Military Psychology

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653681

Evaluating the Psychometric and Measurement Characteristics of a Measure of Sexual Orientation Harassment

Armando X. Estrada\textsuperscript{a}; Tahira M. Probst\textsuperscript{a}; Jeremiah Brown\textsuperscript{a}; Maja Graso\textsuperscript{a}

\textsuperscript{a} Department of Psychology, Washington State University Vancouver, Vancouver, Washington

Online publication date: 04 March 2011

To cite this Article Estrada, Armando X., Probst, Tahira M., Brown, Jeremiah and Graso, Maja(2011) 'Evaluating the Psychometric and Measurement Characteristics of a Measure of Sexual Orientation Harassment', Military Psychology, 23: 2, 220 — 236

To link to this Article DOI: 10.1080/08995605.2011.559394

URL: http://dx.doi.org/10.1080/08995605.2011.559394

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Evaluating the Psychometric and Measurement Characteristics of a Measure of Sexual Orientation Harassment

Armando X. Estrada, Tahira M. Probst, Jeremiah Brown, and Maja Graso

Department of Psychology, Washington State University Vancouver, Vancouver, Washington

We use classical test theory (CTT) and item response theory (IRT) methodologies to examine the psychometric and measurement properties of an instrument designed to assess sexual orientation harassment among military personnel (N = 71,989). CTT analyses indicated that items were unidimensional and exhibited adequate levels of reliability. IRT analyses demonstrated that the items functioned similarly and exhibited appropriate levels of item discrimination. However, the analyses also suggested that the sensitivity of the items may be limited. Differential test functioning analyses provided evidence of the measurement equivalence of the instrument across male and female respondents. The findings provide support for the psychometric properties and measurement equivalence of the instrument for measuring sexual orientation harassment among male and female military personnel. We discuss the implications of our findings for future research on sexual orientation harassment in the workplace.

Empirical research on workplace harassment has increased dramatically in the past two decades. Although studies to date have increased our understanding of the nature and prevalence of sexual harassment (U.S. Merit Systems Protection Board [USMSPB], 1988, 1995) and ethnic harassment (Scarville, Button, Edwards, Lancaster, & Elig, 1999; Schneider, Hitlan, & Radhakrishnan, 2000), research...
has only begun to explore sexual orientation harassment (Croteau, 1996; Ragins, 2004). Ragins noted that though there has been extensive research on the general construct of sexual orientation from psychological, economic, and sociological perspectives, research on sexual orientation in the workplace is still in its infancy. Thus far, research suggests that sexual orientation harassment is as prevalent as other forms of harassment (Button, 2001; Moradi, 2006; Office of the Inspector General, 2000; Ragins, 2004). Other research suggests that the experience of sexual orientation harassment has a negative impact on individual’s job, health, and psychological outcomes (Button; Ragins; Ragins & Cornwell, 2000).

This research has increased our understanding of the nature and impact of sexual orientation harassment. However, it is worth noting that there are methodological and theoretical shortcomings that need to be addressed within this literature. Most notable are issues concerning the validity of existing measures of sexual orientation–based harassment (Croteau, 1996; Herek, Gillis, Cogan, & Glunt, 1997; Ragins, 2004). Ragins and Cornwell (2000) and Croteau have commented on the general lack of psychometrically validated measures of workplace discrimination based on sexual orientation. Indeed, researchers have relied on the use of atheoretical surveys that employ pragmatically driven items. There is a critical need to develop psychometrically sound measures of sexual orientation harassment. This is particularly important given the increasing attention that sexual orientation diversity receives in today’s organizations.

The U.S. Department of Defense (DoD) has tried to address a number of challenges associated with the inclusion of gay service personnel. The DoD has developed programs to curtail harassment and introduced Anti-Harassment Action Plan (AHAP) to addresses anti-harassment training, procedures for reporting harassment, enforcement against harassment, and measurement to assess adherence to anti-harassment policies as well as the effectiveness of such policies (Department of Defense, 2000). Though these efforts represent a step in right direction, it is difficult to assess the efficacy of the policy or anti-harassment training without well-validated measures to assess the nature and prevalence of sexual orientation harassment.

