Published in final edited form as:

Psychol Assess. 2012 June ; 24(2): 444–454. doi:10.1037/a0025831.

Item Analysis and Differential Item Functioning of a Brief Conduct Problem Screen

Johnny Wu,

Department of Psychology, University of Washington

Kevin M. King,

Department of Psychology, University of Washington

Sarah Jensen Racz,

Department of Psychology, University of Washington

Katie Witkiewitz,

Department of Psychology, Washington State University

Robert J. McMahon, and

Department of Psychology, Simon Fraser University and the Child & Family Research Institute

Conduct Problems Prevention Research Group

Abstract

Research has shown that boys display higher levels of childhood conduct problems than girls, and Black children display higher levels than White children, but few studies have tested for scalar equivalence of conduct problems across gender and race. The authors conducted a 2-parameter item response theory (IRT) model to examine item characteristics of the Authority Acceptance scale from the Teacher Observation of Classroom Adaptation-Revised (AA-TOCA-R; L. Larsson-Werthamer, S. G. Kellam, & L. Wheeler, 1991) in 8,820 kindergarten children and estimated the degree of differential item functioning (DIF) by gender and race/urban status. The mean level of latent conduct problems was best represented by behaviors such as being stubborn, breaking rules, and being disobedient, whereas breaking things and taking others' property best represented the construct at one standard deviation above the mean. DIF by gender was detected, such that at equivalent levels of latent conduct problems, males received more endorsements of overt behaviors from teachers, whereas females received more endorsements of nonphysical behaviors. Moreover, overt behaviors were better discriminators of latent conduct problems for males, and nonphysical behaviors were better discriminators of latent conduct problems for females. Differences across race/urban status were not found to be conceptually meaningful. The authors' analyses also suggest that the item scaling of the AA-TOCA-R may be best represented by 5e

© 2011 American Psychological Association

Correspondence concerning this article should be addressed to Johnny Wu, University of Washington, Department of Psychology, Box 351525, Seattle, WA 98195-1525. johnnywu@u.washington.edu.

Members of the Conduct Problems Prevention Research Group, in alphabetical order, include Karen L. Bierman, Department of Psychology, Pennsylvania State University; John D. Coie, Department of Psychology, Duke University; Kenneth A. Dodge, Center for Child and Family Policy, Duke University; Mark T. Greenberg, Department of Human Development and Family Studies, Pennsylvania State University; John E. Lochman, Department of Psychology, the University of Alabama; Robert J. McMahon, Department of Psychology, Simon Fraser University and the Child & Family Research Institute; and Ellen E. Pinderhughes, Department of Child Development, Tufts University.

Karen L. Bierman, John D. Coie, Kenneth A. Dodge, Mark T. Greenberg, John E. Lochman, and Robert J. McMahon are the developers of the Fast Track curriculum and have a publishing agreement with Oxford University Press. Mark T. Greenberg is an author on the PATHS curriculum and has a royalty agreement with Channing-Bete, Inc. Mark T. Greenberg is a principal in PATHS Training, LLC. Robert J. McMahon is a coauthor of *Helping the Noncompliant Child* and has a royalty agreement with Guilford Publications, Inc.; he is also a member of the Treatments That Work Scientific Advisory Board with Oxford University Press.

categories instead of 6. These findings provide support for the use of IRT modeling to examine item characteristics of conduct problem scales and DIF to test for scalar equivalence across diverse subpopulations.

Keywords

conduct problems; Authority Acceptance Scale; Teacher Observation of Classroom Adaptation-Revised; item response theory; differential item functioning

Conduct problems are manifested in children's outward behavior and reflect the child acting negatively toward the external environment (Campbell, Shaw, & Gilliom, 2000). These behaviors include (but are not limited to) fighting, stealing, and breaking rules. Prospective studies have demonstrated that childhood conduct problems can predict juvenile delinquency and other negative outcomes in young adulthood (Farrington, 1991; Loeber & Hay, 1997; Moffitt, 1993). For example, teacher reports of conduct problems from boys at age 8 predicted substance abuse and psychotic disorders in early adulthood (Sourander et al., 2005). Moffitt, Caspi, Harrington, and Milne (2002) showed that boys with a pattern of childhood-onset conduct problems were at increased risk for substance dependence and mental health problems relative to boys who did not exhibit conduct problems. Moreover, Petras and colleagues (2005) found that girls with early onset of conduct problems were more likely to be violent during adolescence relative to girls without conduct problems.

Given that early childhood conduct problems pose a risk for negative outcomes during adolescence and young adulthood, early identification of at-risk children is critical (Lochman & Conduct Problems Prevention Research Group, 1995). Due to the time, organizational, and resource constraints existing in most schools, standard practice is to use a cross-sectional teacher assessment to identify at-risk children. One commonly used tool in identifying childhood conduct problems is the 10-item Authority Acceptance scale from the Teacher Observation of Classroom Adaptation-Revised (AA-TOCA-R; Werthamer-Larsson, Kellam, & Wheeler, 1991), a measure designed so that teachers could rate their students' conduct problem behaviors in the classroom. Teacher ratings on the AA-TOCA-R have predicted later sexual activity and substance abuse (Bradshaw, Schaeffer, Petras, & Ialongo, 2010), violence among adolescents (Petras, Chilcoat, Leaf, Ialongo, & Kellam, 2004), criminality in young adults (Schaeffer, Petras, Ialongo, Poduska, & Kellam, 2003), poor academic performance, and prosocial skill deficits (Flanagan, Bierman, Kam, & Conduct Problems Prevention Research Group, 2003). These results lend support for the utility of the AA-TOCA-R in identifying children at risk for later conduct problems.

Koth, Bradshaw, and Leaf (2009) found evidence for unidimensionality in the factor structure of the AA-TOCA-R and showed that among kindergarteners, boys had a higher overall mean than girls and Blacks had a higher overall mean than Whites (higher score = more frequent conduct problems). However, little other research has been conducted on the psychometric properties of the AA-TOCA-R. In particular, it is unknown whether the AA-TOCA-R, which has the potential to be widely deployed to thousands of classrooms as a potential screening device for high-risk children, exhibits scalar equivalence across different subgroups (e.g., gender or ethnicity).

Scalar equivalence is said to hold when scores on a measure represent the same levels of the construct across diverse subgroups (Van de Vijver & Leung, 1997). When the effects of subgroup factors are accumulated across items, the implications of overall scale scores can become distorted, and observed group differences may not reflect real differences at the latent level (Bolt, Hare, Vitale, & Newman, 2004). For example, if fighting is more

normative among boys than girls, endorsement of an item about fighting may reflect a lower level of conduct problems for boys compared with girls. If this source of measurement noninvariance is ignored, then boys may receive spuriously higher scores on a conduct problems scale. Hence, moderate-risk boys may be over-selected for prevention programs, whereas high-risk girls may be underselected. No study to date, however, has examined the scalar equivalence of the AA-TOCA-R.

