J Psychopathol Behav Assess (2010) 32:157–168
DOI 10.1007/s10862-009-9145-1

Differential Item Functioning on Antisocial Behavior Scale Items for Adolescents and Young Adults from Single-Parent and Two-Parent Families

Young I. Cho · Monica J. Martin · Rand D. Conger · Keith F. Widaman

Published online: 27 June 2009

Abstract We investigated measurement equivalence in two antisocial behavior scales (i.e., one scale for adolescents and a second scale for young adults) by examining differential item functioning (DIF) for respondents from single-parent (n=109) and two-parent families (n=447). Even though one item in the scale for adolescents and two items in the scale for young adults showed significant DIF, the two scales exhibited non-significant differential test functioning (DTF). Both uniform and nonuniform DIF were investigated and examples of each type were identified. Specifically, uniform DIF was exhibited in the adolescent scale whereas nonuniform DIF was shown in the young adult scale. Implications of DIF results for assessment of antisocial behavior, along with strengths and limitations of the study, are discussed.

This research is currently supported by grants from the National Institute of Child Health and Human Development, the National Institute on Drug Abuse, and the National Institute of Mental Health (HD047573, HD051746, and MH051361) (Rand Conger, PI). Support for earlier years of the study also came from multiple sources, including the National Institute of Mental Health (MH00567, MH19734, MH43270, MH59355, MH62989, and MH48165), the National Institute on Drug Abuse (DA05347), the National Institute of Child Health and Human Development (HD027724), the Bureau of Maternal and Child Health (MCJ-109572), and the MacArthur Foundation Research Network on Successful Adolescent Development Among Youth in High-Risk Settings (Rand Conger, PI).

Y. I. Cho (✉) · K. F. Widaman (✉)
Department of Psychology, University of California,
One Shields Avenue,
Davis, CA 95616, USA
e-mail: yicho@ucdavis.edu

M. J. Martin · R. D. Conger
Family Research Group,
Department of Human and Community Development,
Davis, CA 95616, USA

Keywords Differential item functioning · Differential test functioning · Antisocial behavior · Adolescents · Young adults

Social scientists have long been interested in studying group differences on key outcome measures. For examples, gender differences in internalizing symptoms (Hankin and Abramson 2002; Sanders et al. 1999), racial differences in externalizing symptoms (Krueger et al. 2003; Ruchkin et al. 2006), and differences in growth trajectories between children from different family structures (Curran 2000; Beyers and Loeber 2003) are all areas in which researchers have reported differences between groups. However, whether group differences are real or the result of measurement bias is not always clear. When groups are compared on a given construct, the impacts of real group differences and bias should be recognized and differentiated (Dorans and Holland 1993; Millsap and Everson 1993). Detection of differential item functioning (DIF) and correction to items that exhibit DIF are extremely important to implement so that researchers can make more valid comparisons between groups.

In this study we investigated whether items in two scales of antisocial behavior function differently for individuals from single-parent and two-parent families across adolescence and adulthood. Adolescents and young adults from different family structures may recognize and interpret items designed to measure antisocial behavior in different ways because different family structures may alter thresholds of antisocial behavior in the two groups (Dishion and McMahon 1998). We employed an iterative procedure to detect whether items showed significant DIF. We also studied the effects of DIF on differential functioning at the scale level. Finally, we discussed implications of our findings for assessment of antisocial behavior, noting

limitations of the current DIF study on single-parent and two-parent families.

Measuring Antisocial Behavior

The self-report antisocial behavior scale created by Elliott, Ageton, and Huizinga (1985) for the original National Youth Survey (NYS) is perhaps the most widely known measure of antisocial behavior, and its pool of items is arguably the most widely used. In fact, items from this self-report antisocial behavior scale are so well known and widely used that three major longitudinal studies funded by the Office of Juvenile Justice and Delinquency Behavior Prevention—studies centered in Rochester (Thornberry et al. 1993), Denver (Huizinga et al. 1991), and Pittsburgh (Loeber et al. 1998)—and other major studies, including the Dunedin Multidisciplinary Health and Development Study (Moffitt et al. 1996), all use items from the original Elliott et al. scale or modified versions of them to measure antisocial behavior (Piquero et al. 2002).

The NYS began in 1976 as a national probability sample of over 1700 youth (Elliott et al. 1985), and the study protocol included an antisocial behavior inventory with 47 items designed to capture a wide range of behaviors, including Uniform Crime Report offenses. However, most researchers use or adapt items only from the general delinquency scale, a subset of 24 items, ranging in severity from serious and violent offenses (e.g., "had [or tried to have] sexual relations with someone against their will;" "used force to get money or things from other people [not students or teachers]"), to more common offenses such as theft and vandalism, to relatively minor offenses such as skipping classes. In the NYS, 177 respondents were randomly selected and reinterviewed approximately 4 weeks after their initial assessment during the 5th wave of the study. Test-retest correlations were 0.84 and .75 for the general antisocial behavior frequency and variety scores, respectively, and 0.52 to 0.93 for the general delinquency subset, with a mean of 0.74 across 22 estimates of test-retest reliability (Huizinga and Elliott 1986).

Measurement Equivalence Across Groups

When investigating measurement equivalence across groups, the developmental appropriateness of items is a concern. One strength of the antisocial items from the NYS (Elliott et al. 1985, 1989) is the fact that item content was adjusted to the developmental level of respondents. In our study, we investigate measurement equivalence in two different sets of items—one set that is developmentally appropriate for adolescents, and the other designed for young adults. Ideally, these forms should be linked so that scores from the adolescent and adult versions fall on the same scale, making it possible to make comparisons across adolescence and young adulthood. Although a detailed explanation for linking procedures is beyond the scope of the current study, the mandatory data collection design for linking scales and investigating DIF is that two different measures should have common items. The antisocial behavior measures used in this study have six items that are common across the adolescent and young adult forms. Hence, the scales for these two forms may be linked and placed on a common metric when there are no method (age) effects, a good spread of thresholds, and a number of other factors (Reise and Waller 2009). In the current study, these common items are the focus of testing DIF.