This article begins to address this limitation by evaluating the measurement characteristics of a questionnaire designed by the DoD to assess harassment based on perceived sexual orientation. Specifically, we employ both classical test theory (CTT) and item response theory (IRT) to examine psychometric and measurement properties of an instrument designed to assess sexual orientation harassment. Further, we utilize Raju, van der Linden, and Fleer’s (1995) analysis of differential test functioning (DTF) to examine the measurement equivalence of the scale across male and female respondents. Our use of both CTT and IRT approaches provides a rigorous examination of both psychometric properties and measurement equivalence of this instrument.

Before presenting our main findings, we provide a brief review of research on the incidence and prevalence of sexual orientation harassment, outlining concerns
and limitations of this literature. Next, we provide a description of the instrument used to assess sexual orientation harassment in the DoD study (Office of the Inspector General, 2000) and review research on sexual harassment as well as anti-gay prejudice to derive empirical support for examining gender differences in sexual orientation harassment. The final section presents the methodology and results of our study along with a discussion of the implications of our findings.

INCIDENCE AND PREVALENCE OF SEXUAL ORIENTATION HARASSMENT

A small but growing body of research on sexual orientation harassment suggests that it is quite prevalent in our society. Studies of university students (D’Augelli, 1989, 1992; Finn, 2004; Herek, 1993), working adults (Button, 2001; Croteau, 1996; Herek, Cogan, & Gillis, 2002; Ragins, 2004; Ragins & Cornwell, 2000), and military personnel (Moradi, 2006; Office of the Inspector General, 2000) indicate that anywhere between 20% to over 60% of respondents have experienced some form of sexual orientation harassment in their academic or working life. It is worth noting that these figures do not necessarily reflect harassment that is exclusively due to an individual’s sexual orientation because the majority of sexual minorities1 do not tend to disclose their sexual orientation to other individuals (Badgett, 1996; Ragins).

Research examining the nature of these harassment experiences suggests that the most prevalent form of sexual orientation harassment appears to involve verbal abuses and interpersonal threats (Croteau, 1996; Herek, 1989). Illustrative examples of this finding may be gleaned from D’Augelli (1989, 1992), who reported that approximately 75% of college respondents surveyed experienced verbal abuse, as well as Morris and Balsam (2003), who reported that 56.6% of respondents in a large, ethnically diverse sample of sexual minority individuals experienced verbal attacks. Similar rates have also been reported in studies of military personnel (Office of the Inspector General, 2000), with 80% of military respondents reporting that they heard offensive speech, derogatory names, jokes, or remarks about homosexuals in the 12 months preceding the survey. The results of this latter study are noteworthy because these figures are not differentiated by the sexual orientation of the respondents. These figures highlight the ubiquity of sexual orientation harassment in the military workplace.

In sum, the available evidence suggests that sexual orientation harassment appears to be pervasive (Croteau, 1996; Ragins, 2004). However, it is important to note that the preponderance of this research has used atheoretical surveys that

1We use the term sexual minority to be inclusive of lesbian, gay, bisexual, and transsexual individuals.
rely on pragmatically derived items to estimate the incidence and prevalence of sexual orientation harassment (Croteau, 1996; Herek et al., 1997; Ragins). And, though some research has examined the nature and extent of sexual orientation harassment with large-scale samples from diverse groups (Croteau, 1996; Morris & Balsam, 2003; Office of the Inspector General, 2000; Ragins), substantive differences in the conceptualization of sexual orientation harassment and differences in research methodologies employed across these studies make it difficult to draw firm conclusions regarding the nature and extent of sexual orientation harassment across studies (Croteau, 1996; Ragins, 2004). Accordingly, we cannot state any definitive conclusions regarding the types of behaviors and the frequency of their experiences, nor can we begin to systematically compare differences in perceptions of these experiences for men and women or other subgroups (e.g., sexual minorities). In short, there is a critical need for validated measures of sexual orientation harassment.

