Measurement and Structural Models for Children’s Problem Behaviors

Yuk Fai Cheong  
Emory University  

Stephen W. Raudenbush  
University of Michigan

This article considers an analytic strategy for measuring and modeling child and adolescent problem behaviors. The strategy embeds an item response model within a hierarchical model to define an interval scale for the outcomes, to assess dimensionality, and to study how individual and contextual factors relate to multiple dimensions of problem behaviors. To illustrate, the authors analyze data from the primary caregiver ratings of 2,177 children aged 9-15 in 79 urban neighborhoods on externalizing behavior problems using the Child Behavior Checklist 4-18 (T. M. Achenbach, 1991a). Two subscales, Aggression and Delinquency, are highly correlated, and yet unidimensionality must be rejected because these subscales have different associations with key theoretically related covariates.

In studies of social development in children and adolescents, researchers commonly use multiitem scales for behavioral assessment. On the basis of item responses obtained from interviews or direct observations, they compute raw and standardized scale scores as outcomes to assess the genetic and environmental influences on problem behaviors (e.g., Gjone & Stevenson, 1997), to identify risk factors (e.g., Kupersmidt & Patterson, 1991), to identify at-risk children for treatment (e.g., Huesmann et al., 1996), to classify children into different levels of seriousness in a particular problem behavior (e.g., Peeples & Loebel, 1994), and to evaluate the effectiveness of preventive intervention programs (e.g., Nadel, Laussell-Bryant, Spellmann, Landsberg, & Alvarez-Canino, 1996).

Researchers disagree, however, on how to conceive and quantify problem behaviors. Some advocate use of a single general construct, whereas some adopt a more differentiated approach to operationalize and measure a set of problem behaviors (Loebel, Farrington, Stouthamer-Loebel, & Van Kammen, 1998). Still other researchers recommend a more flexible approach that allows the choice between the two different specifications to depend on the purposes of the study. This latter approach is reflected in Achenbach’s (1993) recommendations regarding the use of the various multiitem scale scores obtained from the widely used Child Behavior Checklist (CBCL) and its supplements (Achenbach, 1991a, 1991b, 1991c). However, the research literature is not exactly clear on how to decide which approach is preferable, nor is it clear on how to incorporate a multidimensional conception of behavior into the context of multilevel ecological studies. The present article aims to address these interrelated concerns by proposing an analytic framework for measuring and modeling the multiple aspects of children’s problem behaviors. To illustrate the dilemma, we address the choice between viewing externalizing behaviors of the CBCL either as a single dimension or as constituting two distinct subscales.

In its entirety, the CBCL is intended to measure two second-order or broadband factors—Internalizing and Externalizing. The Internalizing factor is com-

---

1 For brevity, we refer to both children and adolescents as children.
posed of three scales—Withdrawn, Somatic Complaints, and Anxious/Depressed. The Externalizing factor is composed of two narrowband factors—Delinquent Behavior and Aggressive Behavior (Achenbach, 1991a). Internalizing and externalizing behaviors differ from each other in the level of inhibition and anxiety they reflect; internalizing reflects undue anxiety and inhibition, and the latter reflects a lack of these qualities (Greenbaum & Dedrick, 1998).

In a guide for using syndromes and profile types derived from the CBCL and its supplements, Achenbach (1993) maintained that Externalizing and Internalizing scores distinguish children who display problems from one broadband factor or the other. Such information may facilitate clinical decisions regarding individual children as well as research investigating distinct and common causes of the subscales that constitute the two groupings. The use of the narrowband factors or subscales, however, yields more differentiated data for assessment and research than does the internalizing–externalizing classification. Thus, researchers can choose to use scores at various levels of generality according to the purposes of their research. Verhulst and van der Ende (1991) exercised this option with the CBCL in their study of a 6-year developmental course of internalizing and externalizing problem behavior. They used the scores for the subscales, the Internalizing and Externalizing scales, and the overall total.

The legitimacy of the option to choose scores at different levels of generality relies on the existence of a general construct underlying all the behaviors in a set. Thus, according to this framework, delinquent and aggressive behaviors constitute a single behavioral dimension of externalizing problems—a notion not shared by all researchers (Frick et al., 1993; Loeber, 1988; White & Labouvie, 1994).

The debate about dimensionality presumes the existence of a meaningful scale of measurement for each hypothesized dimension. However, researchers have expressed reservations about the meaningfulness of the CBCL scales. For instance, Drotar, Stein, and Perrin (1995) argued that inasmuch as validation studies for the CBCL and its related instruments have relied on use of extreme scores for clinical diagnoses, there is inherent ambiguity in the meaning of scores that lie within the normal range regarding their clinical significance and the interpretation of their variability.

In this article, we consider an analytic strategy for conceiving and quantifying problem behaviors. The approach enables researchers to construct meaningful interval scales for multiple dimensions, to test dimensionality of scales, and to model context effects on multiple dimensions of problem behaviors. The approach borrows item response models from educational testing (Lord, 1980; Rasch, 1980) and embeds these models in a hierarchical model (Bryk & Raudenbush, 1992; Goldstein, 1995; Hedeker & Gibbons, 1996; Longford, 1993) that incorporates multiple levels of children’s ecological context. We apply the strategy to study primary caregiver ratings on externalizing behavior problems of 2,177 Chicago urban children aged 9–15 in 79 neighborhoods, collected by the Project of Harman Development in Chicago Neighborhoods (PHDCN) in 1994–1997 to decide on the appropriateness of a separate versus joint analysis of aggression and delinquency. The PHDCN is a prospective study that investigates how individual risk factors interact with features of social organization to shape propensity for antisocial development.

Item Response Theory

Item response theory postulates that characteristics of a test item, such as its difficulty, interact with a person’s ability or trait to determine the probability of a correct response to that item (Lord, 1980). Assuming all items represent the same ability domain, difficult items will be answered correctly less often than will easy items. Similarly, given the difficulty of the item, more able examinees will obtain a correct response with higher probability than will less able examinees. If the model is sensible, it will generate an interval scale along which every item and every examinee can be located. A visual examination of this “item map” provides useful clues about the construct validity of the test, because one can assess whether the empirically estimated item difficulties conform to cognitive theory regarding the sources of item difficulty. It is also possible to identify misfitting items (e.g., difficult items frequently solved by persons of low ability) and misfitting persons (e.g., able persons who frequently miss easy items). Such analyses form a basis for discarding poor items and assessing the overall quality of the scale. The analysis produces a measure of scale reliability and a standard error of

---

2 In the PHDCN survey, we collected data on self-ratings by the children as well. Whereas it is also of interest to model differential functioning of scales depending on the reporter (primary caregiver vs. participant), we forgo this option for the sake of simplicity.
measurement for each examinee (Wright & Stone, 1979).