Methods for detecting measurement bias may identify either "an observed conditional invariance" or "an unobserved conditional invariance" (Millsap and Everson 1993). In particular, the second category contains likelihood ratio (LR) tests based on either IRT or confirmatory factor analysis (CFA) (see Reise et al. 1993, for comparison of LR tests based on IRT and CFA; Kim et al. 2007, for application of DIF detection methods). For short tests, Finch (2005) claimed that LR tests based on IRT more accurately detected uniform bias—that all categories in an item consistently behave in a fashion favoring one group over another group—than did a variety of other methods. In this study, we used the LR test to detect measurement bias on an antisocial behavior scale for adolescents and young adults.

Parameter estimation in item response theory can be categorized into parametric and nonparametric methods, scaling dichotomous as well as polytomous items. Specifically, parametric IRT model10 are based on either normal ogive or logistic functions, whereas nonparametric IRT model10 do not assume any specific parametric function (see Embretson and Reise 2000; Meijer and Baneke 2004; Sijtsma 1998, for further description of parametric and nonparametric IRT model10). When parametric IRT model10 are employed, likelihood ratio tests may be implemented by BILOG-MG (Zimowski et al. 2002) and MULTILOG-MG (Thissen 2003) for dichotomous and polytomous items, respectively. In contrast, when nonparametric IRT model10 are used, DIF tests may be conduced using TESTGraf (Ramsay 2001).

Thissen, Steinberg, and Gerrard (1986) discussed the LR test in parametric IRT (IRT-LR) in the context of analyses of dichotomous items. Later, Thissen and Steinberg (1988) extended the original parametric IRT-LR method for tests comprised of polytomous items. In this method, the basic idea is to compare the log likelihood values from two

model10, where one model, which is more restricted, is nested within the other, which is less restricted (i.e., has more parameter estimates than the restricted model).

In certain situations, the selection of items that serve as anchor items is obvious; in such situations, the computation of *LR* and investigation of item DIF is relatively simple. More commonly, which items should be considered anchor items and which should be studied items is less clear, and the assessment of item bias is therefore more complicated. Based on a simulation study, Candell and Drasgow (1988) recommended an iterative procedure for linking metrics from two separate groups and detecting which items exhibited DIF when anchor and studied items are mixed or unknown. Using an iterative method, Segall (1983) proposed four steps to detect items exhibiting DIF. First, item parameter metrics from two independently calibrated groups are initially linked. Second, after equating item parameters from the two groups by employing linking coefficients from the previous step, all items from a test are examined for item bias. Third, the item with the most extreme DIF is identified, and linking coefficients are recalculated after excluding this item. Fourth, with linking coefficients that are calculated without the item exhibiting most extreme DIF, G^2 is recomputed for all remaining items in a scale and evaluated with the corresponding critical value based on the alpha level.

Our aim in the present study was to investigate DIF for two scales of antisocial behavior derived from commonly used items from the NYS, one scale for adolescents and the second for young adults. Both scales were assumed to fit the graded response model10 (GRM; Samejima 1969). Children from two-parent families typically exhibit lower levels of antisocial behavior relative to children from single-parent families (Dawson 1991; Hoffmann 2006). However, a direct way of comparing levels of antisocial behavior across these two groups is possible only if antisocial behavior items do not show DIF. Therefore, we conducted DIF analyses on the adolescent and young adult forms of the antisocial behavior scale to determine whether items exhibited DIF across two groups of participants, one from two-parent families and the second group from single-parent families.

Method

Participants

The participants in this study come from the first (1994) and second (1995) waves of the Family Transitions Project (FTP; for additional information on the study, see Conger and Conger 2002). The FTP began in 1994, when most of the target adolescents were seniors in high school, and combines participants from two earlier samples—the Iowa Youth and Families Project (IYFP) and the Iowa Single—Parent Project (SPP).

The IYFP is a multiple-wave study of 451 rural families from eight counties in Central Iowa, with assessments beginning in 1989. The IYFP participants were recruited through 34 public and private schools from these eight counties. Families with a seventh grade child (target adolescent) living with both biological parents and a sibling within 4 years of the target adolescent's age were eligible for the study. About 78% of the families who met the criteria for inclusion agreed to participate in the initial wave of data collection in 1989. Ninety percent of the original 451 families remained in the study in 1992.

The SPP is a multiple-wave study of 107 single-mother families also from rural Iowa, with annual assessments beginning in 1991. SPP participants were identified through lists of students provided by schools in rural areas of Iowa and initially contacted by telephone. To match the chronological age of IYFP participants, a family was eligible to participate if the family had a target adolescent in the ninth grade in 1991 who had a sibling within 4 years of the target adolescent's age. Additionally, the household had to be mother-headed, and the biological parents must have divorced within the past 2 years. Fifteen percent of the women who were telephoned met all the criteria. Of those, 99% agreed to participate. In the first year of the study, all of the target adolescents lived in households headed by the mother, with 33% of the mothers having sole maternal custody, and 58% joint custody (Ge et al. 2006). Ninety-six percent of the original SPP sample remained in the study in 1993.

The FTP combined the two projects, matching the adolescents who were the same age and were in the 9th grade in 1991 (N=558). The two samples were similar on a number of important characteristics in 1991, including target's age, mother's age, number of children, mother's education, percentage of female targets, and percentage of mothers employed, although the SPP families had lower incomes than the IYFP families (Ge et al. 2006; Wickrama et al. 2003).

The first two waves (1994 and 1995) of the combined FTP sample of families provided data for the present study. In the FTP, professional interviewers made home visits to each family for approximately 2 h on two occasions each assessment period. During the first visit, each family member completed a set of questionnaires covering an array of topics including work, finances, family life, mental and physical health

status, friends, and antisocial behavior. Only information gathered from the first home visit in each assessment is used in the current analyses.

