Measurement of Teen Dating Violence Attitudes: An Item Response Theory Evaluation of Differential Item Functioning According to Gender

Maria Orlando Edelen, Daniel F. McCaffrey, Grant N. Marshall and Lisa H. Jaycox

*J Interpers Violence* 2009 24: 1243 originally published online 13 August 2008
DOI: 10.1177/0886260508322187

The online version of this article can be found at:
http://jiv.sagepub.com/content/24/8/1243
Accurate assessment of attitudes about intimate partner violence is important for evaluation of prevention and early intervention programs. Assessment of attitudes about cross-gender interactions is particularly susceptible to bias because it requires specifying the gender of the perpetrator and the victim. As it is likely that respondents will tend to identify with the same-gender actor, items and scales assessing attitudes about intimate partner violence may not have equivalent measurement properties for male and female respondents. This article examines data from 2,575 high school students who participated in a teen-dating violence intervention study. The majority of participants were Latino (91%), and the sample was nearly evenly split with respect to gender (51% female). Items from two scales (boy-on-girl violence, 4 items; girl-on-boy violence, 5 items) reflecting teens’ attitudes about dating violence were calibrated with the graded item response theory (IRT) model and evaluated for differential item functioning (DIF) by gender. A total of three items, two from the girl-on-boy violence scale and one from the boy-on-girl violence scale, were identified as functioning differently for girls and boys. In all cases where DIF was detected, the item’s attitudinal statement was easier to accept for the gender group that was portrayed as victim rather than perpetrator. For both scales, accounting for the identified DIF influenced inferences about the magnitude of mean differences in attitudes between boys and girls. These results support the use of IRT scores that account for DIF to minimize measurement error and improve inferences about gender differences in attitudes about dating violence.

**Keywords:** assessment; domestic violence; dating violence; domestic violence intervention/treatment
The negative impact of intimate partner violence on public health is well documented (Tjaden & Thoennes, 2000), and there is growing consensus that development of effective prevention and early intervention programs is crucial. To date, such programs have attempted to change participants’ knowledge about intimate partner violence, attitudes about violence and aggression, and behavior (Whitaker et al., 2006). As these programs get evaluated for their impact, it is essential that we be able to accurately measure outcomes such as attitudes about violence. Current attitude measurement tools lag behind intervention and prevention development, hampering efforts to evaluate these programs.

One issue not commonly addressed in the assessment of attitudes about intimate partner violence is the possibility that men and women may interpret items about cross-gender interactions differently. Social learning theory posits that attitudes and behavior develop via personal experience as well as the influence of family, peer, and cultural norms (Bandura, 1973, 1977; Garbarino, Schellenback, & Sebes, 1986; Wekerle & Wolfe, 1999), and thus teens would be expected to reflect back many of those norms in assessments. Many empirical studies have documented that boys hold more accepting attitudes about violence as compared with girls (e.g., Avery-Leaf, Cascardi, O’Leary, & Cano, 1997; Foshee, 1996; Lavoie, Vezina, Piche, & Boivin, 1995; Malik, Sorenson, & Aneshensel, 1997; Tontodonato & Crew, 1992; Ulloa, Jaycox, Marshall, & Collins, 2004). However, whether these differences on assessments reflect a true difference in attitudes, or whether boys and girls may be interpreting the items differently by virtue of their gender, has not been explored. In particular, assessment of attitudes about cross-gender interactions is difficult because it requires specifying the gender of the perpetrator and the victim (e.g., Huesmann & Guerra, 1997), and it is likely that respondents will tend to identify with the same-gender actor. For example, one measure of teen dating violence attitudes includes an item asking the respondents to indicate the extent to which they agree that boys sometimes deserve to be hit by the girls they date (Foshee et al., 1996). A girl who considers this item is likely to approach it with the mindset of the perpetrator, whereas a boy is likely to self-identify with the victim.

This difference in identification depending on gender may lead to differences in item interpretation as well as different motivations for choosing responses, which in turn can result in items and scales that do not behave

Authors’ Note: Please address correspondence to Maria Orlando Edelen, RAND Corporation, 1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138; e-mail: maria_Edelen@rand.org.
the same for men and women. More generally, cultural norms and other factors might result in men and women interpreting any item about violence differently, even those where the genders of the perpetrator and victim are not specified. When items and scales do not have equivalent measurement properties across genders, they are said to lack measurement invariance with respect to gender (Reise, Widaman, & Pugh, 1993). If the assumption of measurement invariance (which is usually tacitly assumed) does not hold, inferences about group differences can be biased and lead to potentially erroneous conclusions.