Of course, a cognitive test is quite different from a behavioral rating scale, yet such a rating scale incorporates items that reflect varying severity in the underlying construct, and the children being assessed are presumed to differ quantitatively with respect to aggressiveness, delinquency, anxiety, and so forth. Thus, item difficulty and person ability in testing have clear analogues in behavioral assessment. Not surprisingly, then, researchers increasingly have adopted item response analysis in behavioral assessment (Kindlon, Wright, Raudenbush, & Earls, 1996; Wright & Masters, 1982; Wright & Stone, 1979).

Recently, there has been ongoing research in expanding item response models to include more than one measurement trait. For a succinct historical account of multidimensional item response theory both as a further development of unidimensional item response theory and as a special instance of factor analysis or structural equation modeling for discrete variables, see Reckase (1997).

Hierarchical Models

Hierarchical models, also referred to as multilevel models (Goldstein, 1995; Mason, Wong, & Entwistle, 1983) or random coefficient models (Hedeker & Gibbons, 1996; Longford, 1993; Rosenberg, 1973), address the clustered character of many data sets in social sciences. For instance, in education, students are commonly nested within classrooms, schools, districts, or program sites such that responses within groups are context dependent. Furthermore, when students are administered multiple items measuring one or more traits, dependent item responses from the same child constitute another level, which is nested within students themselves. By explicitly incorporating the clustered structure, hierarchical models allow one to study variability at each level, to evaluate effects within and across different levels, and to avoid the problem of misestimated precision that results from the failure to include components of variance and covariance arising from grouping effects.

Incorporating Item Response Models in a Multilevel Setting

Item response models can be embedded in a hierarchical model to represent measurement errors in a multilevel setting. The models can consist of item responses as Level 1 units and people as Level 2 units. Adams, Wilson, and Wu (1997) and Kamata (1998) discussed and illustrated uses of these two-level models in educational testing. In their studies, an item response model, cast as a within-student model, was used to describe the probability of observing a pattern of item responses given the ability level of a student. The within-student model was combined with a between-student model that investigated the variability among students in the ability level. In the present analysis, we define person-specific traits and item severities at Level 1 and study the person- and neighborhood-level correlates of these traits at Levels 2 and 3.

Using a three-level hierarchical framework, Raudenbush, Rowan, and Kang (1991) proposed and illustrated a similar three-level model. They studied the multiple item responses of teachers nested within schools to assess the statistical properties of school climate perceptions. However, their model assumed Likert responses to lie on an interval scale. Raudenbush and Sampson (1999) formally incorporated item response models in measuring neighborhood characteristics. In these prior studies, the object of measurement was the ecological unit (school or neighborhood). In the present article, the object of measurement is the child, and we seek to study multilevel effects on latent behavioral dimensions of children.

Criteria for Assessing Dimensionality

Theoretical Basis and Empirical Research

Our approach to assessing dimensionality of a measure supposed that the measure has been developed and refined on a theoretical basis and tested in empirical studies. The study of the grouping of externalizing and internalizing syndromes, for instance, has drawn on theoretical perspectives and clinical research from developmental psychopathology (Cicchetti & Toth, 1991). Over more than 30 years, as Achenbach (1991a) reviewed, researchers have studied the two sets of problem behaviors under various names including personality problems versus conduct problems, inhibition versus aggression, and overcontrolled versus undercontrolled behaviors. Such distinctions have been identified in numerous multivariate analyses of children's behavioral and emotional problems.

Interrelationships Among Behavior Items and Scales

Most of the multivariate analyses Achenbach (1991a) referred to in identifying behavioral dimen-
sions for various constructs in CBCL or of problem behaviors in general are factor analytic studies. There are two main types of factor analysis—exploratory and confirmatory—that are used to examine the extent of the interrelationships among the behavioral items as well as subscales. Achenbach (1991a), for example, used an exploratory factor analytic approach to identify the first- and second-order factors in CBCL. Dedrick, Greenbaum, Friedman, Wetherington, and Knoff (1997) and De Groot, Koot, and Verhulst (1994) used confirmatory factor analysis to evaluate and to lend support to Achenbach’s (1991a) first-order factor model. Greenbaum and Dedrick (1998) applied confirmatory factor analysis to test the broadband structure of the CBCL/4-18, and they maintained that the fits of the models supported Achenbach’s (1993) grouping of the syndromes.

As mentioned earlier, Frick et al. (1993) disagreed with this classification scheme. Their analysis illustrates a key point in the assessment of dimensionality. High intercorrelations among subscales are a necessary but not sufficient condition to assert unidimensionality. It is also essential that the subscales relate identically or, at least, very similarly, to theoretically linked covariates.

In the case of the externalizing dimension, there is some reason to suspect that its subscales, Delinquency and Aggression, might relate differently to age. There is a large literature indicating that delinquent propensity increases rapidly during early adolescence (e.g., Hirschi & Gottfredson, 1983). There is not a corresponding literature suggesting such a strong positive age effect for aggression. In fact, the manual for CBCL (Achenbach, 1991a) provides plots of the association between age and prevalence, and these age-prevalence plots tend to differ between the Aggression and Delinquency items. Some Aggression items (e.g., bragging, temper tantrums) appear to indicate social immaturity and decline with age during adolescence. In contrast, prevalence on Delinquency items tends to increase with age, as suggested by literature on the age-crime curve (Gottfredson & Hirschi, 1990). Under these conditions, a high correlation between the Aggression and Delinquency subscales would not be sufficient to warrant the conclusion that these constitute a single dimension. Indirect evidence of a differential association with age would require us to reject unidimensionality and to model the age effect, if any, for each subscale separately to study how risk factors at various ecological levels may interact with the age effect.