Measure of Antisocial Behavior

Target's antisocial behaviors were assessed using self-report items adapted from the NYS study (Elliott et al. 1985, 1989). Targets were asked to indicate how often during the past 12 months they had engaged in a variety of antisocial behaviors using a 5-point scale ranging from 0 (never) to 4 (six or more times). The items are presented in Appendix A. The six boldfaced items in Appendix A were common across both 1994 and 1995, 9 were unique to 1994, and 6 were assessed only in 1995. Therefore, the 15 items used in 1994 constitute the adolescent antisocial behavior scale, and the 12 items used in 1995 represent the young adult antisocial behavior scale. Both sets of items include items ranging in severity, which is necessary to represent the domain of antisocial behavior (Thornberry and Krohn 2000).

Data Analysis

We used SAS 9.1.3 (SAS Institute Inc., 2003) to calculate item frequency and descriptive statistics. Exploratory factor analysis (EFA) was utilized to investigate the dimensionality of the scales at the two measurement occasions. Common factor analysis was employed, with squared multiple correlations as communality estimates and principal axis extraction of factors. Scale unidimensionality was assumed, thus, rotation of factors was not conducted in this study.

MULTILOG (Thissen 2003) was utilized for calibrating item parameters and computing IRT-LR estimates. When computing IRT-LR estimates, one item at a time was considered as a studied (or tested) item exhibiting DIF, and the remaining items were used as anchor items. If at least one item exhibited significant DIF, the item with the largest DIF was discarded, and all other remaining items except the item having the largest G^2 values were kept and then examined for DIF in the following analysis. This procedure was repeated until no item in a scale showed significant DIF. This approach follows the four-step procedure outlined by Segall (1983) and supported by the simulation study by Candell and Drasgow (1988).

Item parameters from the two groups (two-parent and single-parent families) were simultaneously calibrated using anchor items in MULTILOG and IRTLR (Thissen 2001), thus, a linking procedure was not required in this study. If an item exhibited significant DIF, we examined the item discrimination (a) and category difficulty (b) parameters separately to determine which parameter was the source of DIF. If the a parameter was the source of DIF, this form of DIF is known as nonuniform DIF because ICCs across groups will cross at some point. In other words, when the ICCs for an item from two separate groups cross, the pattern of favoring one group to the other group is conditioned on latent ability level. Conversely, if the b parameter was the source of DIF, this form of DIF is known as uniform DIF, because the ICCs will not cross and the pattern favoring one group over the other is seen at all latent ability levels. Finally, effects of items with DIF on the total scale score were investigated through estimating Differential Test Functioning (DTF) with DFIT (Raju et al. 1995). This parametric procedure was intended to identify DTF through the comparison of test characteristic curves (TCCs) among groups. In order to compare TCCs, unlike our comparison of DIF at the item level, linking coefficients using EQUATE 2.0 (Baker 1993) were used.

Results

Sample Descriptive Statistics

Adolescent antisocial behavior scale The simple frequencies and descriptive statistics were calculated on items from the adolescent antisocial behavior scale for 12th grade students, who averaged 18 years of age. These values are reported in Appendix A. Antisocial behaviors were largely positively skewed and had large kurtosis. For eight of the 15 items (del2, del3, del5, del6, del7, del8, del13, del15), no respondent selected the highest category. Among these eight items, three (del2, del13, del15) also had no responses for the second highest category. All eight items had means of less than .10 and *SDs* of less than .42. Seven items (del3, del6, del7, del8, del13, del14, del15) had few responses for anything other than the zero response category. Hence, we restricted our analyses to 9 of the 15 items (del10, del11, del12, del1, del2, del4, del5, del7, del9) in the current study of DIF, because the remaining 6 items did not have sufficient responses beyond the zero response category at any measurement occasions to allow accurate parameter estimation. Among the nine items used in the current analyses, three items had no responses for the highest categories. In the following analyses, del2 was adjusted into an item having three categories (0, 1, or 2), and del5 and del7 were modified to have four categories (0, 1, 2, or 3), instead of the five categories used for the remaining 6 items. Test information and standard errors with and without the six items in the test were estimated via IRT and compared. No obvious differences were found.

Young adult antisocial behavior scale The simple frequencies and descriptive statistics for items from the young adult antisocial behavior scale were calculated for young adults who, on average, were 19 years old. These values are reported in Appendix B. Antisocial behavior items in 1995 had similar distribution patterns—large positive skew and large kurtosis—to items in 1994. Three of 12 items (del21, del13, del15) had no respondents selecting the highest category. Five of the 12 items (del19, del21, del13, del14, del15) had extremely few responses in categories other than the zero response category. Those five items had means of less than .05 and *SDs* of less than .36. Therefore, five of the 12 items (del 21, del14, del15, del19, del21) were deleted in the following analyses because these five items did not have enough responses in response categories other than zero to allow accurate parameter estimation. Therefore, DIF analyses were computed on the remaining seven items from the young adult form. Test information and standard errors with and without the five items in the test were estimated via IRT and compared. No obvious differences between them were found.

Dimensionality

Adolescent antisocial behavior scale Although the assumption of unidimensionality can be relaxed in some IRT model10, our analyses used the GRM, a generalized version of the 2PLM for polytomous items, and use of the GRM requires that items comprising a scale are unidimensional. Thus, our exploratory factor analyses were implemented to investigate whether only one latent construct was measured within each group. Number of factors in each scale was determined through a scree test and a ratio of the eigenvalue of the first factor to that of the second (Fabrigar et al. 1999). The initial eigenvalues and percentages of variances explained by each factor are reported in Table 1. In addition to these values, the scree test was also employed to estimate the number of factors in the antisocial behavior scale in 1994. The eigenvalues for the first three factors were 2.468, .537, and .129 for adolescents from two-parent (IYFP) families, and 3.827, .641, and .264 for adolescents from single-parent (SPP) families. The ratios of eigenvalues for the first factor to second factor were 4.60 for IYFP family and 5.97 for SPP family. These ratio values, which are greater than 3 to 1, supported the contention that the antisocial behavior scale at 12th grade was unidimensional (Hambleton et al. 1991; Lagenbucher et al. 2004; Lord 1980).