**SEXUAL ORIENTATION–BASED HARASSMENT IN THE U.S. MILITARY**

On December 13, 1999, at the request of Secretary of Defense William S. Cohen, the Office of the Inspector General was tasked with assessing the implementation of the policy on homosexual conduct in the military. The assessment examined the extent to which the harassment of servicemembers based on perceived or alleged homosexuality occurred; the extent to which disparaging speech or expression with respect to sexual orientation occurred; and the extent to which disparaging speech or expressions with respect to sexual orientation were tolerated (Office of the Inspector General, 2000). The study included eight items that asked respondents to report the degree to which they had experienced or witnessed various types of harassing behaviors in the previous 12 months. All items employed a common stem that asked respondents to indicate, “How often during the past 12 months have you witnessed or experienced event(s)/behaviors(s) involving military personnel, on or off duty, who harassed another military person(s) because of perceived homosexuality (1) with offensive speech? (2) with offensive or hostile gestures? (3) with threats or intimidation? (4) by graffiti? (5) by vandalism of their property? (6) by physical assault? (7) by limiting or denying training and/or career opportunities? (8) in disciplinary action or punishment?” Respondents indicated the extent to which they had witnessed or experienced each of these events or behaviors using a 5-point response scale ranging from *never* (1) to *very often* (5). Incidence rates were estimated by summing across items such that if an individual gave an affirmative response to at least one item on the survey he or she was classified as experiencing or witnessing sexual orientation harassment. Thirty-seven percent of military respondents indicated that they “had witnessed or experienced one or more of eight specific events or behaviors toward a service member that
they considered to be harassment based on perceived homosexuality” (Officer of the Inspector General, p. 8).

This study was particularly noteworthy because it had several methodological strengths. The study included responses from large, randomly selected, representative samples drawn from military installations and units ($N = 71,500$). The study also incorporated methodological innovations from research on workplace harassment by using behaviorally based questions to assess multiple forms of sexual orientation harassment. Finally, the study was carried out in the largest single employer in the United States—the DoD. Despite these methodological strengths, there were several shortcomings in this study. The most notable one involved the limited examination of the measurement characteristics of the instrument used to assess sexual orientation harassment. Thus, we know little about the psychometric properties of the instrument. Further, it is unclear whether the instrument can be used to make comparison across key demographic indicators such as gender.

There are several reasons why perceptions of sexual orientation harassment experiences may differ by gender. Although public opinion regarding sexual orientation issues has changed in recent years (Herek, 2009; Sherrill & Yang, 2000; Yang, 1997), sexual minorities continue to experience considerable hostility, discrimination, and violence in our society (e.g., Herek, 2000, 2002, in press; Probst, Estrada, & Brown, 2008; Ragins, 2004; Ragins & Cornwell, 2000; Ragins & Wiethoff, 2005; Rubenstein, 2004). Research also shows that heterosexuals’ acceptance of sexual minorities varies by gender, with men expressing less accepting attitudes than women (e.g., Herek, 2000, 2002; Herek & Capitanio, 1999; Kite & Whitley, 1996; Oliver & Hyde, 1993). Other research shows that men and women differ in their perceptions of socio-sexual behavior, with women being more likely to consider a wider range of socio-sexual behaviors as harassment compared with men (Blumenthal, 1998; Rotundo, Nguyen, & Sackett, 2001). Finally, research shows that the nature and extent of harassment varies by gender—women experience higher levels of sexual harassment than men (Donovan & Drasgow, 1999; Illies, Hauserman, Schowochau, & Stibal, 2003), and men are more likely to experience certain forms of harassment than women (Berdahl, Magley, & Waldo, 1996; Stockdale, Visio, & Batra, 1999; Waldo, Berdahl, & Fitzgerald, 1998). Taken together this work suggests that men and women differ in their attitudes toward sexual minorities; their interpretation of socio-sexual behavior; and in the nature and extent of their experiences with various forms of harassing behaviors. Thus, it is reasonable to question whether gender may play a similar role with regard to sexual orientation harassment. However, before we can examine this question, it is important to evaluate whether an instrument used to assess sexual orientation harassment can be used to make accurate comparison regarding sexual orientation harassment experiences of both men and women. Therefore, we propose to evaluate the psychometric properties and measurement characteristics of this instrument to determine whether there
is evidence of differential test functioning. A finding of measurement equivalence would then allow greater confidence in future explorations of relational equivalence between men and women’s harassment reports (Drasgow & Kang, 1984).