A differential association between age and subscales also leads one to hypothesize that social contextual effects, especially the effects of the neighborhood social environment, would differ across subscales. As children grow older, their social interactions occur in an ever-expanding spatial realm, exposing them to peer influences beyond the home and classroom and creating an increased role for neighborhood effects on antisocial behavior (Reiss, 1986). We therefore hypothesized that if age effects are more pronounced for delinquency than for aggression, neighborhood effects would similarly be more pronounced for delinquency than for aggression. Such a finding would undermine the case for using the broad-based Externalizing scale as an outcome in studies of neighborhood effects on antisocial behavior. It would support, instead, a more sharply focused study of neighborhood effects on delinquency.

In addition to differential age effects, the two subscales may relate differently to gender. The criminology literature suggests that gender differences exist for most types of delinquent behaviors at all ages (Steffensmeier & Allan, 1995). In contrast, gender differences in aggression appear to depend on the age of the child and the type of expression of aggression (Loeber & Stouthamer-Loeber, 1998). Results indicating differential gender effects, if any, would lend further support to treat the two subscales as distinct.

We assume that the basic subscales, Delinquency and Aggression, are sound and that their factor analytic structures are invariant over age and gender, as reflected in how the two CBCL scales have been studied, validated, and widely used by researchers and practitioners (see, e.g., Dedrick et al., 1997; Greenbaum & Dedrick, 1998). On the basis of these assumptions, we investigated whether the two subscales can be meaningfully collapsed into a single externalizing construct for assessment and modeling purposes. It would be perfectly reasonable for one to further evaluate our assumption that the same definitions of aggression or delinquency hold for children at different ages and genders. In the final section of this paper, we consider how this evaluation might proceed (see the Discussion section).

Summary

An evaluation of dimensionality presumes a meaningful measurement scale for each hypothesized dimension. Unidimensionality also requires that hypothesized subscales not only be strongly correlated but also that they relate similarly to risk factors. When
the relevant risk factors are defined at different ecological levels, a multilevel strategy is required to assess dimensionality. These requirements suggest a modeling strategy that embeds a multiple item response model in a hierarchical structural model. We propose and test such a model in this article.

Method and Results

Research Design

The analyses are based on a subsample of the 1994–1997 first wave data of the PHDCN. Using a spatial definition of neighborhood—a collection of people and institutions occupying a subsection of a larger community—the project collapsed 847 census tracts in the city of Chicago to form 343 neighborhood clusters (NCs). The predominant guideline in formation of the NCs was that they should be as ecologically meaningful as possible, composed of geographically contiguous census tracts, and internally homogeneous on key census indicators. The project settled on an ecological unit of about 8,000 people, which is smaller than the 77 established community areas in Chicago (the average size is almost 40,000 people) but large enough to approximate local neighborhoods. Geographical boundaries (e.g., railroad tracks, parks, and freeways) and knowledge of Chicago’s neighborhoods guided this process.

The extensive racial, ethnic, and social-class diversity of Chicago’s population was a major criterion in its selection as a research site. Table 1 shows the classifications of the 343 NCs according to ethnicity and a trichotomized measure of socioeconomic status (SES) constructed from indicators of poverty, public assistance, income, and education in the 1990 census. Although there were no low-SES White neighborhoods and no high-SES Latino neighborhoods, there were Black neighborhoods in all three cells of SES, and many heterogeneous neighborhoods varied in SES. Table 1 at once thus confirms the racial and ethnic segregation and yet rejects the common stereotype that minority neighborhoods in the United States are homogeneous.

As part of the longitudinal study of the antisocial development in children, 800–900 participants in each of seven age groups (0–1, 3, 6, 9, 12, 15, 18) were sampled from households in 80 of the 343 NCs. Home-based interviews with parents and children in each cohort were conducted over 30 months from 1994–1997. The 80 NCs were sampled from the 21 strata defined in Table 1 with the aim of representing these 21 cells as nearly equally as possible to eliminate the confounding between ethnic mix and SES. Each family will be followed over a period of 5–7 years for repeated assessments.

Table 2 describes the background of the children and the primary caregivers and a characteristic of the neighborhoods. The average age of the children in the sample was 12.0 years. Nearly half of them were girls, and 43% of the primary caregivers had finished high school. Latino primary caregivers made up 43% of the sample; African Americans 37%; Native Americans, Pacific Islanders, and Other 6%; whereas the remaining 15% were White.

Table 1

<table>
<thead>
<tr>
<th>Racial–ethnic stratum</th>
<th>SES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>&gt;75% Black</td>
<td>77 (9)</td>
</tr>
<tr>
<td>&gt;75% White</td>
<td>0 (0)</td>
</tr>
<tr>
<td>&gt;75% Latino</td>
<td>12 (4)</td>
</tr>
<tr>
<td>≥20% Latino and ≥20% White</td>
<td>6 (4)</td>
</tr>
<tr>
<td>≥20% Latino and ≥20% Black</td>
<td>9 (4)</td>
</tr>
<tr>
<td>≥20% Black and ≥20% White</td>
<td>2 (2)</td>
</tr>
<tr>
<td>NCs not classified above</td>
<td>8 (4)</td>
</tr>
<tr>
<td>Total</td>
<td>114 (27)</td>
</tr>
</tbody>
</table>

Note: The 80 sampled NCs are shown in parentheses. SES was defined by a six-item scale that summed standardized neighborhood-level measures of median income, percentage college educated, percentage with household income over $50,000, percentage of families below the poverty line, percentage on public assistance, and percentage with household income less than $50,000 on the basis of the 1990 decennial census. In forming the scale, the last three items were reverse coded. PHDCN = Project of Human Development in Chicago Neighborhoods.
Table 2
Descriptive Statistics for Person- and NC-Level Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Range</th>
<th>M</th>
<th>SD</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person-level data (N = 2,177)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>7.8–16.9</td>
<td>12.0</td>
<td>2.4</td>
<td>.500</td>
</tr>
<tr>
<td>Female</td>
<td>0 = male, 1 = female</td>
<td>.495</td>
<td></td>
<td>.500</td>
</tr>
<tr>
<td></td>
<td>0 = less than high school, 1 = otherwise</td>
<td>.426</td>
<td></td>
<td>.496</td>
</tr>
<tr>
<td>Latino</td>
<td>0 = no, 1 = yes</td>
<td>.426</td>
<td></td>
<td>.493</td>
</tr>
<tr>
<td>Black</td>
<td>0 = no, 1 = yes</td>
<td>.366</td>
<td></td>
<td>.482</td>
</tr>
<tr>
<td>Other (Pacific Islander, Asian and Native Americans, and other)</td>
<td>0 = no, 1 = yes</td>
<td>.055</td>
<td></td>
<td>.228</td>
</tr>
<tr>
<td>NC-level data (N = 79)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concentrated disadvantage</td>
<td>-1.596–2.663</td>
<td>-0.101</td>
<td>0.793</td>
<td></td>
</tr>
</tbody>
</table>

Note. The concentrated disadvantage variable is constructed from the 1990 decennial census and is based on the percentage of households below the poverty line, percentage of residents on public assistance, percentage of unemployed people, percentage of Black families, percentage of female-headed families, and percentage of children under age 18. NC = neighborhood cluster.