Young adult antisocial behavior scale Dimensionality of the young adult anti-social behavior scale at the 1995 measurement occasion was also scrutinized utilizing EFA. The eigenvalues for extracted factors and percentages of explained variance by the factors are reported in Table 1, and the scree test was once again used to study dimensionality of the young adult antisocial behavior scale. The eigenvalues for the first three factors were 1.905, .146, and .015 for respondents from IYFP families, and 1.612, .638, and .354 for participants from SPP families. The ratio of the first factor to the second factor was 13.048 for IYFP participants and 2.527 for SPP participants. Once again, these ratio values implied unidimensionality of the young adult anti-social behavior scale for 19-years olds (Hambleton et al. 1991; Lagenbucher et al. 2004; Lord 1980)

Differential Item and Test Functioning for Two Antisocial Behavior Scales

Adolescent antisocial behavior scale—DIF A backward elimination iterative method (Kim and Cohen 1995), deleting at each stage the item that exhibited the greatest DIF, was used because of the possible bias arising from deleting multiple items with statistically significant DIF in a single-step procedure. In the first iteration, all items from the scale were assumed to have no DIF between the two family groups and the G^2 was calculated. Only one item (del12) had a significant G^2 value of 11.1, implying DIF on

Table 1 Initial eigenvalues and percentages of variances explained by each factor in 1994 using nine items and in 1995 with seven items

Groups	Years	Eigen values			% of Variance		
		Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
IYFP	1994 (n=403) 1995	2.468	.537	.129	99.4	21.6	5.2
	(n=421)	1.905	.146	.015	124.3	9.5	1.0
SPP	1994 (n=101) 1995	3.827	.641	.264	88.8	14.9	6.1
	(n=86)	1.612	.638	.354	83.6	33.1	18.3

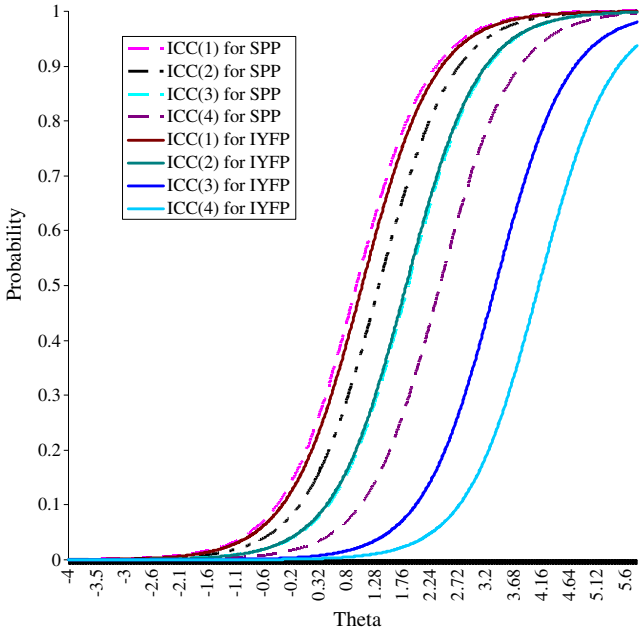

Fig. 1 Item characteristic curves of del12 for respondents from IYFP and SPP Families at 12th Grade (Adolescents)

the item. Item del12 was removed from the following iteration, and G^2 values for the remaining items were estimated. As expected, G^2 values for the eight remaining items changed from the first to the second iteration, but no items now exhibited significant DIF, signified by a significant G^2 statistic, at the second iteration.

The sources of DIF for item del12 were examined. The item response curves for item del12 are displayed in Fig. 1. The IRT-LR estimates for the *a* parameters and *b* parameters were calculated separately across groups. The *a* parameters showed a non-significant G^2 of 3.3 with 1 degree of freedom, so these parameters were constrained to invariance across groups. In contrast, the *b* parameters had a significant G^2 of 10.8 with 4 degrees of freedom, and thus were allowed to vary across groups. Because the *b* parameters varied significantly across groups, the IRT-LR results indicated that del12 showed uniform DIF between the SPP and IYFP groups. In order to specify the direction of DIF on del12, *b* parameter values of the two groups were compared, and the ICCs are shown in Fig. 1. The two groups had the same $b(1)$ of 1.02. The SPP group, however, had smaller $b(2)$ of 1.42, $b(3)$ of 1.96, and $b(4)$ of 2.55, when compared to the respective estimates from the IYFP group, which were $b(2)$ of 1.75, $b(3)$ of 3.22, and $b(4)$ of 3.94. The discrepancy of *b* parameters between the two groups indicated that an individual from a single-parent family was more likely to endorse a higher category on the del12 item than would an individual from a two-parent family even if the two individuals had the same latent variable value (and therefore had the same level on the underlying antisocial behavior latent variable). Item parameter estimates for both single- and two-parent families are reported in Table 2.

Adolescent antisocial behavior scale—DTF The effect of retaining one item with DIF on the entire 9-item scale was investigated by estimating the DTF index with DFIT (Raju et al. 1995). The 9-item adolescent antisocial behavior scale had a DTF value of .049 that was smaller than the cutoff of .792 (which was empirically provided by Raju et al. 1995), implying this scale exhibited non-significant DTF, despite the inclusion of one item that showed significant DIF, or differential functioning at the item level.

Young adult antisocial behavior scale—DIF The same iterative method was employed as in the analyses for the adolescent form. In the first iteration, items del16 and del17 had large and significant G^2 values of 16.8 and 14.7, respectively, $p<.05$. Because item del16 had the larger G^2 value, del16 was removed from consideration, and G^2s were computed for the remaining six items at the second iteration. In the second iteration, item del17 still exhibited a significant G^2 value of 14.7, $p<.05$, indicating DIF on the item. After removing item del17, no remaining item had a significant G^2 statistic. The item parameters for the highest category on item del17 could not be estimated because no SPP targets endorsed the highest category on that item. Therefore, direct comparisons of category parameters for item del17 between the two groups were not possible and it could not be considered in the following analysis.