Item response theory (IRT) provides an appealing framework for studying scalar equivalence (Embretson & Reise, 2000). IRT provides information regarding the performance of items with the purpose of estimating parameters that describe the relationship between each item and the latent construct that the scale intends to measure (Flora, Curran, Hussong, & Edwards, 2008). To date, there have been no studies applying IRT methods to the AA-TOCA-R.

One advantage of IRT is that it can incorporate techniques for evaluating measurement invariance across subpopulations using tests of differential item functioning (DIF; Thissen, Steinberg, & Wainer, 1993). Essential to any DIF analysis is the identification of items that can be assumed to perform invariantly across sub-populations. These items contribute to defining the metric of a latent construct against which the remaining items can be examined for DIF. Items that display DIF have the potential to contribute to bias in the overall scale scores if DIF consistently favors one group across several items.

In the present study, we report on an IRT analysis with over 8,800 teacher reports of classroom conduct problems among kindergarteners. Within this sample, we test for DIF across gender and race/urban status. DIF analyses across gender and race have been extensively applied to childhood substance use outcomes (Gilder, Lau, & Ehlers, 2009) and attention-deficit/hyperactivity disorder symptoms (Hillemeier, Foster, Heinrichs, Heier, & Conduct Problems Prevention Research Group, 2007); however, research on DIF of the AA-TOCA-R has not been done, and only two DIF studies have been conducted on measures of conduct problems. Studts (2008) conducted a survey on a convenience sample of 900 preschool-age children recruited from waiting rooms of pediatric care facilities. Parents were asked to rate their child's externalizing behavior problems, and the researcher tested for DIF across child gender and race. The author concluded that none of the behavior problems demonstrated DIF. Likewise, Gross and colleagues (2006) conducted differential item analysis on the Child Behavior Checklist (Achenbach, 1992) in 682 preschool-age children and also found a lack of DIF across Whites and Blacks in an urban setting. The present study, therefore, aims to examine the scalar equivalence of the AA-TOCA-R on a large school-based sample that is representative of the community across both urban and rural environments.

Research has consistently shown that boys display higher overall levels of childhood conduct problems than girls (Bradshaw et al., 2010; Koth et al., 2009; Petras et al., 2005). Little is known, however, regarding gender differences in how various observed conduct problems discriminate levels of the underlying latent trait. In addition, at the same latent level of conduct problems, are there particular behaviors that are more often endorsed for boys or girls? The same question can be posed for differences across race and urban/rural status. Although these variables would ideally be analyzed separately, the nature of the present sample created a Race $\times$ Environment confound, in that less than 1% of the entire sample consisted of Blacks in rural environments. Therefore, we separated participants into three groups for the purposes of testing DIF across race/urban status: urban White children, urban Black children, and rural White children. Although Black children have been consistently shown to display higher overall levels of childhood conduct problems than White children (e.g., Jones, Dodge, Foster, Nix, & Conduct Problems Prevention Research

Group, 2002; Petras et al., 2004), little is known about the scalar equivalence of conduct problems comparing White children from urban versus rural environments or across White and Black children in an urban environment.

In summary, the goals of the present study were as follows: (a) to use an IRT model to examine item characteristics of the AA-TOCA-R in kindergarten children (i.e., to obtain sample-free estimates of item discrimination and difficulty) and (b) to estimate the degree of DIF by gender and race/urban status.

Method

Participants and Design

The present study is a secondary analysis of data from the screening sample of the Fast Track study¹ (Conduct Problems Prevention Research Group, 1992). Four geographic sites were selected for the original Fast Track study: Durham, NC, a small urban city with a large low-income population that is primarily Black; Nashville, TN, a moderate-sized urban city with a mix of low-to-middle income and Black and White population; Seattle, WA, a moderate-sized urban city with a mix of low-to-middle income and an ethnically diverse population; and central Pennsylvania, a rural area with low-to-middle income White population. Samples from these sites varied widely in ethnicity (most minorities were Black, with some Latino and Asian children) and poverty (as measured by free/reduced lunch rates) as follows: Durham, NC = 90% minority and 80% reduced lunch; Nashville, TN = 54% minority and 78% reduced lunch; Pennsylvania = 1% minority and 39% reduced lunch; and Seattle, WA = 52% minority and 46% reduced lunch.

A multiple-gating screening procedure (Lochman & Conduct Problems Prevention Research Group, 1995) that combined teacher and parent ratings of disruptive behavior was applied to all 9,594 kindergarteners across three cohorts (1991–1993) in 55 schools (Durham = 13, Nashville = 10, Pennsylvania = 17, and Seattle = 15). Children were screened initially for classroom conduct problems by more than 350 teachers, using the AA-TOCA-R (Werthamer-Larsson et al., 1991). Within each site, schools were divided into sets matched for demographics and randomly assigned to control and intervention groups. For purposes of the present study, all urban White ($n = 2,184$), urban Black ($n = 3,960$), and rural White ($n = 2,678$) children were included in analyses because randomization took place after the TOCA-R screen was administered and treatment outcome was not relevant to our proposed hypotheses. In Durham, Nashville, and Pennsylvania, the total number of Latino, Asian, and other ethnicity children accounted for less than 2% of each site's sample. In Seattle, Latino children accounted for 4%, Asian children accounted for 5%, and other ethnicity children accounted for 17%. Hence, because the number of Latino, Asian, and other ethnicity children in the entire Fast Track sample was not large enough for testing DIF across ethnicities, these children were excluded, leaving a total of 8,820 White and Black children (91% of the original 9,594 kindergarteners screened). Black children recruited from Pennsylvania were also excluded from analyses because they consisted of less than 1% of the sample. The site distribution of the 8,820 study sample was as follows: Durham = 2,373; Nashville = 2,126; Pennsylvania = 2,677; and Seattle = 1,644. The mean age of participants was 6.5 years ($SD = 0.48$) at the time of identification.

Measure

The measure for conduct problems was composed of 10 items from the AA subscale of the TOCA-R (Werthamer-Larsson et al., 1991), assessed during kindergarten. The original

¹For additional information concerning Fast Track, see <http://www.fasttrackproject.org>

TOCA (Kellam, Branch, Agrawal, & Ensminger, 1975) was developed by the Woodlawn Research Center (Chicago, IL) to assess social adaptation and not intended to provide ratings of clinical symptomatic behaviors. The TOCA underwent a major revision by the Johns Hopkins Center for Prevention and Early Intervention in 1991. The TOCA-R became a 43-item questionnaire designed for teachers to assess three subscales relevant to the child's behavior in a classroom situation. Each question asks, "In the last three weeks, would you say the following statements were 0 = never, 1 = rarely, 2 = sometimes, 3 = often, 4 = very often, or 5 = always true of this child?" Then each question would be followed by 1 of 43 items that were grouped into (a) disruptive and aggressive (i.e., authority acceptance) behavior; (b) concentration problems; (c) and shy behavior. The alpha coefficient for each subscale exceeded .80 (Werthamer-Larsson et al., 1991). For purposes of the present study, only the AA subscale was analyzed, which includes 10 items representing conduct problem behavior (see Table 1). Internal consistency of the AA-TOCA-R subscale for the present sample was adequate (Cronbach's $\alpha = .94$).