Neighborhood-level concentrated disadvantage is a weighted factor score constructed from the 1990 decennial census. The concentrated disadvantage scale was created from six variables indicating (a) percentage of households that were below the poverty line, (b) percentage of residents in receipt of public assistance, (c) percentage of unemployed people, (d) percentage of Black residents, (e) percentage of female-headed families, and (f) density of children. Sampson, Raudenbush, and Earls (1997) provided a detailed description of the construction of the scale.

Instrument and Measures

The CBCL/4-18 (Achenbach, 1991a) was used to assess externalizing behavior problems. The CBCL/4-18 consists of 113 items. An informant is asked to rate the extent to which an item describes a child’s behavior now or within the past 6 months. Response options are on a 3-point scale (0 = not true, 1 = somewhat or sometimes true, 2 = very true or often true). We combined the last two categories of the 3-point continuum of the items because responses to most items behaved like dichotomous ones with 0 as the modal category. Thus, in the recoded scale, never true = 0 and ever true = 1.

In the checklist, the Aggressive Behavior scale is composed of 20 items and the Delinquent Behavior scale is composed of 13 items. Table 3 gives the frequency distribution of the items for the two multitem scales that constitute the externalizing grouping, arranged in decreasing proportion of affirmative responses. Note that the items behave essentially as one might expect. Less serious indicators of aggressive behaviors—for instance, arguing and being stubborn—are endorsed more frequently than are serious indicators—physical attack and threat—with the temper tantrums being rated with moderate frequency. The same pattern occurs for the indicators of delinquent behaviors.

Modeling Procedure

To begin, we seek to understand how the items function within each construct and to use this information to build an interval scale for each narrowband factor by using an item response model. As mentioned earlier, in the analogy with ability testing, each child is an examinee, each indicator of syndrome is an item, and a correct response occurs when a child receives an affirmative rating on that item. In this setting, item difficulty is the severity of the indicator of a problem behavior, and ability corresponds to participant’s level of aggression or delinquency.

To achieve this purpose, the item response model is recast as a three-level logistic regression model that embeds its item response portion within a hierarchical structure in which the secondary units of measurement, the children, are nested within the NCs. The combined model extends the usual item response model also in allowing for multiple characteristics to be measured—in this case, aggression and delinquency—rather than a single, unidimensional trait, and in allowing for randomly missing responses. The model also allows researchers to gauge the variability of the log-odds of the endorsement of aggression and
### Table 3

**Item Responses for Person-Level Scales**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Frequency</th>
<th>0 = Never</th>
<th>1 = Ever</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aggressive Behavior</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argues a lot</td>
<td>782</td>
<td>1,394</td>
<td></td>
</tr>
<tr>
<td>Stubborn, sullen, or irritable</td>
<td>1,003</td>
<td>1,171</td>
<td></td>
</tr>
<tr>
<td>Bragging, boasting</td>
<td>1,054</td>
<td>1,119</td>
<td></td>
</tr>
<tr>
<td>Demands a lot of attention</td>
<td>1,092</td>
<td>1,081</td>
<td></td>
</tr>
<tr>
<td>Talks too much</td>
<td>1,104</td>
<td>1,069</td>
<td></td>
</tr>
<tr>
<td>Easily jealous</td>
<td>1,189</td>
<td>983</td>
<td></td>
</tr>
<tr>
<td>Showing off or clowning</td>
<td>1,203</td>
<td>969</td>
<td></td>
</tr>
<tr>
<td>Disobedient at home</td>
<td>1,284</td>
<td>889</td>
<td></td>
</tr>
<tr>
<td>Teases a lot</td>
<td>1,298</td>
<td>875</td>
<td></td>
</tr>
<tr>
<td>Temper tantrums or hot temper</td>
<td>1,331</td>
<td>843</td>
<td></td>
</tr>
<tr>
<td>Sudden changes in mood or feeling</td>
<td>1,402</td>
<td>771</td>
<td></td>
</tr>
<tr>
<td>Unusually loud</td>
<td>1,453</td>
<td>721</td>
<td></td>
</tr>
<tr>
<td>Screams a lot</td>
<td>1,548</td>
<td>628</td>
<td></td>
</tr>
<tr>
<td>Disobedient at school</td>
<td>1,579</td>
<td>591</td>
<td></td>
</tr>
<tr>
<td>Cruelty, bullying, or meanness to others</td>
<td>1,790</td>
<td>383</td>
<td></td>
</tr>
<tr>
<td>Destroys things belonging to his/her family or other children</td>
<td>1,873</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>Gets in many fights</td>
<td>1,882</td>
<td>293</td>
<td></td>
</tr>
<tr>
<td>Destroys his/her own things</td>
<td>1,893</td>
<td>281</td>
<td></td>
</tr>
<tr>
<td>Physically attacks people</td>
<td>2,045</td>
<td>130</td>
<td></td>
</tr>
<tr>
<td>Threatens people</td>
<td>2,044</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td><strong>Delinquent Behavior</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prefers being with older kids</td>
<td>1,366</td>
<td>809</td>
<td></td>
</tr>
<tr>
<td>Doesn’t seem to feel guilty after misbehaving</td>
<td>1,446</td>
<td>720</td>
<td></td>
</tr>
<tr>
<td>Lying or cheating</td>
<td>1,482</td>
<td>689</td>
<td></td>
</tr>
<tr>
<td>Swearing or obscene language</td>
<td>1,721</td>
<td>451</td>
<td></td>
</tr>
<tr>
<td>Hangs around with children who get into trouble</td>
<td>1,770</td>
<td>402</td>
<td></td>
</tr>
<tr>
<td>Truancy, skips school</td>
<td>2,062</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td>Thinks about sex too much</td>
<td>2,065</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>Steals at home</td>
<td>2,093</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>Runs away from home</td>
<td>2,103</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Steals outside the home</td>
<td>2,121</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>Sets fire</td>
<td>2,128</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Uses alcohol or drugs for nonmedical purposes</td>
<td>2,145</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Vandalism</td>
<td>2,151</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

delinquent behaviors at the child level as well as at the NC level. It provides information on how those factors correlate with one another and estimates of useful psychometric properties for the factors as well.