The sources of DIF for item del16 were evaluated, however. The ICCs for item del16 for the two groups are displayed in Fig. 2. The IRTLR estimates for the *a* and *b* parameters, like the previous analysis, were calculated separately. The *a* parameter for del16 had a significant G^2 of 4.7, $p<.05$, and the *b* parameters also had a significant G^2 of 14.2, $p<.05$, indicating that del16 showed nonuniform DIF between SPP and IYFP.

The ICCs from the two groups in Fig. 2 show the intersections of ICCs between respondents from the two family groups. To specify the direction of DIF on item del16, *a* and *b* parameter values of the two groups were compared. The *a* parameter was .90 for the SPP group, and 1.72 for the IYFP group, respectively. SPP respondents had smaller difficulty parameters of $b(1) = .37$, $b(2) = .77$, $b(3) = 1.31$, and $b(4) = 1.81$, relative to IYFP respondents with higher difficulty parameters of $b(1) = .79$, $b(2) = .85$, $b(3) = 1.96$, and $b(4) = 2.40$. Because the ICCs intersected each

Table 2 Item parameter estimates for anchor items and studied items in 1994

Items	a_j		$b_j(1)$		$b_j(2)$		$b_j(3)$		$b_j(4)$	
	IYFP	SPP	IYFP	SPP	IYFP	SPP	IYFP	SPP	IYFP	SPP
del1	1.63 (.21)		.64 (.12)		1.21 (.17)		2.13 (.29)		2.55 (.37)	
del2	1.12 (.25)		2.15 (.45)		3.54 (.75)		N/A		N/A	
del4	1.66 (.24)		.83 (.14)		1.18 (.17)		1.77 (.25)		2.44 (.34)	
del5	1.85 (.41)		1.78 (.28)		2.44 (.39)		3.45 (.71)		N/A	
del7	2.62 (.58)		1.61 (.20)		2.10 (.28)		3.24 (.66)		N/A	
del9	1.47 (.20)		.57 (.13)		1.26 (.20)		2.37 (.34)		2.91 (.43)	
del10	1.82 (.23)		.57 (.11)		1.03 (.15)		1.74 (.21)		2.30 (.29)	
del11	1.52 (.20)		1.06 (.23)	.97 (.27)	1.82 (.34)	1.36 (.35)	3.36 (.78)	1.86 (.42)	4.11 (1.06)	2.41 (.58)
del12	2.51 (.35)		.81 (.10)		1.43 (.14)		2.12 (.23)		2.37 (.29)	

other, the interpretation of the probability for each category was complicated. For ICC 1 and 2, an individual with theta of higher than 1.0 from the IYFP group had a higher probability of responding in the higher category than an individual from the SPP group. When considering theta values less than 1.0, however, subjects from the IYFP group had a lower probability of responding in the higher category than subjects from the SPP. For ICC 3 and 4, the same pattern of likelihood of responding in the higher category for the two groups was shown. However, the level of theta that led to a flipping from higher to lower likelihood of higher category endorsement between SPP to IYFP participants now fell at theta of 2.40 for ICC 3 and 2.90 for ICC 4, respectively. Item parameter estimates for single- and two-parent families are reported in Table 3.

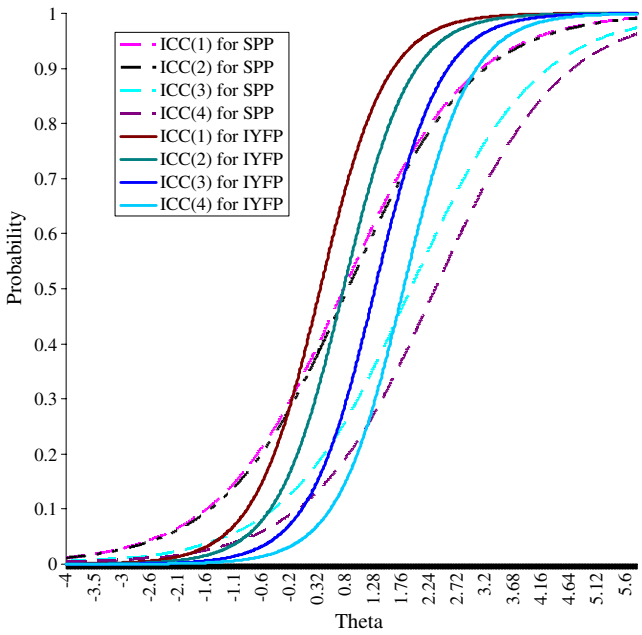

Fig. 2 Item characteristic curves of del16 for IYFP and SPP adolescents in young adults

Young adult antisocial behavior scale—DTF The young adult antisocial behavior scale had a DTF value of .027 that was smaller than the cutoff of .576 which was empirically estimated by Raju et al (1995). Hence, although two items from the antisocial behavior scale for young adults exhibited statistically significant DIF, retaining these two items showing DIF in the seven-item scale resulted in non-significant level of scale DTF. Therefore, individuals with the same true score on the total scale would have the same observed score across the items comprising the scale regardless of family structure.

Discussion

Our analyses revealed different types of DIF—uniform and nonuniform—which must be treated in different manners, as they imply different types of bias. A common approach to managing item *bias* is to delete any item showing DIF from a measure. However, retaining all items is vastly preferable due to the expensive and time-consuming development of a scale and confirmatory tests of factor structure in a scale. Hence, one way to handle scales that exhibit DIF is to correct the *bias* by retaining matching items with opposite *biases* that cancel out the DIF at a scale level (Teresi 2006). In order to match up the appropriate corresponding items and cancel the item DIF at a scale level, the direction and type of DIF should be recognized properly.