Lack of measurement invariance between genders has been observed with a wide range of scales including mathematics achievement (Mendes-Barnett & Ercikan, 2006), physical self-concept (Fletcher & Hattie, 2005), and mental health (Cauffman & MacIntosh, 2006). An intuitive example is the assessment of crying spells as a symptom of depression (Gelin & Zumbo, 2003). Measurement invariance for this item would require that men and women with the same amount of depression are equally likely to have crying spells. However, independent of depression status, women are more likely to cry than men. Thus a woman’s endorsement of the crying symptom is less indicative of depression than a man’s endorsement of that same symptom. Failure to account for this imbalance results in scores that contain measurement bias, confounding interpretation of observed gender differences in depression. Identification of this noninvariance can improve accuracy of measurement when examining gender differences by removing bias and can provide insight into the ways in which individuals interpret or respond to items differently because of their gender.

One common conceptualization of the nonequivalence of scales focuses on the presence of statistical item bias or differential item functioning (DIF). DIF refers to items that have different measurement properties for various subgroups (e.g., genders) after controlling for the overall differences between subgroups on the construct being measured (Holland & Wainer, 1993). Essentially, DIF captures an additional identifiable source of variation associated with the item. An item is said to exhibit DIF if two respondents from distinct subgroups who have equal levels of the psychological trait being measured do not have the same probability of endorsing each response category of that item. For example, a female student may have more difficulty than a male student with a math word problem that contains the word “girder” even though the two students have equal math ability. If the DIF in the “girder” item is not accounted for, the female student would receive a score on the test that is lower than her true math ability.
The practical result at the item level is that responses on an item exhibiting DIF are not equivalent across the groups being studied, leading to potentially misleading group differences and inaccurate bivariate associations involving the items that exhibit DIF. At the level of the scale, the effect of the presence of DIF items within a scale can vary depending on the degree of DIF, the number of items in the scale exhibiting DIF, and the proposed uses of the scale.

A number of approaches to detecting DIF have been developed employing methods from both classical test theory (CTT) and item response theory (IRT; Holland & Wainer, 1993). CTT is a traditional approach to measurement based on the concepts of observed score, true score, and error score. Although CTT methods require few assumptions and are relatively easy to implement, results of these applications are sample specific and as such are not sufficient for ensuring measurement invariance (Budgell, Raju, & Quartetti, 1995; Hulin, Drasgow, & Parsons, 1983). IRT comprises a collection of parametric modeling techniques for scaling item responses that offers many advantages over traditional test theory and can be a powerful tool for constructing and evaluating attitudinal and other measurements (Embretson, 1996).

One of the most basic assumptions of the application of parametric IRT models is that the model is appropriate for the data. Another critical assumption of unidimensional parametric IRT models is that the construct being measured is in fact unidimensional; that is, that the covariance among the items can be explained by a single underlying dimension. When the assumptions for an IRT application are met, the IRT approach to DIF detection offers several advantages. Most notably, results from an IRT application generalize beyond the sample being studied to the population it represents. The IRT approach also offers graphical representations of DIF that are valuable diagnostic tools for evaluating the potential impact of DIF both at the item level and at the level of the entire scale. Additionally, once DIF is detected, the IRT models can be used to generate scores that account for DIF, thereby providing accurate and comparable scores on measures of constructs of interest for all groups. These scores can be compared with raw scores in sensitivity analyses to evaluate the practical impact of DIF on the application of interest.

The basis of IRT is the item characteristic curve (ICC), or trace line, which models the relationship between a person’s response to an item and his or her level on the latent construct measured by the scale (e.g., acceptance of violence). For items with dichotomous response options, the two parameter logistic (2PL) model is often applied. This model specifies the
ICC as a logistic function described by the location \((b)\) and slope \((a)\) parameters. The \(b\) parameter (also called the threshold parameter) is the point along the ICC at which the probability of a positive response for a dichotomous item is 50%. The larger the location parameter, the more of the measured construct a respondent must have to endorse that item. The \(a\) parameter (also called the discrimination parameter) represents the slope of the ICC at the value of the location parameter and indicates the extent to which the item is related to the underlying construct. A steeper slope indicates a closer relationship to the construct and therefore a more discriminating item. An item displays DIF in an IRT analysis if there are significant differences in either the item’s location or slope parameter estimates, according to group membership.

Previous studies have demonstrated that IRT can be useful in examining the structure of measures of violence (Michie & Cook, 2006; Regan, Bartholomew, Oram, & Landolt, 2002; Schafer, 1996). In this article, we use an IRT framework to evaluate gender DIF in scales assessing attitudes about dating violence. Data were collected as part of an intervention study of teen violence (Jaycox et al., 2006). As is described in more detail below, we first tested each of the items for DIF according to gender and interpreted any identified DIF. Then we evaluated the impact of the identified DIF on inferences about group differences by comparing results using scores that accounted for the DIF with results using scores that did not account for the identified DIF. Implications for this study in particular and for the measurement of dating violence outcomes in general, are discussed.