This baseline model constitutes an unconditional three-level hierarchical logistic regression. The Level 1 units are item responses within children, the Level 2 units are children, and the Level 3 units are NCs. The model is unconditional as it does not include any person- or NC-level explanatory variables. To investigate the patterns of age-related changes of the two factors and the relationships of the factors to various risk factors and covariates, we then add the relevant person- and NC-level explanatory variables to the unconditional model and evaluate their associations with the two measures.

**Unconditional Model**

**Level 1 model.** The Level 1 model represents predictable and random variation among item responses within each child. This is a one-parameter item re-
response model and might be termed a Rasch model (Rasch, 1980) with random effects.

Level 1 sampling model. Let \( Y_{ijk} \) take on a value of unity if the \( i \)th response is endorsed for child \( j \) of neighborhood \( k \), with \( Y_{ijk} = 0 \) if not; and \( \mu_{ijk} \) denote the probability that \( Y_{ijk} = 1 \). This probability varies randomly over participants. However, conditioning on this probability, we have \( Y_{ijk} \mid \mu_{ijk} \sim \text{Bernoulli}(\mu_{ijk}) \), \( \text{E}(Y_{ijk} \mid \mu_{ijk}) = \mu_{ijk} \), \( \text{Var}(Y_{ijk} \mid \mu_{ijk}) = \mu_{ijk}(1 - \mu_{ijk}) \). As is standard in logistic regression, we define \( \eta_{ijk} \) as the log-odds of the probability that the \( i \)th response is affirmative for child \( j \) of neighborhood \( k \). Thus, we have

\[
\eta_{ijk} = \log\left( \frac{\mu_{ijk}}{1 - \mu_{ijk}} \right).
\]

The structural model at Level 1 accounts for predictable variation within children across responses. It views the log-odds of endorsement on response \( i \) as depending on which aspect of externalizing behavior problems is of interest (aggression or delinquency) and which specific item is involved. Let \( D_{AGG_{ijk}} \) take on a value of 1 if the \( i \)th response is to an item measuring aggression and a value of 0 otherwise, and let \( D_{DEL_{ijk}} = 1 - D_{AGG_{ijk}} \) similarly indicate whether that response is to an item measuring delinquency. Then we have

\[
\eta_{ijk} = D_{AGG_{ijk}} \left( \pi_{AGG_{ijk}} + \sum_{m=1}^{19} \alpha_{mijk} X_{mijk} \right) + D_{DEL_{ijk}} \left( \pi_{DEL_{ijk}} + \sum_{m=1}^{12} \delta_{mijk} Z_{mijk} \right),
\]

where \( X_{mijk} \) and \( Z_{mijk} \) are indicator variables representing the items in the scales; each is centered around the mean of the corresponding scale (1/20, or 0.05, for the aggression scale and 1/13, or 0.0769, for delinquency); \( \pi_{AGG_{ijk}} \) and \( \pi_{DEL_{ijk}} \) are aggression and delinquency traits, respectively, defined as the adjusted log-odds of the endorsement of the behavior on a "typical item" for child \( j \) of NC \( k \); and \( \alpha_{mijk} \) and \( \delta_{mijk} \) reflect the difficulty or severity level of behavioral item \( m \) within the aggression and delinquency syndromes, respectively. Note that 19 item indicators represent the 20 aggression items and that 12 item indicators represent the 13 delinquency items. The item that is not included in each of the two scales serves as the reference item.

It may appear that the response index \( i \) and the item index \( m \) are redundant. However, both are needed to allow for the possibility of missing data. Thus, for example, let Item 1 (\( m = 1 \)) be "Argues a lot" and Item 2 (\( m = 2 \)) be "Stubborn, sullen, or irritable." If a participant failed to rate a child on the first item, his or her first response \( (i = 1) \) would be to Item 2 (\( m = 2 \)). The adjustment in the two intercepts, \( \pi_{AGG_{ijk}} \) and \( \pi_{DEL_{ijk}} \), is achieved by centering \( X_{mijk} \) and \( Z_{mijk} \) around the mean of the corresponding scale. As a result of the coding scheme, \( \pi_{AGG_{ijk}} \) and \( \pi_{DEL_{ijk}} \) are the expected log-odds of endorsement for the two subscales for child \( j \) in neighborhood \( k \) adjusted for differences among the children in \( X_{mijk} \) and \( Z_{mijk} \) arising from incomplete responses.

Level 2 model. The Level 2 model accounts for variation between children within NCs on the latent aggression and delinquency:

\[
\pi_{AGG_{ijk}} = \beta_{AGG} + \mu_{AGG_{ijk}} \]
\[
\pi_{DEL_{ijk}} = \beta_{DEL} + \mu_{DEL_{ijk}},
\]

where \( \beta_{AGG} \) and \( \beta_{DEL} \) are the intercepts for NC \( k \) on the two latent traits. The random effects \( \mu_{AGG_{ijk}} \) and \( \mu_{DEL_{ijk}} \) account for variation across children.