Statistical Analyses

The present study aimed to test an IRT model of the 10 AA-TOCA-R items and DIF across gender and race/urban status. A core assumption underlying the IRT model is that the set of observed items represents a single dimension (Embretson & Reise, 2000). Hence, to empirically evaluate dimensionality, a confirmatory factor analysis (CFA) was estimated using Mplus version 5.1 (Muthén & Muthén, 2008). The factor mean and variance were set to equal 0 and 1, respectively. In line with theory and previous studies (e.g., Koth et al., 2009), one latent factor was set to be indicated by all 10 items, and loadings were freely estimated.

In the present study, students in each classroom were rated by the same teacher. Teachers were nested within schools, and schools were nested within research sites. This nonindependence was accounted for at the teacher, school, and site level by adjusting the standard errors and chi-square for nonindependence of observations (Asparouhov, 2005). Weighted least square mean and variance (WLSMV) estimator with delta parameterization was used because the items were ordered categories. One can also use the robust maximum likelihood estimator (MLE), which produces slightly more efficient estimates than WLSMV when dealing with item response data (Muthén & Lehman, 1985). MLE, however, can be quite time-consuming if many dimensions of integration are needed.

A two-parameter graded response IRT model (Samejima, 1969) was estimated using a latent variable framework (described by Takane & de Leeuw, 1987) in Mplus version 5.1. Samejima's graded response model for polytomous items was used because a prespecified order of response categories (i.e., from "never" to "always") was assumed. The model includes three main parts. First, the theoretical construct of interest (θ), represents the underlying level of conduct problems. Second, the discrimination parameter a (i.e., slope) indicates how well an item discriminates across individuals that differ on θ (Reeve, Hays, Chang, & Perfetto, 2007). Last, the difficulty parameter b indicates the level of θ at which an individual would have a 50% chance of endorsing above a particular item threshold. For a recent technical discussion of the IRT model, see Kamata and Bauer (2008).

To scale θ , the latent factor was standardized to have $M = 0$ and variance = 1. Because Mplus uses metrics common to factor analysis with categorical indicators, one has to convert loadings and item thresholds into discrimination and difficulty parameters, respectively. Kamata and Bauer (2008) offer formulas for these transformations using the WLSMV estimator:

$$a = \lambda / \sqrt{1 - \lambda^2}, \quad (1)$$

where a is the discrimination parameter and λ is the factor loading, and

$$b = \tau / \lambda, \quad (2)$$

where b is the difficulty parameter and τ is the item threshold. For each item, one a parameter and five b parameters were estimated (0 = *never*, 1 = *rarely* [b_1], 2 = *sometimes* [b_2], 3 = *often* [b_3], 4 = *very often* [b_4], and 5 = *always* [b_5]). A total of 60 parameters (10 a and 10 $\times$ 5 b) were estimated.

In addition, item information curves were estimated to determine the range of θ explained by the items and how much information an item contributed within its range. Greater information, determined by the discrimination parameter, denotes more precision for measuring θ , whereas the difficulty parameters locate the information curve along the range of θ . The total information curve was also estimated, which is estimated on the basis of the discrimination and difficulty parameters of all of the items and illustrates how much information the scale yields across the entire range of θ .

Next, gender and race/urban status were incorporated as predictors of θ . *Race/urban status* was defined by the interaction between race and urban/rural status at the four Fast Track data collection sites. White participants were sampled from both urban and rural areas, whereas Black participants were primarily sampled from urban areas. As noted above, Black participants recruited from rural areas consisted of less than 1% of the sample and were therefore excluded from analyses.

Finally, a multiple-group approach was used to test DIF across gender and race/urban status using likelihood ratio testing (Satorra, 2000) to determine the extent of DIF across groups. Latent variable methods for DIF testing use either the multiple-group approach or multiple-indicator-multiple-cause (MIMIC) models, each of which is a type of structural equation model (Woods, 2009). Multiple-group analysis is often preferable to the MIMIC approach because more types of hypotheses can be tested (e.g., DIF with respect to item discrimination). Under the multiple-group model, invariance of measurement parameters is assumed to hold for different subgroups. To detect DIF, group membership was identified and the estimated item parameters for a given level of conduct problems were compared across groups. Building on the unconditional model, a fully restricted multiple-groups model was specified, which forced all parameters to be equal across groups, while also allowing θ to differ. This was done by fixing mean of $\theta = 0$ and variance of $\theta = 1$ for the reference group and allowing the mean and variance of θ to be freely estimated for the other group(s). We then tested whether freeing each parameter significantly improved model fit compared with the fully restricted baseline model by starting with the a parameters for each item and then testing the b parameters (one at a time) for each item. A significant chi-square difference using the log-likelihood from each model (Satorra, 2000) would indicate that the parameter in question differed across groups, thereby exhibiting DIF. Due to the large number of comparisons, DIF was determined to be significant across groups only if the chi-square difference test indicated the model fit significantly better at $p < .01$ (McCarthy, Pedersen, & D'Amico, 2009). It is important to note that, up to this point, reference to DIF included both uniform and nonuniform DIF. Mellenbergh (1982) offers an interpretable definition of the two types of DIF. *Uniform DIF*, which refers to a significant difference in the difficulty (b) parameter without significant differences in the discrimination parameter (a), exists when the level of endorsement for an item is greater for one group relative to another group, and this is uniform over the entire range of θ . However, *nonuniform DIF*,

which refers to a significant difference in the discrimination (a) parameter, exists when the level of endorsement for an item is greater for one group relative to another group only at particular values of θ ; hence, there is an interaction between group membership and level of underlying conduct problems.

DIF was tested for gender (males as the reference group) and race/urban status (urban Whites as the reference group). Differences in parameter estimates indicates that, at a given level of θ , group membership influences item characteristics. Not accounting for DIF could potentially lead to biased estimates of θ based on group membership.

Results

Table 1 provides descriptive information regarding endorsement of items as a function of gender and race/urban status. Overall, defiant behaviors such as being stubborn, disobedient, and breaking rules had the highest means, whereas aggressive and deceitful behaviors such as harming others, breaking things, and taking property had the lowest means. This pattern was consistent between males and females, with males having a higher mean on each item, and also among urban White children, urban Black children, and rural White children, with urban Black children having the highest mean and rural White children having the lowest mean on each item.