Method

Data and Sample

**Participants.** Data are from a study evaluating a dating violence prevention curriculum. This evaluation employed an experimental design with random assignment by cluster of 9th grade Health classes in Los Angeles United School District. Eligible high schools were those serving majority Latino populations. Clusters were assigned to an immediate or delayed intervention (control group) and both groups were followed for 6 months. We obtained active parental consent for participation in the research project and obtained student assent on each day of survey administration. This study uses responses from the first survey administration prior to any student receiving the intervention, for a sample of \(N = 2,575\). The sample
was nearly evenly split according to gender with 1,263 males and 1,312 females. The majority (91%) of the participants were Latino, with an average age of 14.5 years ($SD = 1.04$). More information about the study and participants can be found in Jaycox et al. (2006).

**Measures.** Analyses involve two outcome measures reflecting adolescents’ attitudes about aggression in dating situations (Orlando, Jaycox, McCaffrey, & Marshall, 2006). The survey instrument included the Prescribed Norms scale (Foshee et al., 1996) and the six items on cross-gender retaliation from the NOBAGS scale (Huesmann & Guerra, 1997). These scales were chosen from measures used in prior studies on dating violence (e.g., Foshee et al., 1998) to match the attitudes the program under evaluation was attempting to change (acceptance of aggression). The Prescribed Norms scale asks respondents to indicate on a 4-point scale (1 = *strongly agree* to 4 = *strongly disagree*) their extent of agreement with statements about dating violence (e.g., “Boys sometimes deserve to be hit by the girls they date”). The NOBAGS items ask participants to indicate on a 4-point scale (1 = *really wrong* to 4 = *perfectly ok*) the extent to which the response to a situation was acceptable (e.g., “Suppose a girl says something bad to a boy, do you think it’s wrong for the boy to scream at her?”). The Prescribed Norms items were reverse-scored for analyses so that a higher score indicated more acceptance of violence for all items.

Using 9 of the 15 items on the Prescribed Norms and NOBAGS scales, two distinct outcome measures were derived, acceptance of female-perpetrated violence on males (girl-on-boy; GB; 5 items; Cronbach’s $\alpha = .71$) and acceptance of male-perpetrated violence on females (boy-on-girl; BG; 4 items; Cronbach’s $\alpha = .55$). Table 1 lists the items in each outcome measure with the proportion of male and female responses in each category.

**Analytic Approach**

In what follows we provide a brief description of the IRT model and the IRT-DIF detection and evaluation approach. More information about IRT is provided elsewhere (e.g., Hambleton & Swaminathan, 1985) and a detailed step-by-step description of the IRT-DIF detection approach can be found in Edelen, Thissen, Teresi, Kleinman, and Ocepek-Weliikson (2006).

The two outcome measures were examined separately using Samejima’s (1997) graded IRT model. Each ICC of Samejima’s graded IRT model specifies the probability of choosing a response as a function of one item slope parameter ($a$) and a location ($b$) parameter for each response category.
### Table 1
Boys’ and Girls’ Proportions of Responses to Items Comprising the Two Dating Violence Attitude Scales (N = 2,575)