In constructing subscales, we adopt Achenbach's (1993) advice and define the two subscales Aggression and Delinquency of the broad-based externalizing domain. We used the option of equal weighting (Wright & Masters, 1982), and we required that an item is indicative of only one scale here. Two options exist for constructing a measurement model in addition to the choice of the two subscales and equal weights for each item. First, one could use an exploratory item factor analysis (Bock, Gibbons, & Muraki, 1988; Muraki & Carlson, 1995) or a statistically equivalent threshold model (Muthén, 1978, 1984) to identify the number of scales and to estimate weights. Second, one could choose a set of specific weights, which may be unequal, based on previous research and the theoretical knowledge and practical field experience of a researcher (e.g., Hoijtink, Roos, & Wilmink, 1999). Our model with equal weights is, in essence, a Rasch or one-parameter item response model that defines a "severity" level for each item within each scale. However, it is a two-dimensional Rasch model because each participant is viewed as possessing a true level of aggression and delinquency. Moreover, it is a Rasch model with random effects because it views the Aggression and Delinquency true scores to vary randomly over participants and across neighborhoods. It is also a confirmatory model in that the factor structure is specified a priori.
\( u_{DEL,a} \) are assumed bivariate normally distributed with zero means, person-level variances \( \tau_{AGG} \) and \( \tau_{DEL} \), and covariance \( \tau_{AGGDEL} \). We constrain the item severities, the values \( \alpha_{mjk} \) and \( \delta_{mjk} \), to be invariant across the children and across the NCs for the sake of parsimony; that is, \( \alpha_{mjk} = \alpha_m \) and \( \delta_{mjk} = \delta_m \), for all \( j, k \). By lifting the constraint, researchers can investigate whether individual item severity varies at the child level and/or NC level and can study its correlates. Such variation is known as differential item functioning, or DIF in the literature in educational testing (van der Linden & Hambleton, 1997).

**Level 3 model.** The Level 3 model accounts for variation between NCs on the latent means of the problem behaviors:

\[
\begin{align*}
\beta_{AGG} &= y_{AGG} + \nu_{AGG} \\
\beta_{DEL} &= y_{DEL} + \nu_{DEL},
\end{align*}
\]

where \( y_{AGG} \) and \( y_{DEL} \) are the grand mean levels of the two latent scores on aggression and delinquency, respectively, in Chicago neighborhoods. The random effects \( \nu_{AGG} \) and \( \nu_{DEL} \) are assumed bivariate normally distributed with zero means, NC-level variances \( \omega_{AGG} \) and \( \omega_{DEL} \), and covariance \( \omega_{AGGDEL} \).

**Combined model.** The above models can be combined through substitutions of terms and expressed as

\[
\eta_{ijk} = D_{AGGijk}(y_{AGG} + \sum_{m=1}^{19} \alpha_m X_{mijk} + u_{AGGjk})
\]

\[
+ \nu_{AGGk} + D_{DELijk}(y_{DEL} + \sum_{m=1}^{12} \delta_m Z_{mijk})
\]

\[
+ u_{DEL,k} + \nu_{DELk}.
\]

Equation 4 shows that the log-odds of endorsement of an item for a particular scale depends on item severities plus individual child and NC contributions to the aggression and delinquency traits.

**Results for the unconditional model.** Table 4 shows the partial results for the unconditional model. The estimates of the item severities, \( \alpha_m \) and \( \delta_m \), are not included here but are displayed in Figure 1 which we will discuss. The results show that the expected logit of endorsement of a typical aggression behavior item for a typical child is \(-1.060\) and that there is statistically significant between-child and between-NCs variation in the log-odds of an affirmative response. The predicted odds are \( \exp(-1.060) = .35 \). The odds of the endorsement to nonendorsement are thus 1 to 2.89. The corresponding probability is \( 1/(1 + \exp(1.060)) = .26 \). The logit of endorsement of a typical delinquent behavior item for a typical child is \(-3.149\), and there is also statistically significant between-child and between-NCs variation in the log-odds of an affirmative response. The predicted odds are \( \exp(-3.149) = .04 \). The odds of the endorsement to nonendorsement are 1 to 23.31. The corresponding probability is \( 1/(1 + \exp(3.149)) = .04 \), much less than for aggression.

The two syndromes manifest high intercorrelations at the between-children and between-NC level, .92 and .87, respectively—a result that suggests there may be a general single externalizing construct. These correlations are adjusted for measurement errors specified in the Level 1 sampling model. Note the

---

4 All the models in this study were analyzed with the module for three-level random effects logistic models of the Hierarchical Linear Model (HLM) program (Bryk, Raudenbush, & Congdon, 1996). Using the Windows version of the program, we proceeded according to the following procedure:

1. Specification of the outcome of the Level 1 sampling model, which was a participant’s response to an individual item of the two behavioral scales;
2. entering the indicators \( X_{mijk} \) and \( Z_{mijk} \) as the Level 1 predictors;
3. entering the relevant Level 2 and Level 3 predictors for the conditional models; and
4. specifying only the two intercepts, \( \pi_{AGG,i} \) and \( \pi_{DEL,k} \), as randomly varying.

In principle, the same models can be estimated using similar packages like MLwiN (Goldstein et al., 1998) and the macro GLIMMIX used in conjunction with the procedure PROC MIXED in SAS (Littell, Milliken, Stroup, & Wolfinger, 1997).

5 The formulas for the between-children and between-NC correlations are

\[
\rho_B = \text{corr}(\pi_{AGG,i} \pi_{DEL,k}) = \frac{\tau_{AGGDEL}}{\sqrt{\tau_{AGG}\tau_{DEL}}}
\]

and

\[
\rho_B = \text{corr}(\beta_{AGG,i} \beta_{DEL,k}) = \frac{\omega_{AGGDEL}}{\sqrt{\omega_{AGG}\omega_{DEL}}}
\]

respectively.
Table 4
Partial Results for the Unconditional Models

<table>
<thead>
<tr>
<th>CBCL subscale</th>
<th>Intercept</th>
<th>Variance</th>
<th>Average reliability</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>SE</td>
<td>t</td>
<td>df</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggression</td>
<td>-1.060</td>
<td>0.030</td>
<td>-24.70*</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>1.866 (0.069)</td>
<td>0.054 (0.022)</td>
<td>.83</td>
<td>.38</td>
</tr>
<tr>
<td>Delinquency</td>
<td>-3.149</td>
<td>0.056</td>
<td>-56.49*</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>1.817 (0.089)</td>
<td>0.064 (0.027)</td>
<td>.60</td>
<td>.36</td>
</tr>
</tbody>
</table>

Note. CBCL = Child Behavior Checklist 4-18; Coeff. = coefficient; Approx. = approximate; NC = neighborhood cluster; CI = confidence interval.
* p < .05.

The sampling variance of $z$ is approximately $\chi^2$ d.f., where $d$ is the vector of first derivatives of $z$ with respect to the variance-covariance, and $V$ is the estimated sampling variance-covariance matrix of the estimates of the variance-covariance terms.