METHOD

Data and Procedures

Data for this study came from 71,989 active members of the United States Armed Forces (N = 59,882 males; N = 11,568 females; N = 539 nonreported gender) who responded to the DoD Inspector General’s Don’t Ask, Don’t Tell, Don’t Pursue Survey of servicemembers administered in January–February of 2000. Respondents included members from each of the military departments (34% Army, 24% Navy, 12% Marine Corps, and 29% Air Force). Forty-seven percent of the respondents were junior enlisted rank (e.g., E1–E4), 38% were senior enlisted rank (e.g., E5–E9), 9% were junior grade officers (e.g., W01–W02/O1–O3), and 6% were senior grade officers (e.g., W03–W05/O4–O10). The demographic and military characteristics of the participants were comparable to those of the active duty population serving in the U.S. military during the same time period (Office of the Inspector General, 2000).

Representative samples of active duty servicemembers were drawn from 38 randomly selected military installations and units within each of the military departments. At each site, a team of representatives from the Office of the Inspector General collected data from randomly selected units identified. The team explained the general purpose of the study and provided each respondent with a copy of the survey. Respondents were informed that their responses were completely anonymous and no identifying information was collected in the survey. Respondents were also instructed to seat themselves so that enough space was provided to ensure the privacy of their responses. Respondents placed their completed surveys in a box near the exit of the room.

Measures

The Don’t Ask, Don’t Tell, Don’t Pursue (DADTDP) policy survey included several multi-item scales assessing background and workplace information, individual perceptions regarding the DADTDP policy and training, and respondents’ experiences of direct and indirect (witnessing) sexual orientation–based harassment. Because the primary purpose of this article was to assess the measurement characteristics of the Sexual Orientation Harassment (SOH) Scale, we describe only those items used in our analyses.
Sexual Orientation Harassment Scale. Sexual orientation–based harassment was assessed with eight items described previously. All items employed a common stem and asked respondents to report the degree to which they had experienced or witnessed various types of harassing behaviors in the previous 12 months (see Table 1). Responses were provided on a 5-point response scale that ranged from never (1) to very often (5).

Analytical Approach

Our analyses employ CTT and IRT to examine the psychometric properties and assess the measurement characteristics of this instrument. Accordingly, we report psychometric properties derived from CTT (e.g., item-total correlations, factor analytical results, and reliability estimates) as well as measurement characteristics derived from IRT and DTF analyses (e.g., measurement equivalence statistics). IRT has two primary advantages over CTT when evaluating the measurement equivalence of scales across populations. First, parameters characterizing the psychometric properties of the items are not sample dependent (i.e., items are not dependent on the subpopulation of respondents), which is the case for parameter estimates derived from CTT (e.g., proportion correct and item-total correlations). Second, the IRT estimate of an individual’s standing on the latent trait of interest is not dependent on the specific items used in a scale, which is the case for “true score” estimates derived from CTT. This invariance of item and ability parameters allows direct comparisons of item and test parameters across two different populations (e.g., males and females) that CTT estimates would not afford.