<table>
<thead>
<tr>
<th>Response Category</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Girl-on-boy violence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GB1&lt;sup&gt;a&lt;/sup&gt; Boys sometimes deserve to be hit by girls they date.</td>
<td>56.0</td>
<td>25.9</td>
<td>13.5</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>62.6</td>
<td>22.4</td>
<td>11.4</td>
<td>3.5</td>
</tr>
<tr>
<td>GB2&lt;sup&gt;a&lt;/sup&gt; It is OK for a girl to hit a boy if he hit her first.</td>
<td>43.3</td>
<td>22.6</td>
<td>17.3</td>
<td>16.8</td>
</tr>
<tr>
<td></td>
<td>45.0</td>
<td>26.1</td>
<td>16.1</td>
<td>12.8</td>
</tr>
<tr>
<td>GB3&lt;sup&gt;b&lt;/sup&gt; Suppose a boy says something bad to a girl, do you think it is wrong for her to scream at him?</td>
<td>17.8</td>
<td>46.2</td>
<td>27.8</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>19.3</td>
<td>45.0</td>
<td>25.7</td>
<td>10.1</td>
</tr>
<tr>
<td>GB4&lt;sup&gt;b&lt;/sup&gt; Suppose a boy says something bad to a girl, do you think it is wrong for her to hit him?</td>
<td>64.3</td>
<td>24.4</td>
<td>8.9</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>66.5</td>
<td>23.0</td>
<td>7.4</td>
<td>3.1</td>
</tr>
<tr>
<td>GB5&lt;sup&gt;b&lt;/sup&gt; Suppose a boy hits a girl, do you think it is wrong for her to hit him back?</td>
<td>25.3</td>
<td>15.7</td>
<td>25.7</td>
<td>33.4</td>
</tr>
<tr>
<td></td>
<td>28.5</td>
<td>19.3</td>
<td>24.4</td>
<td>27.8</td>
</tr>
<tr>
<td><strong>Boy-on-girl violence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG1&lt;sup&gt;a&lt;/sup&gt; It is OK for a boy to hit a girl if she hit him first.</td>
<td>74.8</td>
<td>17.2</td>
<td>4.8</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>77.9</td>
<td>16.5</td>
<td>3.8</td>
<td>1.8</td>
</tr>
<tr>
<td>BG2&lt;sup&gt;b&lt;/sup&gt; Suppose a girl says something bad to a boy, do you think it is wrong for him to scream at her?</td>
<td>31.9</td>
<td>51.6</td>
<td>13.8</td>
<td>2.7</td>
</tr>
<tr>
<td></td>
<td>37.4</td>
<td>46.7</td>
<td>14.0</td>
<td>1.9</td>
</tr>
<tr>
<td>BG3&lt;sup&gt;b&lt;/sup&gt; Suppose a girl says something bad to a boy, do you think it is wrong for him to hit her?</td>
<td>95.7</td>
<td>3.1</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>96.6</td>
<td>2.0</td>
<td>0.4</td>
<td>1.0</td>
</tr>
<tr>
<td>BG4&lt;sup&gt;b&lt;/sup&gt; Suppose a girl hits a boy, do you think it is wrong for him to hit her back?</td>
<td>77.4</td>
<td>14.6</td>
<td>5.2</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>73.3</td>
<td>17.6</td>
<td>6.8</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Note: For each item, boys’ response proportions are listed above, followed below by girls’ response proportions. Response categories for items with superscript a are 1 = Strongly Disagree, 2 = Disagree Somewhat, 3 = Agree Somewhat, 4 = Strongly Agree. Response categories for items with superscript b are 1 = Really Wrong, 2 = Sort of Wrong, 3 = Sort of OK, 4 = Perfectly OK.
transition. Thus for items with four response categories (as in this application) three location parameters are estimated.

The scales were evaluated for gender DIF using the likelihood ratio approach implemented with the freeware program IRTLRDIF (Thissen, 2001). We used an iterative purification process (e.g., Baker, Al-Karni, & Al-Dosary, 1991; Raju, van der Linden, & Fleer, 1995) to identify items that were free of gender DIF. These items, commonly referred to as “anchor items,” served as the basis to link the groups and estimate the group mean difference (Embretson, 1996). Once the set of anchor items for each scale was established, each of the remaining items was tested for DIF relative to the now-specified anchor items. The statistical significance tests for DIF were adjusted for multiple comparisons using the Benjamini–Hochberg adjustment (Benjamini & Hochberg, 1995) and tests used an overall alpha level of .01. Although these criteria for significance are rather stringent, it is appropriate to control for multiple comparisons because we had no a priori hypothesis about which items and parameters would exhibit DIF.

On the basis of the DIF analyses, we used MULTILOG (Thissen, 1991) to estimate a version of Samejima’s graded IRT model that allowed item parameters with DIF to vary by gender for each measure. Using the parameters from these final models, we also generated plots of category, item, and scale response functions separately for girls and boys to obtain a visual representation of the effect size of the identified gender DIF.

We performed two distinct checks to determine whether the final IRT model was an adequate reflection of the data. First, we visually examined the concordance between the observed data and the model-predicted category responses for each item using bar charts. Strong concordance in these charts would imply that the model is able to re-create the data. Second, we examined nonparametric models of the item responses using TESTGRAPH (Ramsay, 1993) to evaluate whether the functional form imposed by the parametric IRT model was reasonable. In this examination, similarity between the shape of the category response functions from the parametric and nonparametric analyses would lend support for the IRT model.

Comparing scoring approaches. One of the advantages to using IRT to evaluate DIF is that observed DIF can be accounted for in a final IRT model and used to generate scores. However, the impact of identified DIF at the score level is frequently so minimal that it is not worthwhile to adjust for it with this complex scoring approach. To determine whether the observed DIF in this study was influencing inferences about group differences, we
generated three distinct sets of scores for each scale and evaluated \( t \) tests comparing males and females for each of the outcomes using each of the score types. Specifically, we generated observed scores (OBS) equal to the average of the items in the measure. We also generated two types of IRT scores. One type was based on IRT models that accounted for the identified DIF (IRT-D) and a second was based on IRT models that ignored the identified DIF (IRT-ND). Comparisons among these three sets of scores allowed us to separate the impact of IRT scaling from the impact of accounting for the DIF. That is, differences in inferences based on analyses using IRT-ND and OBS scores would reflect the influence of IRT alone, whereas differences in inferences based on analyses using IRT-ND and IRT-D scores would reflect the influence of controlling for DIF.