As correlations tend to have rather skewed distributions it is recommended that one work with the Fisher-$z$ transformation rather than the correlation itself. Thus, in the case of between-children correlation one would first transform the correlation coefficient, $r$, to $z$ using the Fisher-$z$ transformation. The formula is:

$$z = \frac{1}{2} \log \left( \frac{1 + r}{1 - r} \right)$$
indicate the reliability with which we can discriminate between children within an NC. Between NCs, the corresponding average reliabilities are .38 and .36.\textsuperscript{7,8} The results suggest that one cannot distinguish among NCs in the general level of aggression and delinquency with high reliability, which may be due to the relatively small number of children per NC and the small between-cluster variance for both latent measures. Note also the very wide 95% confidence interval of .45–.98 for the correlation between mean aggression and mean delinquency at the neighborhood level. A benefit of viewing the two scales as unidimensional is thus that it allows one to pool the items together to construct a scale and to obtain higher internal consistency. However, this pooling requires that the subscales operate in a unidimensional fashion.

Figures 1 and 2 give the item maps or “rulers” for the two measures obtained from estimates of $\alpha_m$ and $\delta_m$. For aggression, the item “Argues a lot” was chosen as the reference item. All the other items display more severity or seriousness, and the item “Threatens people” is on the far end of the scale, suggesting that endorsement of this item reflects relatively severe aggression. The item distribution for delinquency seems

\begin{itemize}
  \item Item difficulties for delinquency.
\end{itemize}

\textsuperscript{7} An approximate formula for computing the reliabilities for the aggression measure at the child level is

$$
\lambda_{AGGjk} = \frac{\omega_{AGG} + \tau_{AGG}}{\omega_{AGG} + \tau_{AGG} + \frac{1}{n_j w_k}}, \tag{7}
$$

where $\lambda_{AGGjk}$ is the internal consistency of the aggression measure for child $j$ in NC $k$; $n_j$ is the number of items on aggression rated for child $j$ in NC $k$; and $w_k$ is the average within child $j$ in NC $k$ of $\mu_{jk}(1 - \mu_{jk})$ on aggression items. Thus, the reliability depends on the intra-NC correlation but also on the number of items per scale and the item severities. Our summary measure of reliability is the average of the person-level reliabilities.

\textsuperscript{8} An approximate formula for computing the reliability for the aggression measure at the NC level is

$$
\lambda_{AGG} = \frac{\text{Var}(\hat{\beta}_{AGG})}{\text{Var}(\hat{\beta}_{AGG})} = \frac{\omega_{AGG} + \tau_{AGG} + \frac{1}{n_j w_k}}{\omega_{AGG} + \frac{\tau_{AGG} + \frac{1}{n_j w_k}}{J_k}}, \tag{8}
$$

where $\lambda_{AGG}$ is the internal consistency of the aggression measure for NC $k$; $n_k$ is the average number of items per child in NC $k$; $J_k$ is the number of children sampled within NC $k$; and $w_k$ is the average within NC $k$ of $\mu_{jk}(1 - \mu_{jk})$ on aggression items. Thus, the reliability depends on the intra-NC correlation but also on the number of children sampled, the number of items per scale, and the item severities. The approximation is exact when all participants provide responses to all items. The summary measure of reliability at the NC level is the average of the NC-specific reliabilities.
to suggest there are two groups of clustered items, with the items having severities larger than 2 (6th to 13th items from the top of the scale) forming a group that generally involves status violations and breaking of laws, and the other group including those less severe items (1st to 5th items) that do not. By computing empirical Bayes residuals at the child or NC level, one can easily compute scores at the corresponding levels as well, which share the same metric used in Figures 1 and 2 (see, e.g., Raudenbush & Sampson, 1999).

**Conditional Models**

The results from the unconditional model suggest that the correlation between the latent aggression and delinquency variables is high, giving support to the notion that these subscales constitute a single dimension. However, this assumption of unidimensionality also requires that subscales have similar associations with theoretically linked variables such as age and neighborhood disadvantage. To test this assumption requires that such covariates be formally incorporated into multivariate models. Because these covariates are measured at different levels, such models must be hierarchically structured. We therefore consider such models, which we call the *conditional models* because they take into account or “condition on” the relevant person-and/or NC-level covariates.

The analysis proceeds in two stages. We first study a model that includes only person-level covariates and investigate how person-level covariates relate to the log-odds of endorsement of aggression and delinquency. Of particular interest are the age and gender effects for the two subscales. We then enter an NC-level covariate into the model to study how it relates to the subscales and whether it moderates age effects.\(^9\)

**Level 1 model.** The Level 1 model remains the same as given in Equation 1. It models the log-odds of endorsement for item \(i\) as a function of the type of externalizing behavior problem, aggression or delinquency, and the specific item involved.

**Level 2 model.** At the person level, child age, a quadratic term for child age (Age sq.), child sex, primary caregiver (PC) ethnicity, and PC educational level (Educ. lvl.) are entered.\(^11\) The Level 2 model for latent aggression, \(\pi_{AGG_{jk}}\), and latent delinquency, \(\pi_{DEL_{jk}}\), are as follows:

\[
\pi_{AGG_{jk}} = \beta_{AGG} + \theta_{AGG1}(Age)_{ijk} + \theta_{AGG2}(Age sq.)_{ijk} + \theta_{AGG3}(Female)_{ijk} + \theta_{AGG4}(Female_{jk}) + \theta_{AGG5}(Educ. lvl.)_{ijk} + \theta_{AGG6} (Latino)_{ijk} + \theta_{AGG7} (Black)_{ijk} + \theta_{AGG8} (Other)_{ijk} + \mu_{AGG_{jk}}
\]

\[
\pi_{DEL_{jk}} = \beta_{DEL} + \theta_{DEL1}(Age)_{ijk} + \theta_{DEL2}(Age sq.)_{ijk} + \theta_{DEL3} (Female)_{ijk} + \theta_{DEL4} (Educ. lvl.)_{ijk} + \theta_{DEL5} (Latino)_{ijk} + \theta_{DEL6} (Black)_{ijk} + \theta_{DEL7} (Other)_{ijk} + \mu_{DEL_{jk}}
\]

\(\beta_{AGG}\) and \(\beta_{DEL}\) are the intercepts for NC \(k\) on the two latent traits, respectively, adjusted for the various background characteristics of the children and the primary caregivers in the models. (All covariates are centered at their sample grand means). \(\theta_{AGGq}\) and \(\theta_{DELq}\), for \(q = 1\) to \(7\), capture the relationship between the various personal characteristics and the latent traits of aggression and delinquency of child \(j\) within NC \(k\). The random effects \(\mu_{AGG_{jk}}\) and \(\mu_{DEL_{jk}}\) are assumed bivariate normally distributed with zero means and with residual person-level variances \(\tau_{AGG}\) and \(\tau_{DEL}\) and covariance \(\tau_{AGGDEL}\).