In our study, we found evidence of uniform *bias* in item del12. However this bias appeared to be counteracted at the

Table 3 Item parameter estimates for anchor items and studied items in 1995

Items	a_j		$b_j(1)$		$b_j(2)$		$b_j(3)$		$b_j(4)$	
	IYFP	SPP	IYFP	SPP	IYFP	SPP	IYFP	SPP	IYFP	SPP
del16	1.75	.93	.36	.77	.76	.84	1.30	1.92	1.79	2.34
	(.36)	(.35)	(.10)	(.41)	(.12)	(.42)	(.15)	(.73)	(.19)	(.89)
del17	1.90	.49	.89	2.49	1.27	3.79	1.75	6.26	2.26	17.45
	(.25)	(.49)	(.11)	(2.20)	(.14)	(3.34)	(.18)	(6.58)	(.24)	(***)
del18	1.39 (.16)		−.05 (.11)		.25 (.11)		.89 (.13)		1.50 (.18)	
del20	.62 (.20)		2.99 (.91)		3.73 (1.14)		5.23 (1.63)		6.73 (2.29)	
del10	2.16 (.29)		1.18 (.10)		1.54 (.14)		2.17 (.20)		2.80 (.33)	
del11	1.45 (.25)		1.65 (.21)		2.20 (.30)		3.16 (.52)		3.67 (.66)	
del12	2.53 (.38)		1.28 (.10)		1.69 (.14)		2.24 (.20)		2.64 (.30)	

scale level because the *bias* by which an adolescent from SPP has a higher probability of selecting a higher category than an individual from IYFP was cancelled out by slight bias favoring IYFP over SPP on remaining items in the adolescent scale, even though the *bias* from the remaining items was not statistically significant. Because significant uniform bias on one item (item del12) from the adolescent scale was counteracted by small and nonsignificant bias in the opposite direction on the remaining items, across-group DTF on the total scale was negligible. Similarly, the non-significant DTF for the antisocial behavior scale for young adults suggests that any *bias* in items del16 and del17 was also counteracted by opposite types and directions in biases on remaining items.

Despite the item DIF revealed in our analyses, comparisons between participants from single- and two-parent families using these scales appear sound, given our non-significant results with regard to scale DTF on these instruments. In general, when analyses reveal significant DTF, a researcher should scrutinize the magnitude and direction of DIF for all items in a scale and ensure the inclusion of items showing opposite directions and corresponding sizes of DIF to the original items in a test in order to have non-significant DTF. Because our analyses revealed non-significant DTF even in the presence of some significant differential functioning at the item level, this extra step of balancing bias across items was not necessary in the current study.

Similar to Dunifon and Kowaleski-Jones (2002), we found that adolescents from single-parent families exhibit more antisocial behavior than adolescents from two-parent families. Specifically, for 1994, the average estimated delinquency levels were −.04 (*SD*=.67) and .18 (*SD*=.82) in IYFP and SPP, respectively. For 1995, the average values were −.10 (*SD*=.63) and .11 (*SD*=.76), respectively. Two additional findings were noteworthy. First, items exhibiting DIF in 1994 were items representing more extreme forms of antisocial, whereas items exhibiting DIF in 1995 reflected milder forms of antisocial behavior. In 1994, for a given level of theta, adolescents from one-parent families had a higher probability of endorsing a given option on the extreme item del11 (beat up somebody because they made you angry) than adolescents from two-parent families. Conversely, in 1995, again for a given level of theta, adolescents from one-parent families had lower probability of endorsing a given option on the relatively mild items del16 (drive a car recklessly) and del17 (cheat at school or other places) compared to adolescents from two-parent families. Second, the two items showing DIF in 1995 indicate a different relationship with the construct of antisocial behavior depending on the family structure (i.e., single-parent and two-parent family). That is, the two DIF items representing relatively mild antisocial behaviors are better indicators of antisocial behavior in two-parent families than one-parent families: a_j parameters of 1.75 vs .93 and 1.90 vs .49 for del16 and del17, respectively.

The present study has several strengths and is of practical importance for longitudinal research on antisocial behavior in adolescents and young adults. First, this study highlights the importance of testing for DIF in the adolescent and young adult antisocial behavior scales, which are widely used in studies of problem behavior. In order to reveal the actual magnitude of the mean difference on the two scales between two groups such as single-parent and two-parents families, DIF tests at the item and scale levels were implemented. Second, having established the importance of detecting both item and scale bias, our study illustrated methods to make these comparisons. To conduct DIF tests on the two scales, we employed an iterative procedure for detecting DIF using the likelihood method. The iterative approach enabled us to identify bias arising from DIF on certain items and isolate this from any bias arising from DIF items on remaining items. Third, we showed how the two types of DIF—uniform and nonuni-

form—can be differentiated and treated. Uniform DIF was identified for one item on the adolescent scale, and nonuniform DIF was detected for two items on the young adult scale. Fourth, we suggested that scale level tests, such as DTF, are important and may support a conclusion that the overall scale is not biased even though significant item-level DIF bias is found.

Furthermore, the present study has another practical implication for longitudinal studies on antisocial behavior in adolescents and young adults. The two scales employed in this study have three common items with which the two scales can be linked onto a common scaling metric. By eliminating or controlling for DIF on the three common items, a researcher in a longitudinal study on antisocial behavior can compare trajectories of individuals from adolescence through young adulthood for individuals from both single-parent and two-parent families. That is, one could estimate scores on the underlying antisocial behavior latent variable that were on a comparable metric across adolescence and young adulthood, enabling one to study change in antisocial behavior tendencies across these age periods even though only a relatively small number of items are in use across age levels.

The current study also has limitations that should be noted. First, the sample size (n=556) of the present study was not extremely large, and the number of individuals from single-parent families was relatively small (n=109). Given the sample size, we had lower power to detect DIF than if we had a larger sample of participants available for analysis. Thus, it is possible that important levels of item DIF existed, but went undetected in our analyses. Additionally, item del17 was not completely investigated for DIF because no individuals from single-parent families responded using the highest category on del17. Not all items in the two antisocial behavior scales were investigated for DIF and DTF due to insufficient responses above the lowest category on the response scale. As a result, future research with larger samples that exhibit higher levels of antisocial behavior would be able to extend our research by examining DIF on the items we had to delete due to low frequencies of response. Finally, even for items with sufficient numbers of responses in higher categories, the frequency of these responses was not large. In such situations, a researcher might re-score item responses on a 0-to-3 scale into 0-1 scoring. Then, future research could investigate whether use of dichotomous IRT model10 including 2PLMs or 3PLMs to model these responses leads to approximately the same levels of measurement precision as does the use of polytomous IRT model10 such as the GRM.