Prior to performing our analyses, we followed recommendations proposed by Hulin, Drasgow, and Parsons (1983) and Candell and Drasgow (1988) and scored all items dichotomously as either 0 (never experiencing or witnessing sexual orientation based harassment) or 1 (experiencing or witnessing sexual orientation based harassment at least once) because the base rate of responses was low. This procedure is consistent with existing research and practice in the area of sexual harassment (e.g., Donovan & Drasgow, 1999). Below we describe steps taken in the IRT and DTF analyses. IRT and DTF analyses were performed separately with two sets of data that were randomly selected from the original sample of 59,882 usable cases. A set of data was allocated for calibration (n = 35,994), and a second set of data was allocated for validation (n = 35,995).

Selection and assessment of IRT model. The first step in an IRT analysis involved selecting an IRT model that provided a reasonable fit to the data. A two-parameter logistic model was used to model the data in this study because previous research has indicated that two-parameter models are appropriate for items
<table>
<thead>
<tr>
<th>Item</th>
<th>Endorsement Rate (%)</th>
<th>Factor Loading</th>
<th>Women</th>
<th>Men</th>
<th>Combined Calibration Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offensive speech</td>
<td>33.9</td>
<td>.92</td>
<td>2.08</td>
<td>0.42</td>
<td>2.40</td>
</tr>
<tr>
<td>Offensive/hostile gestures</td>
<td>20.4</td>
<td>.93</td>
<td>2.41</td>
<td>0.94</td>
<td>3.32</td>
</tr>
<tr>
<td>Threats or intimidation</td>
<td>12.9</td>
<td>.95</td>
<td>2.75</td>
<td>1.28</td>
<td>3.15</td>
</tr>
<tr>
<td>Graffiti</td>
<td>10.2</td>
<td>.86</td>
<td>1.45</td>
<td>1.81</td>
<td>1.15</td>
</tr>
<tr>
<td>Vandalism</td>
<td>5.9</td>
<td>.96</td>
<td>2.23</td>
<td>1.78</td>
<td>2.15</td>
</tr>
<tr>
<td>Physical assault</td>
<td>6.0</td>
<td>.96</td>
<td>2.22</td>
<td>1.72</td>
<td>2.31</td>
</tr>
<tr>
<td>Limit career opportunities</td>
<td>5.9</td>
<td>.94</td>
<td>1.70</td>
<td>1.69</td>
<td>1.98</td>
</tr>
<tr>
<td>Discipline or punishment</td>
<td>6.5</td>
<td>.94</td>
<td>1.78</td>
<td>1.66</td>
<td>1.94</td>
</tr>
</tbody>
</table>


*NC-DIF = noncompensatory DIF.
that are unlikely to be affected by guessing (Hulin, 1987; Hulin et al., 1983), like those involving harassing behaviors (Donovan & Drasgow, 1999). Before applying this two-parameter IRT model it was necessary to assess the fundamental assumptions underlying IRT—unidimensionality, local independence, and model fit. We performed a principal axis factor (PAF) analysis, using tetrachoric correlations computed from the dichotomously scored item responses, to evaluate the unidimensionality of the scale. A single-factor extraction and/or large loadings on the first principal factor are indicative of scale unidimensionality (Drasgow & Parsons, 1983). The second and third assumptions were assessed utilizing MODFIT (Stark, 2001), which employs procedures outlined by Drasgow, Levine, Tsien, Williams, and Mead (1995) to evaluate model fit. MODFIT compares item parameter estimates derived from the calibration sample for the selected IRT model (in this case, a 2 parameter-logistic model [2-PL]) to the empirical proportions derived from the validation sample. Graphical fit plots are then constructed to compare the actual proportion of respondents endorsing an item in a given theta interval to the estimated proportion based on the 2-PL model item parameters. A $\chi^2$ goodness of fit statistic (adjusted to a sample size of $N = 3,000$ as recommended by Drasgow et al.) can then be computed for each item to compare the actual number of correct responses to the number expected based on the IRT model. Additional $\chi^2$ statistics can be computed for item pairs and item triples, which provide a test of the IRT assumption of local independence. A general rule of thumb is that adjusted $\chi^2/df$ statistics should ideally be below 3.00 (Stark, Chernyshenko, Chuah, Lee, & Wadlington, 2001).