Unidimensionality

A total number of 60 parameters (10 total loadings and five thresholds for each of the 10 items) were estimated. The model fit was adequate, $\chi^2(25, N = 8,820) = 2428.55, p < .001$, comparative fit index (CFI) = .972, Tucker-Lewis Index (TLI) = .964, and root-mean-square error of approximation (RMSEA) = .091. The standardized factor loadings for the 10 items ranged from .821 (yells at others) to .906 (breaks rules). The R^2 values for the 10 items ranged from .674 (yells at others) to .820 (breaks rules). Thus, a single factor was successful in accounting for the variability in the items, which is consistent with the results of Koth and colleagues (2009). We also tested the model with multiple factors, but the results (not shown) favor the one-factor structure based on empirical parsimony.

Two-Parameter-Graded Response IRT Model

In the next step, a two-parameter IRT model was fitted to the six-category AA-TOCA-R items. An examination of several individual trace lines, however, suggested that the number of categories used in the original AA-TOCA-R may not be optimal (see Figure 1). Specifically, the trace lines for Categories 3 and 4 were confounded. Recall that the categories were defined by 0 = *never*, 1 = *rarely*, 2 = *sometimes*, 3 = *often*, 4 = *very often*, 5 = *always*. On the basis of the trace lines, we concluded that the categories of *often* and *very often* were relatively indistinguishable. For example, looking at the upper left and right corners of Figure 1, we see the trace lines for the six categories of the items “takes property” and “breaks rules” do not follow a linear trend. In a typical graded response two-parameter polytomous IRT model, trace lines for subsequent categories of an item should follow an increasing or decreasing trend if the latent variable that the item is assumed to measure is unidimensional. In other words, we assume that when a teacher is asked how often the child “takes others’ property” or “breaks rules,” the category responses from *never*, *rarely*, *sometimes*, *often*, *very often* to *always* (Categories 0–5, respectively) take on an increasing severity of conduct problems. However, our data suggested that *often* and *very often* (Categories 3 and 4, respectively) were confounded. Semantically, it is relatively easy to distinguish between *never* versus *rarely*, or *sometimes* versus *often*, but the disparity between *often* versus *very often* may be not so clear.

To test this assumption, we recoded all 10 items such that all responses of Categories 3 and 4 would be combined, resulting in a five-category scale: 0 = *never*, 1 = *rarely*, 2 = *sometimes*, 3 = *often/very often*, and 4 = *always*. A total number of 50 parameters (10 total loadings and four thresholds per item) were estimated. The model fit was adequate, $\chi^2(35, N = 8,820) = 2265.581$, $p < .001$, CFI = .974, TLI = .967, and RMSEA = .082. The standardized factor loadings for the 10 items ranged from .825 (yells at others) to .904 (breaks rules). The R^2 values for the 10 items ranged from .681 (yells at others) to .818 (breaks rules). The results suggest that the six-category and five-category CFA models were virtually identical in terms of model fit. The five-category model, however, contained 10 fewer parameters, making it the more parsimonious model. In addition, the observed versus expected percentages of responses (given the estimated IRT parameters) in each category across every AA-TOCA-R item were nearly identical.

In the next step, a two-parameter graded response IRT model was fitted to the five-category AA-TOCA-R items. The bottom half of Figure 1 shows trace lines for the items “takes property” and “breaks rules.” For both items, the combination of *often* and *very often* produced a smoother progression of trace lines. This disparity between the fit of six- versus five-category trace lines was detected for all 10 items (results not shown). Figure 2 provides the CFA models, evidence of smoother trace lines, and equivalent total information curves for the five- and six-category IRT models. From the plots, we can conclude that the two models are not significantly different. Hence, with the similarity of fit indices from the in Table 1 reflect the five-category models (i.e., range = 0–4).

Figure 3 shows the individual trace lines for all items. Overall, item discriminations seem to be quite similar (range of $a = 1.460$ “yells at others” to 2.127 “harms others”). There are, however, minor differences when examined closely. The items “breaks rules,” “harms others,” “fights,” and “disobedient” all have a parameter values above 2.0 (as can be seen in their steeper slopes), whereas the other six items all have a parameter values below 2.0. This implies that a one-unit increase in θ is associated with larger increases in the mean levels of “breaks rules,” “harms others,” “fights,” and “disobedient” relative to items with lower discrimination values.

The next results to consider are the individual item difficulties. On the one hand, for the items “harms others,” “breaks things,” and “takes property,” most trace lines are located to the right of $\theta = 0$. On the other hand, for items “stubborn,” “breaks rules,” and “disobedient,” there are more trace lines centered at $\theta = 0$. That is, a much higher level of θ is needed to have a 50% probability of endorsing “harms others,” “breaks things,” and “takes property,” relative to items with lower difficulty values.

Figure 4 shows the item information curves. From these plots, we can examine three main questions: (a) Which items provide the most information? (b) For what level of θ does each item provide the most information? and (c) What is the range of θ for which each item space covers? First, the items “breaks rules,” “harms others,” “fights,” and “disobedient” have peaks around 0.6, indicating that these items provide the most information. However, the other six items have peaks just around or below 0.4. For items “stubborn,” “breaks rules,” and “disobedient,” most information is given at $\theta = 0$. For items “harms others,” “fights,” “lies,” “teases,” and “yells at others,” most information is given at $0 < \theta < 1$. For items “breaks things” and “takes property,” most information is given at θ around 1. Hence, for individuals with mean levels of θ , being stubborn, breaking rules, and being disobedient are most indicative of conduct problems, whereas for individuals who are one standard deviation above the mean, breaking things and taking others’ property are most indicative of

conduct problems. Last, to see the range of θ associated with each item more easily, vertical axes at -2 and 2 standard deviations from the mean have been drawn. The items “stubborn” and “breaks rules” are most symmetric about the mean so they cover a relatively normal θ space, whereas the other eight items tend to cover θ space that is shifted toward the more severe end.

DIF

To evaluate the degree to which the item parameters are equal across subgroups, the single-group two-parameter IRT model was expanded to a multiple-groups analysis to test DIF across gender and race/urban status. Before formally testing DIF, gender and race/urban status were included as covariates in a single group model by regressing θ on these covariates. The model provided a reasonable fit to the observed data, $\chi^2(28, N = 8,820) = 1762.659, p < .01$, CFI = .941, TLI = .983, and RMSEA = .080. Females had a significantly lower θ than males ($\beta = -.225, p < .01$), urban Black children had a significantly higher θ than urban White children ($\beta = .287, p < .01$), and rural White children had a significantly lower θ than urban White children ($\beta = -.231, p < .01$).