**Results**

**IRT DIF Evaluation and Model Calibration**

Table 2 shows the results of the DIF analyses for the two outcome measures. For the girl-on-boy outcome, we identified 3 anchor items (items GB1, GB3, and GB4) for which there was no strong evidence of DIF. Items GB2 and GB5 displayed significant \( b \) DIF—the location parameters for boys and girls were not the same for each of these items. For the boy-on-girl outcome, we identified 2 anchor items (items BG1 and BG3). Item BG4 displayed significant \( b \) DIF. Although there was sufficient evidence of DIF in item BG2 to preclude it from being treated as an anchor item, the final test for DIF did not reach statistical significance. More details on the nature of the DIF for items GB2, GB5, and BG4 are presented below.

The next step was to estimate the parameters of the final model for each measure using MULTILOG. The item parameter estimates and their standard errors from these final calibrations are reported in Table 3. The items for both measures tended to be fairly discriminating with all but one slope estimate exceeding 1.2. In this context, a slope of 1 corresponds to a factor loading of about .5, and a slope of 2 to a factor loading of about .76 (Takane & de Leeuw, 1987). The broad range of \( b \)-parameter values indicates that the items, particularly in the girl-on-boy measure, represent an extensive segment of the underlying acceptance of violence continua they are measuring. The location parameters for each outcome are estimated relative to a mean of 0 for the boys.
Table 2
Results of DIF Analyses for Girl-on-Boy and Boy-on-Girl Violence

<table>
<thead>
<tr>
<th>Item</th>
<th>Test for item DIF (4 df)</th>
<th>Test for a DIF (1 df)</th>
<th>Test for b DIF (3 df)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>−2LL</td>
<td>p</td>
<td>−2LL</td>
</tr>
<tr>
<td>GB2</td>
<td>17.9</td>
<td>.001*</td>
<td>3.5</td>
</tr>
<tr>
<td>GB5</td>
<td>18.5</td>
<td>.001*</td>
<td>3.9</td>
</tr>
</tbody>
</table>

Boy-on-Girl Violence (BG1, BG3 Are Anchor Items)

<table>
<thead>
<tr>
<th>Item</th>
<th>Test for item DIF (4 df)</th>
<th>Test for a DIF (1 df)</th>
<th>Test for b DIF (3 df)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>−2LL</td>
<td>p</td>
<td>−2LL</td>
</tr>
<tr>
<td>BG2</td>
<td>11.1</td>
<td>.025</td>
<td>na</td>
</tr>
<tr>
<td>BG4</td>
<td>19.6</td>
<td>.001*</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Note: −2LL is the −2 × log-likelihood value for the nested model comparison and is distributed as $\chi^2$ with the specified df.
*Statistically significant at $p < .01$ after controlling for multiple tests with the Benjamini–Hochburg adjustment.

Table 3
Final Item Parameter Estimates and Their Standard Errors for the Two Measures

<table>
<thead>
<tr>
<th>Item</th>
<th>a</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Girl-on-boy violence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GB1</td>
<td>1.10 (0.05)</td>
<td>0.36 (0.06)</td>
<td>1.72 (0.11)</td>
<td>3.30 (0.20)</td>
</tr>
<tr>
<td>GB3</td>
<td>1.27 (0.06)</td>
<td>−1.55 (0.08)</td>
<td>0.53 (0.05)</td>
<td>2.22 (0.11)</td>
</tr>
<tr>
<td>GB4</td>
<td>1.91 (0.09)</td>
<td>0.47 (0.04)</td>
<td>1.61 (0.07)</td>
<td>2.62 (0.12)</td>
</tr>
<tr>
<td>GB2 – girls</td>
<td>1.72 (0.06)</td>
<td>−0.28 (0.05)</td>
<td>0.68 (0.06)</td>
<td>1.51 (0.08)</td>
</tr>
<tr>
<td>GB2 – boys</td>
<td>1.72 (0.06)</td>
<td>−0.25 (0.05)</td>
<td>0.55 (0.06)</td>
<td>1.34 (0.07)</td>
</tr>
<tr>
<td>GB5 – girls</td>
<td>1.57 (0.05)</td>
<td>−0.96 (0.06)</td>
<td>−0.20 (0.06)</td>
<td>0.76 (0.06)</td>
</tr>
<tr>
<td>GB5 – boys</td>
<td>1.57 (0.05)</td>
<td>−0.98 (0.07)</td>
<td>−0.34 (0.06)</td>
<td>0.62 (0.06)</td>
</tr>
<tr>
<td>Boy-on-girl violence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG1</td>
<td>1.58 (0.09)</td>
<td>0.96 (0.06)</td>
<td>2.16 (0.11)</td>
<td>2.95 (0.17)</td>
</tr>
<tr>
<td>BG2</td>
<td>0.86 (0.06)</td>
<td>−0.95 (0.09)</td>
<td>2.06 (0.15)</td>
<td>4.64 (0.34)</td>
</tr>
<tr>
<td>BG3</td>
<td>1.66 (0.18)</td>
<td>2.53 (0.18)</td>
<td>3.31 (0.28)</td>
<td>3.55 (0.32)</td>
</tr>
<tr>
<td>BG4 – girls</td>
<td>2.54 (0.10)</td>
<td>0.55 (0.04)</td>
<td>1.40 (0.06)</td>
<td>2.22 (0.12)</td>
</tr>
<tr>
<td>BG4 – boys</td>
<td>2.54 (0.10)</td>
<td>0.90 (0.05)</td>
<td>1.72 (0.07)</td>
<td>2.41 (0.12)</td>
</tr>
</tbody>
</table>
Assessment of fit strongly supported the use of the IRT models. For both outcome measures, the model-predicted category responses for each item corresponded very closely with the observed frequencies, lending support for the models. The nonparametric item response functions were also reasonably similar to their model-based counterparts.