Despite these limitations, the present study has practical implications for applied researchers: non-significant DTF can exist at the scale level, despite significant levels of item DIF on common items in two scales of antisocial behavior for adolescents and young adults. Using these results, we have a useful way of linking scores on the two antisocial behavior scales across age groups. Our study also illustrated how the two types of DIF—uniform and nonuniform—differ in terms of item and category parameters in GRMs. Given the non-significant levels of DTF on the total scale scores, comparing respondents from single- and two-parent families on raw scale scores appears justified.

Appendix A

Antisocial behavior items

During the past 12 months, how often have you:	1994	1995
Cut classes or stayed away from school without permission	del1	
Taken a car or other vehicle without the owner's permission, just to drive around	del2	
Snatched someone's purse or wallet without hurting them	del3	
Been drunk in a public place	del4	
Broke in or tried to break into a building just for fun or to look around	del5	
Broke in or tried to break into a building to steal or damage something	del6	
Thrown objects such as rocks or bottles at people to hurt or scare them	del7	
Set fire to a building or field or something like that just for fun	del8	
Sneaked into a movie, ballgame or something like that without paying	del9	
Steal money or take something that did not belong to you	**del10**	**del10**
Beat up on someone or fought someone physically because they made you angry	**del11**	**del11**
Purposely damaged or destroyed property that did not belong to you	**del12**	**del12**
Attack someone with a weapon trying to seriously hurt them	**del13**	**del13**
Sold illegal drugs such as pot, grass, has, LSD, cocaine, or other drug	**del14**	**del14**
Used a weapon, force or strong arm methods to get money or things from someone	**del15**	**del15**
Drive a car recklessly		del16
Cheat at school or other places		del17
Tell lies to people		del18
Sell stolen goods		del19
Write bad checks		del20
Use someone else's credit card without permission		del21

Appendix B

1994 Frequency and Descriptive Statistics (*N*=556)

		del1	del2	del3	del4	del5	del6	del7	del8	del9	del10	del11	del12	del13	del14	del15
	0	338	496	515	309	486	507	497	514	393	397	426	438	515	510	517
	1	59	16	4	58	22	7	16	5	58	46	44	39	2	6	3
	2	61	10	1	70	10	6	6	1	53	42	40	26	5	1	2
Freq.	3	26	0	2	34	4	2	3	2	9	14	6	12	0	0	0
	4	38	0	0	50	0	0	0	0	9	22	6	7	0	5	0
	Total	522	522	522	521	522	522	522	522	522	521	522	522	522	522	522
	Missing	34	34	34	35	34	34	34	34	34	35	34	34	34	34	34
	M	.79	.07	.02	.96	.10	.05	.07	.02	.43	.50	.32	.30	.02	.05	.01
	SD	1.26	.32	.22	1.36	.42	.30	.35	.23	.87	1.04	.76	.78	.20	.41	.15
	Skewness[a]	1.45	4.96	11.28	1.15	4.66	7.08	5.82	10.81	2.15	2.17	2.66	2.94	9.16	8.80	11.81
	Kurtosis[b]	.84	24.78	137.2	−.07	23.22	53.44	37.15	127.6	4.30	3.79	7.13	8.46	84.32	79.88	147.6

0 indicates no time in past year; 1 indicates 1 time in past year; 2 indicates 2–3 times in past year; 3 indicates 4–5 times in past year; 4 indicates more than 5 times in past year. [a]skewness for all variables has standard error values of .11 [b]kurtosis for all variables has standard error values of .21

Appendix C

1995 Frequency and Descriptive Statistics (*N*=556)

		del16	del17	del18	del19	del20	del21	del10	del11	del12	del13	del14	del15
	0	316	388	251	502	433	503	421	438	438	497	498	506
	1	44	41	37	3	24	4	32	31	30	5	1	2
	2	60	36	76	1	29	1	34	27	24	3	3	0
Freq.	3	35	23	58	0	13	0	14	6	9	3	4	0
	4	52	19	85	2	9	0	7	6	7	0	1	0
	Total	507	507	507	508	508	508	508	508	508	508	507	508
	Missing	49	49	49	48	48	48	48	48	48	48	49	48
	M	.94	.51	1.39	.03	.31	.01	.33	.25	.26	.04	.05	.00
	SD	1.39	1.01	1.57	.28	.83	.13	.83	.71	.75	.29	.36	.06
	Skewness[a]	2.64	3.24	3.21	8.35	8.31	15.89	1.18	2.08	.57	12.79	2.87	11.86
	Kurtosis[b]	6.42	10.89	10.34	73.66	71.34	251.48	−.10	3.27	−1.28	174.36	7.71	155.61

0 indicates no time in past year; 1 indicates 1 time in past year; 2 indicates 2–3 times in past year; 3 indicates 4–5 times in past year; 4 indicates more than 5 times in past year. [a]skewness for all variables has standard error values of .11 [b]kurtosis for all variables has standard error values of .22

References

Baker, F. B. (1993). *EQUATE 2.0: A computer program for equating two metrics in item response theory*. Madison: University of Wisconsin, Laboratory of Experimental Design.

Beyers, J. M., & Loeber, R. (2003). Untangling developmental relations between depressed mood and antisocial behavior in male adolescents. *Journal of Abnormal Child Psychology, 31*, 247–266.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253–260.

Conger, R. D., & Conger, K. J. (2002). Resilience in Midwestern families: selected findings from the first decade of a prospective, longitudinal study. *Journal of Marriage and Family, 64*, 361–373.

Curran, P. J. (2000). A latent curve framework for the study of developmental trajectories in adolescent substance use. In J. S. Rose, L. Chassin, C. C. Presson & S. J. Sherman (Eds.), *Multivariate applications in substance use research: New methods for new questions* (pp. 1–42). Mahwah, NJ: Lawrence Erlbaum Associates.