Results from the DIF analysis are reported in Table 2. Two multiple-group models were estimated—one for gender and another for race/urban status. One by one, each of the 10 discrimination parameters (a) and 40 difficulty parameters (b) were freed across groups. For each parameter, we compared the less restrictive model (with items freed) with the baseline model, with all item parameters constrained to be equal across groups, via a chi-square difference test using log-likelihoods (Satorra, 2000). The final gender DIF model, in which all significantly different parameters were allowed to vary between groups, provided an adequate fit to the observed data, $\chi^2(48, N = 8,820) = 1735.999, p < .01$, CFI = .950, TLI = .992, and RMSEA = .089. Males were set to be the reference group with mean of $\theta = 0$ and variance of $\theta = 1$. Six discrimination parameters showed differences as a function of gender: “harms others,” “fights,” “stubborn,” “breaks things,” “takes property,” and “yells at others.” These differences are indicative of nonuniform DIF by gender. Overt behaviors like “harms others,” “fights,” and “breaks things” were better discriminators of latent conduct problems for males, and nonphysical behaviors like “stubborn,” “takes property,” and “yells at others” were better discriminators of latent conduct problems for females. In addition, 22 difficulty parameters showed differences as a function of gender. Recall that each item has four difficulty parameters due to the ordinal nature of the responses: never versus rarely, never versus sometimes, never versus often/very often, and never versus always. Controlling for the mean level of θ , males were more likely to “break rules” (sometimes and often/very often), “harm others” (rarely, sometimes, often/very often, and always), “break things” (sometimes), and “fight” (rarely, sometimes, and often/very often), whereas females were more likely to be “stubborn” (rarely, sometimes, often/very often, and always), “take property” (rarely, sometimes, and often/very often), “lie” (rarely), and “yell at others” (rarely, sometimes, often/very often, and always). This implies that at equivalent levels of latent conduct problems, males were rated as having more overt behaviors, whereas females were rated as having more nonphysical behaviors. In summary, the items “stubborn,” “harms others,” “breaks things,” “takes property,” “fights,” and “yells at others” showed nonuniform DIF, whereas the items “breaks rules” and “lies” showed uniform DIF.

However, we did not find consistent support for DIF $\times$ Race/ Urban Status. The final race/urban status DIF model, in which all significantly different parameters were allowed to vary across groups, provided an adequate fit to the observed data, $\chi^2(60, N = 8,820) = 1094.335, p < .01$, CFI = .962, TLI = .993, and RMSEA = .077. Urban White children were set to be the reference group, with mean of $\theta = 0$ and variance of $\theta = 1$. Three discrimination parameters showed differences as a function of race/urban status: “breaks rules,” “harms others,” and “lies.” Four difficulty parameters showed differences as a function of race/

urban status. Controlling for mean level of θ , urban Black children were more likely to break rules (sometimes and often/very often) relative to urban and rural White children, whereas rural White children were more likely to fight (rarely and sometimes) relative to urban White and Black children. In summary, the items “breaks rules,” “harms others,” and “lies” showed nonuniform DIF, whereas only the item “fights” showed uniform DIF. Due to the limited number of either discrimination or difficulty parameters exhibiting DIF, however, no discernable pattern was detected.

Discussion

The goals of this study were to use a two-parameter IRT model to examine item characteristics of the 10 AA-TOCA-R items in kindergarten children and estimate the degree of DIF by gender and race/urban status. A sample of 8,820 children (2,184 urban White, 3,959 urban Black, and 2,677 rural White) with a mean age of 6.5 years at the time of identification were included in all analyses. Our results suggested that the AA-TOCA-R provides a range coverage between -3 and $+4$ standard deviations of underlying conduct problems during early childhood. Moreover, the AA-TOCA-R exhibited considerable DIF by gender, but not by race/urban status, suggesting that considerable caution should be used in applying this scale equally across genders.

Behaviors such as being stubborn, disobedient, and breaking rules had the highest means, whereas aggressive and deceitful behaviors such as harming others, breaking things, and taking others' property had the lowest means. These results are consistent with the developmental stage of the sample (approximately 6 years old), when conduct problems of a more delinquent and severe nature are less prevalent (Campbell et al., 2000). This pattern of behavior was similar between males and females, with males having a higher observed mean on each item, and also similar among urban White, urban Black, and rural White children, with urban Black children having the highest observed mean and rural White children having the lowest observed mean on each item. We found similar results when we conducted tests of DIF and allowed the latent means of these groups to differ, such that males had a higher latent score than females, and urban Black children had the highest latent score, whereas rural White children had the lowest latent score. These results support previous studies in which these differences between gender (Bradshaw et al., 2010) and race (Petrus et al., 2004) have also been observed.

Although the AA-TOCA-R did function fairly well as a measure of conduct problems, the original scaling was not ideal. Our results suggested that Categories 3 (*often*) and 4 (*very often*) did not differentiate individuals with higher and lower levels of conduct problems. Combining Categories 3 and 4 produced a smoother progression of trace lines, such that each category appeared to be more distinct relative to the six-category model. This finding highlights the need to check the uniqueness (or lack thereof) of each category when using ordered categorical items in an IRT framework. Edelen and Reeve (2007) suggested that individual trace lines can be used to determine whether one needs to modify the response category format in order to optimize the coverage of the item categories for measuring distinct portions of the construct continuum. Therefore, in the present study, it could be that perhaps *often* and *very often* were difficult to distinguish semantically. Alternatively, it could be that the six-category scale simply did not require as many categories at the most severe end to adequately distinguish among children who were more or less severe on the latent trait.

Prior studies had shown evidence of gender and race/urban status differences in conduct problems, but few studies had tested for DIF across gender and race/urban status (Gross et al., 2006; Studts, 2008). Our results indicated substantial DIF, both uniform and nonuniform

across gender, but not across race/urban status. Because overt behaviors like “harms others,” “fights,” and “breaks things” were better discriminating items for boys, clinical assessors could target these specific behaviors when evaluating conduct problems in boys. Likewise, results from the present study also suggest that nonphysical behaviors like “stubborn,” “takes property,” and “yells at others” could be targeted when evaluating conduct problems in girls. In addition, boys and girls who were actually at the same level of the latent conduct problems factor would be rated differently by teachers, but this depended on the level of conduct problems. Specifically, teachers tended to rate males as higher than females on overt conduct problem behaviors, but this was only true at higher levels of the latent conduct problems factor. Similarly, girls tended to be rated as having higher levels of nonphysical conduct problem behaviors than males at the same level of the latent conduct problems factor, but only at higher levels of underlying conduct problems. In other words, when the level of the latent trait was held constant, teachers still rated boys and girls differently on certain types of conduct problems at higher levels of those problems. Thus, clinical assessors should be aware that not all conduct problem behaviors are exhibited in equal levels between boys and girls who are on the severe end of underlying latent spectrum.