Figure 1 shows two model-based sets of curves for girls and boys for item GB5 (“Suppose a boy hit a girl, do you think it’s ok for the girl to hit him back?”) from the acceptance of girl-on-boy violence scale. For each gender, the top panel displays the item characteristic curves, the probability of choosing each category as a function of the latent trait being measured by the item set, acceptance of girl-on-boy violence; the bottom panel shows the item response function, the expected item score as a function of acceptance of girl-on-boy violence. The three location parameters for this item are all lower for boys than for girls (see parameter values in Table 3). Thus the DIF in this item indicates that, given the same level of overall acceptance of girl-on-boy violence, boys are slightly more accepting than girls of girls retaliating by hitting boys who have hit them first. That is, among boys and girls with the same acceptance of violence, boys were less likely than girls to respond that it is “really wrong” and more likely to say it is “perfectly ok” for a girl to retaliate and hit a boy who hit her first. Item GB2 from the girl-on-boy measure also shows this pattern of DIF (see parameter values in Table 3).

In contrast, the observed DIF in item BG4 from the boy-on-girl measure (“If a girl hits a boy, do you think it’s OK for the boy to hit her back?”) indicates that controlling for the overall acceptance of boy-on-girl violence, girls are slightly more accepting than boys of this behavior. That is, girls were less likely to say it is “really wrong” and more likely to say it is “perfectly ok” for a boy to hit a girl if she hit him first. It is interesting to note that in all cases where DIF was detected, the effect indicated that the item was easier to endorse for the gender group that was portrayed as victim rather than perpetrator.

Figure 2 displays girls’ and boys’ expected total scores for each measure as a function of the underlying trait (overall acceptance of girl-on-boy violence, top panel, and overall acceptance of boy-on-girl violence, bottom panel). The expected total scores differ for boys and girls because this relationship is generated based on the IRT models that specify separate item parameters for the DIF items (i.e., the parameters in Table 3). If the model did not account for DIF, then the two lines would be completely coincident. Modeling the DIF in the girl-on-boy measure resulted in higher expected total scores for boys relative to girls. The effect is slight, but constant,
Figure 1
Girls’ and Boys’ Category and Item Response Functions for Item GB5 From the Girl-on-Boy Measure

GB5 girls (solid) and boys (dotted)

Acceptance of Girl-on-Boy Violence

GB5 girls (solid) and boys (dotted)

Acceptance of Girl-on-Boy Violence

Downloaded from jiv.sagepub.com by Barbara Ball on August 20, 2012
Figure 2
Girls’ (Solid Lines) and Boys’ (Dotted Lines) Response Functions for Each of the Measures
across the latent continuum of acceptance of girl-on-boy violence. The effect of DIF on the expected total score for the boy-on-girl measure is in the opposite direction; for this outcome, modeling the DIF resulted in higher expected total scores for girls compared with boys. The effect is small, but discernable, for acceptance of boy-on-girl violence latent scores between the values of 0 and 2.

Comparing Scoring Approaches

A final set of descriptive analyses was conducted to inform the decision regarding the most appropriate type of scores to use in future analyses, the observed (OBS), the DIF-adjusted IRT scores (IRT-D), or IRT scores that ignore the DIF (IRT-ND).

