**−1.0**	**−0.5**	**0.0**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**
1	hotel/motel	0.00	0.01	0.01	0.04	0.10	0.24	0.54	0.90	0.97	0.66	0.32	0.13	0.05
2	friend's house/apt	0.00	0.00	0.00	0.00	0.01	0.03	0.12	0.39	1.02	1.61	1.18	0.49	0.15
3	hanging out	0.00	0.00	0.00	0.00	0.01	0.03	0.11	0.37	1.05	1.75	1.28	0.50	0.15
4	his house/her house	0.00	0.00	0.01	0.02	0.04	0.10	0.24	0.49	0.78	0.85	0.61	0.33	0.15
5	at a motel, feels pleasant	0.00	0.00	0.01	0.02	0.05	0.13	0.32	0.65	0.95	0.89	0.54	0.25	0.10
6	at a bar, feels pleasant	0.00	0.00	0.00	0.01	0.02	0.04	0.10	0.23	0.47	0.76	0.84	0.62	0.34
7	Friday night, having fun, getting drunk	0.00	0.00	0.00	0.01	0.03	0.08	0.20	0.45	0.81	0.99	0.75	0.39	0.17
8	Friday night, pleasant feelings, getting high	0.00	0.00	0.00	0.01	0.02	0.06	0.13	0.29	0.54	0.78	0.77	0.52	0.27
9	forgetting problems	0.00	0.00	0.01	0.01	0.03	0.07	0.17	0.35	0.61	0.79	0.69	0.43	0.22
10	feeling relaxed	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.22	0.50	0.87	0.99	0.70	0.35
11	having fun	0.00	0.00	0.00	0.01	0.02	0.06	0.15	0.33	0.62	0.86	0.79	0.49	0.24
12	Friday night, having fun	0.00	0.00	0.00	0.00	0.01	0.05	0.21	0.76	1.79	1.82	0.78	0.22	0.06
13	Friday night, excitement	0.00	0.00	0.00	0.00	0.01	0.07	0.29	1.05	2.21	1.68	0.56	0.14	0.03
14	Friday night, feelings of pleasure	0.00	0.00	0.01	0.02	0.06	0.20	0.59	1.29	1.51	0.86	0.32	0.10	0.03
15	Friday night, pleasing others	0.00	0.00	0.00	0.01	0.02	0.04	0.09	0.20	0.40	0.64	0.75	0.61	0.37
16	new partner	0.01	0.02	0.04	0.07	0.13	0.22	0.33	0.43	0.45	0.38	0.26	0.16	0.09
17	excitement	0.00	0.00	0.01	0.02	0.04	0.11	0.28	0.59	0.94	0.95	0.61	0.29	0.12
18	feels good	0.00	0.00	0.01	0.02	0.05	0.13	0.35	0.76	1.13	0.98	0.53	0.22	0.08
19	pleasant	0.00	0.00	0.00	0.01	0.02	0.03	0.08	0.17	0.33	0.56	0.71	0.63	0.42
	Total Information	1.02	1.05	1.11	1.27	1.68	2.74	5.38	10.92	18.12	19.67	14.28	8.21	4.37
	Test Reliability	0.02	0.04	0.10	0.21	0.41	0.64	0.81	0.91	0.94	0.95	0.93	0.88	0.77