These results are consistent with other research on gender differences in childhood conduct problems. For example, Cairns and Cairns (1994) found that boys showed more physical, overt aggression relative to girls. Without accounting for gender differences in the expression of conduct problems, the AA-TOCA-R will not provide accurate estimates of the level of conduct problems for boys or girls, which in turn may influence researchers’ findings. However, we found no discernable pattern among the limited number of discrimination (i.e., “breaks rules,” “harms others,” and “lies”) and difficulty (i.e., “breaks rules” and “fights”) parameters exhibiting DIF across race/urban status. Gross and colleagues (2006) also tested for DIF of conduct problems across race and concluded that conduct problem scales were equivalent across Black, White, and Latino children. Future research should examine the impact of measurement invariance in the use of the AA-TOCA-R to determine the degree to which gender differences in item functioning influence predictions made by the AA-TOCA-R.

The findings of this study indicate that the AA-TOCA-R appears to be a sound measure of conduct problems for children that functions differently across males and females; however, limitations must be addressed. First, IRT is inherently limited by the item set at hand, and as noted, the AA-TOCA-R does not cover the low end of the latent conduct problems trait. Other measures that include more, less severe conduct problems may be more sensitive to this end of the trait. Second, despite the large sample size, there were still insufficient numbers to test DIF across all ethnicities represented in the Fast Track screening sample. It is unknown whether or how the AA-TOCA-R items might function differently in these other ethnic subgroups. Likewise, the sample was not large enough to examine DIF among rural Black children. This is important, considering recent research that has identified higher rates of violence and delinquency among rural Black youth in comparison to urban Black youth (Vazsonyi, Trejos-Castillo, & Young, 2008). Third, the screening data used in the present analyses came from a cohort of children who were first assessed from 1991 to 1993. Future research should be conducted to determine whether the present findings generalize to children in other parts of the United States and other countries and children born in the 21st century.

Despite these limitations, the present study adds knowledge regarding the measurement of childhood conduct problems using the AA-TOCA-R. The present study was the first in which an examination of the individual item characteristics was conducted and comparisons made across gender and race/urban status. Results suggested that the mean level of latent conduct problems was best represented by behaviors such as being stubborn, breaking rules,

and being disobedient, whereas breaking things and taking others' property best represented the construct at one standard deviation above the mean. Overt behaviors tend to be better discriminators of latent conduct problems for males, and nonphysical behaviors tend to be better discriminators for females. At equivalent levels of latent conduct problems, males received more endorsements of overt behaviors from teachers, whereas females received more endorsements of nonphysical behaviors. Differences across race/urban status were not found to be conceptually meaningful. Our analyses also suggest that the item scaling of the TOCA-R may be best represented by five instead of six categories. The results of this study provide encouraging support for the use of IRT modeling to examine item characteristics of conduct problems scales and DIF to test for scalar equivalence across diverse sub-populations.

Acknowledgments

This research was supported by National Institute of Mental Health (NIMH) Grants R18 MH48043, R18 MH50951, R18 MH50952, and R18 MH50953 and by National Institute on Drug Abuse (NIDA) Grant 1 RC1 DA028248-01. The Center for Substance Abuse Prevention and NIDA also provided support for Fast Track through a memorandum of agreement with the NIMH. This work was also supported in part by Department of Education Grant S184U30002, NIMH Grants K05MH00797 and K05MH01027, and NIDA Grants DA16903, DA015226, and DA017589. The authors are grateful for the close collaboration of the Durham Public Schools, the Metropolitan Nashville Public Schools, the Bellefonte Area Schools, the Tyrone Area Schools, the Mifflin County Schools, the Highline Public Schools, and the Seattle Public Schools. We greatly appreciate the hard work and dedication of the many staff members who implemented the project, collected the evaluation data, and assisted with data management and analyses.

References

- Achenbach, T. Manual for the Child Behavior Checklist/2–3 and 1992 Profile. Burlington: University of Vermont, Department of Psychiatry; 1992.
- Asparouhov T. Sampling weights in latent variable modeling. *Structural Equation Modeling*. 2005; 12:411–434.10.1207/s15328007sem1203_4
- Bolt DM, Hare RD, Vitale JE, Newman JP. A multigroup item response theory analysis of the Psychopathy Checklist-Revised. *Psychological Assessment*. 2004; 16:155–168.10.1037/1040-3590.16.2.155 [PubMed: 15222812]
- Bradshaw CP, Schaeffer CM, Petras H, Ialongo N. Predicting negative life outcomes from early aggressive– disruptive behavior trajectories: Gender differences in maladaptation across life domains. *Journal of Youth and Adolescence*. 2010; 39:953–966.10.1007/s10964-009-9442-8 [PubMed: 19688587]
- Cairns, RB.; Cairns, BD. Life lines and risks: Pathways of youth in our time. New York, NY: Cambridge University Press; 1994.
- Campbell SB, Shaw DS, Gilliom M. Early externalizing behavior problems: Toddlers and preschoolers at risk for later maladjustment. *Development and Psychopathology*. 2000; 12:467–488.10.1017/S0954579400003114 [PubMed: 11014748]
- Conduct Problems Prevention Research Group. A developmental and clinical model for the prevention of conduct disorders: The FAST Track program. *Development and Psychopathology*. 1992; 4:509–527.10.1017/S0954579400004855
- Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*. 2007; 16:5–18.
- Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah, NJ: Erlbaum; 2000.
- Farrington, DP. Childhood aggression and adult violence: Early precursors and later-life outcomes. In: Pepler, DJ.; Rubin, KH., editors. The development and treatment of childhood aggression. Hillsdale, NJ: Erlbaum; 1991. p. 5-29.
- Flanagan KS, Bierman KL, Kam CM. Conduct Problems Prevention Research Group. Identifying at-risk children at school entry: The usefulness of multibehavioral problem profiles. *Journal of*