**Item parameter estimates.** Because a large proportion of the sample experienced and/or witnessed no harassment, we estimated item parameters using Bayesian expected a posteriori (EAP) estimation procedures (Bock & Mislevy, 1982) as implemented in BILOG 3 (Mislevy & Bock, 1991). EAP estimation was selected instead of maximum likelihood (ML) estimation because ML estimates for “perfect” response patterns (e.g., all 0s or all 1s) are infinite, whereas EAP estimates are always finite. Upper limits of 100 Expeted-Maximization cycles, 10 Newton-Raphson cycles, and a convergence criterion of .001 were set. Initially, we computed parameter estimates using a 50% random calibration sample and then compared these estimates with the empirical response data from the remaining 50% of the sample to assess model–data fit (see below). Once model–data fit was established, we computed parameter estimates for male and female samples separately.

**Iterative linking of metrics.** In order to compare responses in the female and male samples, it was necessary to link the scales (i.e., place item parameter estimates for both samples on a single common scale). Thus, item parameters determined in the sample of females (focal group) were transformed to the metric
of the sample of males (referent group) using the test characteristic curve equating procedure developed by Stocking and Lord (1983) and further outlined by Candell and Drasgow (1988) as implemented in ITERLINK06 (Stark, 2006). This procedure requires that initial linking coefficients are produced to transform the parameter estimates from the focal group (e.g., men) to the metric of the reference group (e.g., women). After linking metrics, items identified as exhibiting differential item functioning (DIF), using Lord’s (1980) chi-square DIF analysis, are removed and linking recurs using only unbiased items. DIF items are removed and linking recurs using only unbiased items. The DIF analysis and linking repeat until the same DIF items are flagged in two successive iterations.

Assessment of DTF. We used DFITD4 (Raju, 1999) to perform the DTF analysis after iterative linking was completed. DTF procedures were developed to assess whether a test or scale in its entirety functions differently across different groups (Raju et al., 1995). DTF allows for the examination of measurement equivalence at the level of the total test or scale so that item-level DIFs in opposite directions (i.e., compensatory DIF) are allowed to cancel each other out. A nonsignificant DTF index indicates that although some items may exhibit DIF, the scale as a whole provides equivalent measurement for the groups being compared. DTF is said to be present when both the DTF index is greater than .006 and the chi-square statistic assessing the DTF between groups is significant (Fleer, 1993). DFITD4 employs procedures developed by Raju et al. to evaluate measurement equivalence of tests. When the DTF statistic is significant, DFITD4 identifies and removes items successively until either the DTF index drops below .006 or the chi-square becomes nonsignificant.

RESULTS

CTT Analyses

We computed tetrachoric correlation coefficients for the dichotomously scored data to perform a principal axis factor analysis. Results of principal axis factor analysis clearly suggested that items loaded into a single unidimensional factor—item loadings ranged from .85 to .95. Examination of eigenvalue statistics and a scree plot further supported this interpretation—eigenvalue for the factor was 7.63 (accounting for 88% of the common variance); a single factor was also evident in the scree plot. The Cronbach’s alpha coefficient of .83 provided similar evidence of internal consistency.²

²Cronbach’s alpha coefficient was computed using the continuous rather than the dichotomous form of the scale (i.e., 5- vs. 2-point response scale).
IRT Analysis

To investigate the 2-PL model’s fit to the data we examined chi-square statistics and ratios of the chi-square statistics to their degrees of freedom ($\chi^2/df$) for the single items, item pairs, and item triples for the entire data set. As noted previously, we split the data into a calibration sample ($n = 35,994$) and validation sample ($n = 35,995$) for parameter estimation and model evaluation. We employed similar procedures in our analyses of gender differences in test-level responses and divided the sample into equivalent sets for each gender separately.