Dawson, D. A. (1991). Family structure and children's health and well-being: Data from the 1988 national health interview survey on child health. *Journal of Marriage and the Family, 53*, 573–584.

Dishion, T. J., & McMahon, R. J. (1998). Parental monitoring and the prevention of child and adolescent problem behavior: a conceptual and empirical formulation. *Clinical Child and Family Psychology Review, 1*, 61–75.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale NJ: Erlbaum.

Dunifon, R., & Kowaleski-Jones, L. (2002). Who's in the House? Race differences in cohabitation, single parenthood, and child development. *Child Development, 73*, 1249–1264.

Ge, X., Natsuaki, M. N., & Conger, R. D. (2006). Trajectories of depressive symptoms and stressful life events among male and female adolescents in divorced and nondivorced families. *Development and Psychopathology, 18*, 253–273.

Elliott, D. S., Huizinga, D. E., & Ageton, S. S. (1985). *Explaining delinquency and drug use*. Beverly Hills, CA: Sage.

Elliott, D. S., Huizinga, D. E., & Menard, S. (1989). *Multiple problem youth: Delinquency, substance use, and mental health problems*. New York: Springer-Verlag.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Earlboum Associates.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272–299.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: comparsion with Mantal-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278–295.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage: Newbury Park, CA.

Hankin, B. L., & Abramson, L. Y. (2002). Measuring cognitive vulnerability to depression in adolescence: reliability, validity and gender differences. *Journal of Clinical Child and Adolescent Psychology, 31*, 491–504.

Hoffmann, J. P. (2006). Family structure, community context, and adolescent problem behaviors. *Journal of Youth Adolescence, 35*, 867–880.

Huizinga, D., & Elliott, D. S. (1986). Reassessing the reliability and validity of self-report delinquency measures. *Journal of Quantitative Criminology, 2*, 293–327.

Huizinga, D., Esbensen, F. A., & Weiher, A. W. (1991). Are there multiple paths to delinquency? *Journal of Criminal Law and Criminology, 82*, 83–118.

Kim, S. H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education, 8*, 291–312.

Kim, S. H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*, 93–116.

Krueger, R. F., Chentsova-Dutton, Y. E., Markon, K. E., Goldberg, D., & Ormel, J. (2003). A cross-cultural study of the structure of comorbidity among common psychopathological syndromes in the general health care setting. *Journal of Abnormal Psychology, 112*, 437–447.

Lagenbucher, J. W., Labouvie, E., Martin, C. S., Sanjuan, P. M., Bavly, L., Kirisci, L., & Chung, T. (2004). An application of item response theory analysis to alcohol, cannabis, and cocaine criteria in DSM-IV. *Journal of Abnormal Psychology, 113*, 72–80.

Loeber, R., Farrington, D. P., Stouthamer-Loeber, M., Moffitt, T. E., & Caspi, A. (1998). The development of male offending: key findings from the first decade of the Pittsburgh youth study. *Studies on Crime and Crime Prevention, 7*, 141–71.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillside, NJ: Erlbaum Associates.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychological Methods, 9*, 354–368.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297–334.

Moffitt, T. E., Caspi, A., Dickson, N., Silva, P., & Stanton, W. (1996). Childhood-onset versus adolescent-onset antisocial conduct problems in males: natural history from ages 3 to 18Years. *Development and Psychopathology, 8*, 399–424.

Piquero, A. R., Macintosh, R., & Hickman, M. (2002). The validity of a self-reported delinquency scale: comparisons across gender, age, race, and place of residence. *Sociological Methods and Research, 30*, 492–529.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement, 19*, 353–368.

Ramsay, J. O. (2001). TestGraf. A program for the graphical analysis of multiple-choice tests and questionnaire data [Computer software and manual]. Retrieved from http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27–48.

Reise, P. R., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566.

Ruchkin, V., Sukhodolsky, D. G., Vermeiren, R., Koposov, R. A., & Schwab-Stone, M. (2006). Depressive symptoms and associated psychopathology in urban adolescents: a cross-cultural study of three countries. *Journal of Nervous and Mental Disease, 194*, 106–113.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement Number, 17*, 34. Part 2.

Sanders, D. E., Merrel, K. W., & Cobb, H. C. (1999). Internalizing symptoms and affect of children with emotional and behavioral disorders: a comparative study with an urban African American sample. *Psychology in the Schools, 36*, 187–197.

SAS Institute, Inc. (2003). *SAS/STAT software: Changes and enhancements, release 9.1*. Cary, NC: SAS Institute, Inc.

Segall, D. O. (1983). *Test characteristic curves, item bias, and transformation to a common metric in item response theory: A methodological artifact with serious consequences and a simple solution*. Unpublished manuscript, University of Illinois, Department of Psychology, Urbana-Champaign.

Sijtsma, K. (1998). Methodology review: nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22*, 3–32.

Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications. *Medical Care, 44*, S39–S49.

Thissen, D. (2001). *IRTLR v. 2.0 b: Software for the computation of the statistics involved in item response theory likelihood-ratio*

tests for differential item functioning. Chapel Hill: University of North Carolina.

Thissen, D. (2003). *MULTILOG user's guide*. Mooresville, IN: Scientific Software.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*, 385–395.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: the concept of item bias. *Psychological Bulletin, 99*, 118–128.

Thornberry, T. P., Krohn, M. D., Lizotte, A. J., & Chard-Wierschem, D. (1993). The role of juvenile gangs in facilitating delinquent behavior. *The Journal of Research in Crime and Delinquency, 30*, 55–87.

Thornberry, T. P., & Krohn, M. D. (2000). The self-report method for measuring delinquency and crime. *Criminal Justice, 4*, 33–83.

Wickrama, K. A. S., Conger, R. D., Wallace, L. E., & Elder, G. H., Jr. (2003). Linking early social risks to impaired physical health during the transition to adulthood. *Journal of Health and Social behavior, 44*, 61–74.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). *BILOG-MG 3*. Lincolnwood, IL: Scientific Software International.