- Clinical Child & Adolescent Psychology. 2003; 32:396–407.10.1207/S15374424JCCP3203_08 [PubMed: 12881028]
- Flora DB, Curran PJ, Hussong AM, Edwards MC. Incorporating measurement nonequivalence in a cross-study latent growth curve analysis. *Structural Equation Modeling*. 2008; 15:676–704.10.1080/10705510802339080 [PubMed: 19890440]
- Gilder DA, Lau P, Ehlers CL. Item response theory analysis of lifetime cannabis-use disorder symptom severity in an American Indian community sample. *Journal of Studies on Alcohol and Drugs*. 2009; 70:839–849. [PubMed: 19895760]
- Gross D, Fogg L, Young M, Ridge A, Cowell JM, Richardson R, Sivan A. The equivalence of the Child Behavior Checklist/ 1½–5 across parent race/ethnicity, income level, and language. *Psychological Assessment*. 2006; 18:313–323.10.1037/1040-3590.18.3.313 [PubMed: 16953734]
- Hillemeier MM, Foster EM, Heinrichs B, Heier B. Conduct Problems Prevention Research Group. Racial differences in parental reports of attention-deficit/hyperactivity disorder behaviors. *Journal of Developmental and Behavioral Pediatrics*. 2007; 28:353–361.10.1097/DBP.0b013e31811ff8b8 [PubMed: 18049317]
- Jones D, Dodge KA, Foster EM, Nix R. Conduct Problems Prevention Research Group. Early identification of children at risk for costly mental health service use. *Prevention Science*. 2002; 3:247–256.10.1023/A:1020896607298 [PubMed: 12458763]
- Kamata A, Bauer DJ. A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*. 2008; 15:136–153.10.1080/10705510701758406
- Kellam, SG.; Branch, JD.; Agrawal, KC.; Ensminger, ME. *Mental health and going to school: The Woodlawn Program of assessment, early intervention, and evaluation*. Chicago, IL: University of Chicago Press; 1975.
- Koth CW, Bradshaw CP, Leaf PJ. Teacher Observation of Classroom Adaptation—Checklist: Development and factor structure. *Measurement and Evaluation in Counseling and Development*. 2009; 42:15–30.10.1177/0748175609333560
- Lochman JE. Conduct Problems Prevention Research Group. Screening of child behavior problems for prevention programs at school entry. *Journal of Consulting and Clinical Psychology*. 1995; 63:549–559.10.1037/0022-006X.63.4.549 [PubMed: 7673532]
- Loeber R, Hay D. Key issues in the development of aggression and violence from childhood to early adulthood. *Annual Review of Psychology*. 1997; 48:371–410.10.1146/annurev.psych.48.1.371
- McCarthy DM, Pedersen SL, D’Amico EJ. Analysis of item response and differential item functioning of alcohol expectancies in middle school youths. *Psychological Assessment*. 2009; 21:444–449.10.1037/a0016319 [PubMed: 19719355]
- Mellenbergh GJ. Contingency table models for assessing item bias. *Journal of Educational Statistics*. 1982; 7:105–118.10.2307/1164960
- Moffitt TE. Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*. 1993; 100:674–701.10.1037/0033-295X.100.4.674 [PubMed: 8255953]
- Moffitt TE, Caspi A, Harrington H, Milne BJ. Males on the life-course-persistent and adolescence-limited antisocial pathways: Follow-up at age 26 years. *Development and Psychopathology*. 2002; 14:179–207.10.1017/S0954579402001104 [PubMed: 11893092]
- Muthén B, Lehman J. Multiple-group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*. 1985; 10:133–142.10.2307/1164840
- Muthén, LK.; Muthén, BO. *Mplus (Version 5.1)*. Los Angeles, CA: Author; 2008.
- Petras H, Chilcoat HD, Leaf PJ, Ialongo NS, Kellam SG. Utility of TOCA-R scores during the elementary school years in identifying later violence among adolescent males. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2004; 43:88–96.10.1097/00004583-200401000-00018 [PubMed: 14691364]
- Petras H, Ialongo N, Lambert SF, Barrueco S, Schaeffer CM, Chilcoat H, Kellam S. The utility of elementary school TOCA-R scores in identifying later criminal court violence among adolescent females. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2005; 44:790–797.10.1097/01.chi.0000166378.22651.63 [PubMed: 16034281]

- Reeve BB, Hays RD, Chang C-H, Perfitto EM. Applying item response theory to enhance health outcomes assessment. *Quality of Life Research*. 2007; 16:1–3.10.1007/s11136-007-9220-6 [PubMed: 17033892]
- Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969; 34:100–114.
- Satorra, A. Scaled and adjusted restricted tests in a multi-sample analysis of moment structures. In: Hiejmans, RDH.; Pollock, DSG.; Satorra, A., editors. *Innovations in multivariate statistical analysis: A festschrift for Heina Heudecker*. London, England: Kluwer Academic Publishers; 2000. p. 233-247.
- Schaeffer CM, Petras H, Ialongo N, Poduska J, Kellam S. Modeling growth in boys' aggressive behavior across elementary school: Links to later criminal involvement, conduct disorder, and antisocial personality disorder. *Developmental Psychology*. 2003; 39:1020–1035.10.1037/0012-1649.39.6.1020 [PubMed: 14584982]
- Sourander A, Multimäki P, Nikolakaros G, Haavisto A, Ristkari T, Helenius H, Almqvist F. Childhood predictors of psychiatric disorders among boys: A prospective community-based follow-up study from age 8 years to early adulthood. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2005; 44:756–767.10.1097/01.chi.0000164878.79986.2f [PubMed: 16034277]
- Studs CR. Improving screening for externalizing behavior problems in very young children: Applications of item response theory to evaluate instruments in pediatric primary care. *Dissertation Abstracts International Section A: Humanities and Social Sciences*. 2008; 69(3-A): 1160.
- Takane Y, de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*. 1987; 52:393–408.10.1007/BF02294363
- Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holland, PW.; Wainer, H., editors. *Differential item functioning*. Hillsdale, NJ: Erlbaum; 1993. p. 67-113.
- Van de Vijver, F.; Leung, K. Methods and data analysis of comparative research. In: Berry, JW.; Poortinga, YH., editors. *Handbook of cross-cultural psychology, Vol. 1: Theory and method*. 2. Needham Heights, MA: Allyn & Bacon; 1997. p. 257-300.
- Vazsonyi AT, Trejos-Castillo E, Young M. Rural and non-rural African American youth: Does context matter in the etiology of problem behaviors? *Journal of Youth and Adolescence*. 2008; 37:798–811.10.1007/s10964-007-9239-6
- Werthamer-Larsson L, Kellam SG, Wheeler L. Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*. 1991; 19:585–602.10.1007/BF00937993 [PubMed: 1755437]
- Woods CM. Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*. 2009; 44:1–27.10.1080/00273170802620121

Figure 1. Six- and five-category trace lines for items “takes property” and “breaks rules.”

Figure 2.
Total information curves for the six- and five-category item response theory (IRT) models.

Figure 3.
Individual trace lines for all 10 items in the five-category item response theory model.

Figure 4. Individual information curves for all 10 items in the five-category item response theory model.