Chi-squares calculated for single items can lead to overfitting; however, the fits for item pairs and item triples provide a challenging test of the IRT model. Examination of $\chi^2/df$ ratios indicated a satisfactory fit of the two-parameter model to the data. For the initial analyses involving the comparison of the calibration and validation samples for the entire data set, the mean $\chi^2/df$ ratio was 1.24 for the singlets, 2.91 for the doublets, and 4.14 for the triplets.

To further investigate the two-parameter model’s fit to the data, fit plots comparing two-parameter logistic IRFs to empirically determined IRFs were constructed for each item (Drasgow et al., 1995). Visual inspection of the fit plots indicated that the estimated IRFs for the two-parameter model were quite similar to the empirically determined IRFs. For example, the fit plot for item 1 (see Figure 1) shows that the estimated IRF for the two-parameter model is reasonably similar to the empirical IRF. These results suggest that a two-parameter model provides an adequate fit to the data.

Given that a good-fitting IRT model was identified, we examined parameter estimates for individual items. Table 1 presents parameter estimates for the entire

![Fit Plot for Item 1](image)

**FIGURE 1** Fit plot comparing item response function to empirical curve from responses to item 1.
sample and for each gender separately (after the metrics were first equated). We estimated two parameters: (a) item discrimination and (b) item difficulty. Item discrimination index can range from 0 to >3, with parameters >3 being too discriminating. The item difficulty index on a normalized scale can range from −3 to +3 and thus indicates the latent level of SOH needed for 50% of respondents to endorse the item. A good scale would have items that span that range.

As shown in Table 1, the majority of the items exhibited appropriate levels of item discrimination, though item 2 may be too discriminating. As is demonstrated by the b-parameter estimates, items 1 and 2 were relatively “easy” items (i.e., higher likelihood of endorsement even among individuals with lower latent trait levels of SOH). Thus, offensive speech and offensive/hostile gestures were more commonly endorsed. Items such as physical assault and vandalism were much more difficult items, indicating that endorsement was only likely among individuals scoring high on the latent variable SOH.

Finally, we also calculated the overall TRF for the calibration sample (see Figure 2). Examination of the TRF indicates that the sensitivity of the items in assessing varying levels of sexual orientation–based harassment may be somewhat limited. For example, a 50% probability of experiencing or witnessing harassment is not reached until a theta level of nearly 1.50, again suggesting a somewhat difficult test.

DTF Analysis

In the final analyses, we used the data from the male and female samples to assess DTF for the scale across the two groups. DTF analyses were based on men and

![TCC Plot](image)

FIGURE 2 Test characteristic curve for the SOH scale.
FIGURE 3  Male and female item response functions for item 4.

women’s item-level parameters, women’s individual-level theta estimates, and the A and K linking metric values obtained from the ITERLINK analyses. Although the $\chi^2$ was 2,217.64, $p < .001$, the DTF index was .003 (well below the cutoff of .006), indicating that the instrument functions equivalently for men and women. Indeed, individual item contributions to DTF were quite small (see Table 1). Figure 3 shows the IRFs for men and women for item 4, which exhibited the highest level of noncompensatory DIF.

DISCUSSION

We examined the psychometric and measurement characteristics of an instrument designed to assess sexual orientation harassment. Specifically, we utilized CTT and IRT methodologies to examine the psychometric properties of the instrument and DTF methodology (Raju et al., 1995) to examine the measurement equivalence of the instrument for male and female respondents.