Results of the $t$ tests evaluating group mean differences for each of the score types are presented in Table 4. For the girl-on-boy measure, comparisons between genders are significant regardless of score type. In all cases, boys display a slightly higher acceptance of girl-on-boy violence than do girls. However, the gender difference appears to be highly significant for the two score types that do not account for the observed DIF (OBS and IRT-ND) whereas the effect is diminished considerably, though still significant, when using the IRT-D scores that adjust for DIF. Although for this measure the DIF adjustment does not alter inferences about group differences, the DIF analysis indicates that among boys and girls with equal overall acceptance of girl-on-boy violence, boys are more likely to positively endorse two items on girls hitting boys. When this difference is ignored in the OBS and IRT-ND measures, the DIF is conflated with true differences in the genders on the underlying construct and these unadjusted measures exaggerated gender differences.

For the boy-on-girl violence measure, the effect of score type on gender differences is more pronounced and does alter inferences. The two score types that do not account for DIF result in no gender difference in attitudes about boy-on-girl violence. However, the inference based on the IRT-D score that models the observed gender DIF indicates that boys have a significantly higher acceptance than girls of boy-on-girl violence. In this scale, among boys and girls with equal overall acceptance of boy-on-girl violence, girls are more likely to positively endorse two items accepting boys’ retaliation against girls, resulting in the unadjusted OBS and IRT-D scales exaggerating the overall acceptance of boy-on-girl violence by girls and masking the true gender differences. Correcting for the DIF reveals the significant gender differences.
Measurement of teen dating violence attitudes poses difficult challenges because social learning theory suggests the measurement properties of dating violence attitude scales may not be equivalent for male and female youths. In particular, items assessing attitudes about dating violence must necessarily describe scenarios involving cross-gender interactions. When the genders of the victim and perpetrator of violent or aggressive behaviors are specified, respondents are likely to identify with the same-gender role. The difference in identification may lead to different response processes for male and female respondents and lack of item invariance according to gender. Identifying and accounting for this potential noninvariance are essential when examining outcomes associated with dating violence intervention efforts and determining whether potential group differences are reflected by actual differences in the constructs being measured. In short, accurate measurement of constructs of interest is critical for providing scientific input to inform decisions of policymakers and prevention and intervention specialists on how best to influence change.

IRT provides a useful framework for the evaluation of measurement invariance through DIF analyses. The parametric assumptions of the IRT framework allow for the estimation of different item parameters for different groups, provided the groups can be linked with one or more items that are free of DIF. In the IRT framework, any significant lack of measurement invariance that is...
identified can easily be examined and accounted for in both the parameter estimates and resultant scores. In this article, we used IRT methods to identify DIF, evaluate its impact, and generate adjusted scores that accounted for the identified DIF. We found some invariant items that could act as anchor items and other items that appear to be interpreted somewhat differently by boys and girls. Given that societal norms are more lenient about female-on-male violence than male-on-female violence (e.g., Price & Byers, 1999), it is not surprising that boys and girls might respond to some items differently based on looking at the problem from differing perspectives.

For the girl-on-boy violence scale, we found that three items behaved similarly across gender, whereas two items displayed DIF. The two items that exhibited DIF had very similar content pertaining to a girl hitting a boy in retaliation after being hit. The DIF in these items indicated that given the same underlying level of acceptance of girl-on-boy violence, boys were more accepting than girls of this particular behavior. The other three items, which referred to other types of retaliatory violence directed by girls against boys (screaming at or hitting in response to the boy saying something bad, or hitting because the boy deserved it), had the same measurement properties for boys and girls.

Results were similar for the boy-on-girl violence scale, although only one of the two hitting in retaliation items displayed DIF. In this case, girls were more accepting than boys (with the same underlying level of acceptance of boy-on-girl aggression) of boys hitting in retaliation. The three items that did not display DIF were the other hitting in retaliation item and screaming at or hitting in response to the girl saying something bad. The distinction between the two hitting in retaliation items is that the one that displayed DIF asked whether the respondent thinks it is wrong (with answer choices from really wrong to perfectly ok), whereas the one that did not show DIF asked for an extent of agreement with this behavior being ok. Perhaps the inclusion of the word wrong activated the cultural message that boys hitting girls is wrong no matter what. This message is strongly instilled in boys and may explain why this item, and not the other hitting in retaliation item, was interpreted differently according to gender.

Taken together, the DIF results indicate that adolescents may feel more accepting of retaliatory hitting behavior when the depicted victim is of their own gender. Assuming respondents are identifying with the actor of their own gender, they appear to be more forgiving of their partner than of themselves. This bias could lead to underreporting of dating violence or minimization of the seriousness of the incident. Intervention studies may be able to reduce this bias by being aware that adolescents can tend to be
forgiving of their partner and modifying intervention curriculum to directly address this.