Table 1

Means (and Standard Deviations) of the Items From the Authority Acceptance Scale of the TOCA-R as a Function of Gender and Race/Urban Status

Variable	N	Stubborn	Breaks rules	Harms others	Breaks things	Takes property	Fights	Lies	Disobedient	Teases	Yells at others
Full sample	8,820	1.50 (1.24)	1.36 (1.18)	0.65 (0.95)	0.44 (0.78)	0.51 (0.87)	0.90 (1.08)	0.75 (1.01)	1.09 (1.20)	0.95 (1.07)	0.85 (1.05)
Male	4,561	1.65 (1.24)	1.64 (1.20)	0.85 (1.05)	0.58 (0.87)	0.60 (0.93)	1.16 (1.15)	0.88 (1.08)	1.31 (1.25)	1.14 (1.12)	0.96 (1.08)
Female	4,259	1.33 (1.22)	1.06 (1.06)	0.43 (0.76)	0.29 (0.62)	0.41 (0.79)	0.63 (0.94)	0.60 (0.91)	0.86 (1.10)	0.75 (0.97)	0.74 (0.99)
Urban White	2,184	1.44 (1.21)	1.21 (1.13)	0.62 (0.92)	0.37 (0.71)	0.42 (0.79)	0.79 (1.04)	0.62 (0.92)	1.01 (1.15)	0.87 (1.02)	0.75 (1.01)
Urban Black	3,959	1.78 (1.26)	1.71 (1.21)	0.87 (1.04)	0.59 (0.88)	0.70 (0.99)	1.12 (1.16)	1.02 (1.11)	1.37 (1.28)	1.22 (1.13)	1.05 (1.12)
Rural White	2,677	1.12 (1.13)	0.97 (1.04)	0.35 (0.70)	0.27 (0.60)	0.30 (0.67)	0.68 (0.93)	0.44 (0.44)	0.74 (1.01)	0.62 (0.89)	0.63 (0.90)

Note. TOCA-R = Teacher Observation of Classroom Adaptation-Revised.

Table 2

Results From Gender and Race/Urban Status Differential Item Functioning (N = 8,820)

Item	Parameter	Male	Female	$\Delta\chi^2$	Urban White	Urban Black	Rural White	$\Delta\chi^2$
Stubborn	<i>a</i>	1.47	2.15	8.20	1.55	1.40	1.45	1.17
	<i>b</i> ₁	-0.85	-0.89	9.52	-0.64	-0.54	-0.58	4.50
	<i>b</i> ₂	-0.15	-0.29	20.47	0.11	0.09	0.13	0.39
	<i>b</i> ₃	0.73	0.49	22.20	0.90	1.01	1.07	6.11
Breaks Rules	<i>b</i> ₄	1.79	1.41	13.62	2.01	2.08	2.09	0.29
	<i>a</i>	2.05	2.04	3.13	2.17	2.00	2.02	12.04
	<i>b</i> ₁	-0.81	-0.79	2.02	-0.41	-0.48	-0.46	1.78
	<i>b</i> ₂	-0.16	-0.01	23.45	0.32	0.17	0.29	12.54
Harms Others	<i>b</i> ₃	0.69	0.83	16.87	1.08	1.05	1.16	4.52
	<i>b</i> ₄	1.81	1.83	1.04	2.30	2.12	2.37	10.16
	<i>a</i>	2.10	1.22	34.62	2.25	1.91	2.18	10.14
	<i>b</i> ₁	0.03	0.15	7.81	0.33	0.40	0.49	5.02
Breaks Things	<i>b</i> ₂	0.72	1.13	42.66	1.02	1.18	1.23	8.46
	<i>b</i> ₃	1.46	1.88	8.77	1.72	1.88	1.94	7.24
	<i>b</i> ₄	2.44	3.25	7.66	2.85	2.87	2.98	0.59
	<i>a</i>	1.60	1.23	7.56	1.77	1.47	1.92	4.57
Takes Property	<i>b</i> ₁	0.37	0.43	1.15	0.73	0.75	0.67	0.81
	<i>b</i> ₂	1.20	1.52	15.16	1.57	1.67	1.47	6.50
	<i>b</i> ₃	2.03	2.24	0.88	2.34	2.45	2.32	1.39
	<i>b</i> ₄	2.90	3.23	1.67	3.44	3.33	3.16	1.34
Fights	<i>a</i>	1.38	2.21	13.66	1.40	1.32	1.59	1.42
	<i>b</i> ₁	0.44	0.23	11.43	0.74	0.69	0.68	0.95
	<i>b</i> ₂	1.18	0.88	13.79	1.53	1.44	1.42	4.06
	<i>b</i> ₃	1.97	1.52	14.76	2.24	2.26	2.20	0.39
Fights	<i>b</i> ₄	2.85	2.49	0.35	3.09	3.25	3.07	1.36
	<i>a</i>	2.00	1.32	15.49	2.01	1.91	2.17	4.23
	<i>b</i> ₁	-0.31	-0.16	26.51	0.14	0.20	-0.07	26.04

Item	Parameter	Male	Female	$\Delta\chi^2$	Urban White	Urban Black	Rural White	$\Delta\chi^2$
Lies	b_2	0.35	0.60	30.06	0.79	0.79	0.63	12.82
	b_3	1.15	1.49	17.84	1.49	1.58	1.47	3.62
	b_4	2.22	2.55	2.38	2.54	2.62	2.62	0.19
	a	1.68	1.69	1.10	1.63	1.63	1.75	10.89
Disobedient	b_1	0.03	-0.12	15.19	0.35	0.23	0.35	6.61
	b_2	0.70	0.65	5.28	1.10	0.93	1.05	7.58
	b_3	1.49	1.46	1.46	1.85	1.79	1.90	3.24
	b_4	2.47	2.47	0.04	2.98	2.81	2.83	3.28
Teases	a	2.02	2.03	0.62	2.28	1.94	1.97	1.30
	b_1	-0.38	-0.40	0.72	-0.09	-0.01	-0.09	4.02
	b_2	0.21	0.23	1.07	0.53	0.57	0.58	0.96
	b_3	0.93	0.93	0.03	1.19	1.27	1.35	4.77
Yells at Others	b_4	1.82	1.79	2.26	2.13	2.14	2.14	0.09
	a	1.43	1.42	0.05	1.49	1.44	1.41	5.61
	b_1	-0.34	-0.35	0.02	-0.01	-0.04	0.04	1.33
	b_2	0.40	0.42	0.80	0.75	0.71	0.85	4.67
	b_3	1.31	1.32	0.14	1.62	1.63	1.78	3.73
	b_4	2.63	2.62	0.06	3.10	2.93	3.08	2.70
	a	1.46	2.95	21.42	1.55	1.30	1.62	8.87
	b_1	-0.08	-0.25	26.19	0.21	0.20	0.01	8.95
	b_2	0.62	0.38	16.97	0.86	0.95	0.76	8.58
	b_3	1.48	1.10	15.95	1.65	1.81	1.69	5.14
	b_4	2.73	2.12	7.41	3.06	3.08	2.75	5.94

Note. $\Delta\chi^2$ = chi-square difference between groups for parameter. Boldface indicates that groups differed by $p < .01$.