CTT and IRT analysis provide strong support for the psychometric properties of the instrument. CTT-based analyses indicated that items were unidimensional and exhibited adequate levels of reliability. IRT-based analyses revealed that the items functioned quite similarly, with the majority of items exhibiting appropriate levels of item discrimination. However, these analyses also suggest that the sensitivity of these items for assessing varying levels of sexual orientation harassment may be somewhat limited. One explanation for this finding may be gleaned by examining the nature of the items contained in this instrument. Three
items assessed verbal behaviors, three items assessed physical behaviors, and two items assessed discriminatory behaviors. Of these, the three items involving verbal behaviors (e.g., offensive speech, threats) were endorsed more frequently than items involving physical or discriminatory behaviors (e.g., vandalism, assault, exclusion from training or career opportunities). Thus, it would appear that the instrument does a far better job at assessing some forms of sexual orientation harassment than others. Though we certainly acknowledge this limitation, we do note that the pattern of results involving verbal behaviors is consistent with studies examining the nature of sexual orientation harassment (Croteau, 1996; Herek, 1989, 2009). Moreover, the item content of this instrument is similar to that of other surveys within this literature. Herek (1989, 2009) and Berrill (1990) have relied on similar types of items to assess anti-gay violence and victimization among sexual minority populations.

DTF analyses provided evidence of the measurement equivalence of the instrument for male and female respondents. Analysis of the instrument using Raju’s (1999) DFITD4 program found no evidence of DTF. Thus, the instrument functioned the same across male and female respondents. Visual comparison of IRFs also indicated that the instrument functioned virtually identically for men and women, and the comparison of TRFs revealed that the overall measurement properties of the instrument were remarkably similar across male and female respondents. Taken together these results provide compelling evidence of the psychometric and measurement characteristics of this instrument.

Though these results are encouraging, it is important to acknowledge the limitations of this study. First, it is important to acknowledge that the instrument confounds both direct and indirect experiences of sexual orientation harassment. Recall that the instrument asked respondents “How often during the past 12 months have you experienced or witnessed event(s)/behavior(s) involving military personnel . . . because of perceived homosexuality?” Thus, respondents may have witnessed (i.e., had indirect experience with sexual orientation harassment) but not experienced any of the event(s) or behavior(s) assessed (i.e., had direct experience with sexual orientation harassment). This distinction is important because estimates derived from this instrument may not accurately depict the true incidence of these event(s)/behavior(s) for actual targets of this form of harassment within the military workplace. Though this is an important methodological limitation, it is worth noting that the estimates derived using this instrument provide clear evidence that sexual orientation harassment is a problem within the military workplace. Secondly, it is important to note that the sexual orientation of the participants was not known. Accordingly, estimates of the prevalence of sexual orientation harassment utilizing this instrument may not accurately capture actual harassment experiences of sexual minority personnel. Although it may be desirable for future researchers to ask respondents to provide information concerning their sexual orientation, we recognize that disclosure of sexual orientation among
gay service personnel is a complex matter. Finally, our findings must be tempered by the fact that the sampling strategy may limit the generalizability of the findings. Random selection occurred with respect to military installations and units and not at the level of the individual. Though it is hard to imagine that different results may be obtained with a random sample from the military population, we cannot rule this out as a plausible rival hypothesis.

These limitations notwithstanding, our study begins to address the need for psychometrically sound measures of sexual orientation harassment (Croteau, 1996; Ragins, 2004). Our methodological approach provides a rigorous assessment of the psychometric and measurement properties of this instrument. To our knowledge, no studies have examined the characteristics of this instrument; nor have other studies employed IRT/DTF methodologies to examine item and scale characteristics within the sexual orientation harassment literature (Ragins, 2004) or the broader literature on workplace harassment (for an exception, see Donovan & Drasgow, 1999). These results provide strong support for the psychometric and measurement equivalence of the instrument and allow for greater confidence in future exploration of relational equivalence between men and women’s harassment reports.

REFERENCES