In this study, the capacity of IRT to account for DIF proved useful in increasing measurement accuracy and evaluating hypotheses about basic gender differences, differential program impact on boys and girls, and interactions with gender. In particular, accounting for DIF revealed that boys were not only more accepting than girls of girl-on-boy violence but also of boy-on-girl violence. This troubling finding, which would have been missed in a less thorough evaluation of the item responses, is consistent with social learning theory as well as previous empirical work (see Ulloa, Jaycox, Marshall, & Collins, 2004, for a review). Although the full implications of these gender differences are somewhat unclear because the overall acceptance of any violent act tended to be low for both boys and girls, these differences do suggest a potential leverage point for prevention and intervention programs targeting boys. These results also indicate that both male-on-female and female-on-male violence need to be addressed.

The goal of most prevention programs is to change students’ attitudes about dating and intimate partner violence. These interventions might not only affect the overall attitudes but also might differentially affect responses to individual items and DIF. Hence, studies of measurement invariance and DIF from assessments before and after an intervention might improve estimation of the effects of prevention programs. In this study, we report only on measures prior to the intervention. Nonetheless, the methods applied here could be extended to more general settings by identifying anchor items that are invariant across both groups and time. Using these anchor items, it would then be possible to estimate models with item parameters that are specific to both gender and time of the assessment.

The generalizability of our findings is limited in several respects. The study sample consists of primarily Latino high school students and the analyses examined only a small set of dating violence attitude items. Future research is required to determine whether these findings extend to populations of different ethnicities and age ranges. Additional investigation should also examine whether similar noninvariance trends are evident in other measures of attitudes toward dating violence.

Although we were able to adjust for gender DIF in this study, it is still the case that at least some commonly used measures to assess attitudes related to dating violence are not optimal. To address this issue, further research to improve the assessment of dating violence should consider non-standard approaches to item generation. To understand the nuances in teen attitudes toward cross-gender aggression, for example, our research team...
has begun work that allows teens to articulate their thoughts about dating violence scenarios as they unfold. Preliminary examination of the data collected to date shows us that (a) a variety of factors influence teens’ perceptions of dating violence (e.g., whether the perpetrator is male or female, a friend or a stranger, gender of respondent) and (b) teens express approval and disapproval of dating violence in a number of ways (Rayburn et al., 2007). The insights provided by soliciting adolescents’ thoughts in the moment may be translated into more appropriate and discriminating items that teens will respond to with greater variability.

In addition to considering novel approaches such as this, future development of assessments of attitudes about intimate partner violence should include evaluation of measurement invariance according to gender during instrument construction. Whether the developer wants to exclude variant items or include them and adjust for them in the scoring, this practice will provide useful information and minimize biased inferences about gender differences in this important outcome.

Notes

1. The analyses reported in Jaycox et al. (2006) use a sample of 2,540 students because those analyses exclude blocks of schools (including both those randomized to intervention or control condition) where at least one school failed to provide data or complete the program as assigned. The responses from all students provide information on the measurement model and are included in the current application.

2. Preliminary analyses indicated that five of the items from the Prescribed Norms scale had very little variance (i.e., >90% endorsement of a single response category) and were not useful in measuring acceptance of either girl-on-boy or boy-on-girl violence; those items are not used in the analyses presented here.

References


Maria Orlando Edelen, PhD, is a quantitative psychologist and an assistant professor at Brown University Medical School, department of psychiatry and human behavior at the University of North Carolina at Chapel Hill. She has extensive knowledge of test theory including item response theory and of advanced multivariate analysis methods such as structural equation modeling and latent growth mixture modeling. She has applied her measurement and methodological skills in a variety of health research contexts.

Daniel F. McCaffrey, PhD, is a senior statistician and head of the RAND Statistics Group. His current research interests focus on methods for the analysis of longitudinal data and estimation of value-added measures of teacher performance. Recently, he completed evaluations of school-based substance use prevention and dating violence prevention programs and studies on the effectiveness of adolescent substance abuse treatment programs.

Grant N. Marshall, PhD, received his doctorate from UCLA in clinical psychology and obtained postdoctoral training in survey research methods and psychiatric epidemiology from the Institute for Social Research at the University of Michigan. He has been a behavioral scientist at the RAND Corporation since 1992. His research interests include attitudes about violence, exposure to trauma, and psychological adjustment following life-threatening experiences.

Lisa H. Jaycox, PhD, obtained her doctorate from the University of Pennsylvania. She is a senior behavioral scientist at the RAND Corporation in Arlington, Virginia. Her research focuses on interventions for youth at risk for developing depression or posttraumatic stress disorder, with a focus on youth exposed to violence and trauma. In particular, she studies dissemination of effective interventions into community settings aimed at vulnerable youth. Her prior projects include prevention of depression in school settings, school-based interventions for youth exposed to community violence, linkages between trauma exposure and retention in substance abuse programs, and the study of Break the Cycle’s dating violence prevention and early intervention program highlighted in this article. Current projects include examination of depression treatment for adolescents in primary care settings, evaluation of programs designed to ameliorate the impact of violence on young children, and interventions for children exposed to disaster.