**Level 3 model.** For the intercepts, \(\beta_{AGG}\) and \(\beta_{DEL}\), \(\beta_{AGG} = \gamma_{AGG} + \nu_{AGG}\) and \(\beta_{DEL} = \gamma_{DEL} + \nu_{DEL}\). \(\gamma_{AGG}\) and \(\gamma_{DEL}\) are the grand mean levels of the two latent scores on aggression and delinquency in Chicago neighborhoods. The random effects \(\nu_{AGG}\) and \(\nu_{DEL}\) are assumed bivariate normally distributed

\(^9\) An empirical Bayes estimate of the adjusted log-odds of endorsement for each scale for each child or NC is an optimal composite of an estimate based on the data from that child or NC and an estimate based on data from another similar child or NC. Empirical Bayes residuals are computed as the deviations between the empirical Bayes estimates and the corresponding generalized least squares estimates of \(\beta_{AGG}\) and \(\beta_{DEL}\) in Equation 2: and \(\gamma_{AGG}\) and \(\gamma_{DEL}\) in Equation 3 for each child and NC respectively (see Bryk & Raudenbush, 1992, for details). The generalized least squares estimates of \(\gamma_{AGG}\) and \(\gamma_{DEL}\) are given in Table 4.

\(^{10}\) Our interest is in investigating and comparing the person-specific traits for each subscale across different groups of children and NCs, assuming measurement properties to be invariant across children and NCs. An alternative multilevel factor analysis might study whether these measurement properties vary across different groups and different NCs (see, e.g., Muthén, 1994).

\(^{11}\) We estimated another model that included interaction terms to test whether age-outcome trajectories were similar across levels of gender, ethnicity, and education and to examine whether gender effect, ethnicity effect, and education effects were mutually independent. The results showed no significant interaction effects, and thus the interaction terms were dropped for simplicity.
with zero means and with NC-level variances $\omega_{AGG}$ and $\omega_{DEL}$ and covariance $\omega_{AGGDEL}$.

For simplicity we postulated that the structural coefficients $\theta_{AGGqk}$ and $\theta_{DELqk}$, $q = 1, \ldots, 7$, are invariant across the NCs, that is, $\theta_{AGGqk} = \gamma_{AGGq}$ and $\theta_{DELqk} = \gamma_{DELq}$ for all $k$, though the model can be readily expanded to investigate variation in these coefficients across NCs.

**Combined model.** The combined model now becomes

$$
\eta_{ijk} = D_{AGGijk} \left( \gamma_{AGG} + \sum_{q=1}^{7} \gamma_{AGGq} (P)_{qjk} 
+ \sum_{m=1}^{19} \alpha_m X_{mijk} + u_{AGGjk} + v_{AGGk} \right) 
$$

$$
+ D_{DELijk} \left( \gamma_{DEL} + \sum_{q=1}^{7} \gamma_{DELq} (P)_{qjk} 
+ \sum_{m=1}^{12} \delta_m Z_{mijk} + u_{DELjk} + v_{DELk} \right),
$$

where $P$ denotes a person-level covariate in Equation 9 and there are a total of seven of them.

**Results for the conditional model with only person-level covariates.** Table 5 gives partial results for the model. The adjusted estimates of item severities $\alpha_m$ and $\delta_m$ are not reported here as our major focus was on the results of the simultaneous estimation of the effects of the covariates and risk factors on the two true latent scores. The outcomes of the fitted model listed in Table 5 are the two latent variables, $\pi_{AGGijk}$ and $\pi_{DELijk}$. These latent variables are, in the language of hierarchical models, the Level 1 intercepts (see Equations 1 and 9).

Controlling for gender, ethnicity, and educational level, there exist linear ($\gamma_{DEL1} = 0.106, SE = 0.015$) and quadratic ($\gamma_{DEL2} = 0.017, SE = 0.008$) effects of age for delinquency. However, no significant age-related effects exist for aggression. To gauge the age effects, we computed the probability of endorsement of delinquency as a function of age. At age 9, the probability of endorsement for a typical child is

$$
\frac{1}{1 + \exp[-(-3.277 + (-3 \times 0.106) + (9 \times 0.017))] = 0.03;
$$
at age 12, it is

$$
\frac{1}{1 + \exp[-(-3.277)]} = 0.04;
$$
and at age 15, it is

$$
\frac{1}{1 + \exp[-(-3.277 + (3 \times 0.106) + 9 \times 0.017))] = 0.06.
$$

The fact that age is related significantly to delinquency linearly and quadratically but not to aggression does not imply that the two corresponding coefficients for each subscale ($\gamma_{AGG1}$ and $\gamma_{AGG2}$ for aggression, and $\gamma_{DEL1}$ and $\gamma_{DEL2}$ for delinquency in Equation 10) are significantly different from each other. We therefore tested the null hypothesis that there was no difference in the age effects for aggression and delinquency. Specifically,

<table>
<thead>
<tr>
<th>Predictor of Level 1 intercept</th>
<th>Approx. df</th>
<th>Coeff.</th>
<th>SE</th>
<th>$t$</th>
<th>Aggression</th>
<th>Coeff.</th>
<th>SE</th>
<th>$t$</th>
<th>Delinquency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>78</td>
<td>-1.066</td>
<td>0.056</td>
<td>-19.05*</td>
<td>-3.277</td>
<td>0.069</td>
<td>-47.78*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>2,169</td>
<td>-0.024</td>
<td>0.013</td>
<td>-1.77</td>
<td>0.106</td>
<td>0.015</td>
<td>7.03*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age sq.</td>
<td>2,169</td>
<td>0.001</td>
<td>0.007</td>
<td>0.09</td>
<td>0.017</td>
<td>0.008</td>
<td>2.09*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>2,169</td>
<td>-0.097</td>
<td>0.064</td>
<td>-1.51</td>
<td>-0.287</td>
<td>0.073</td>
<td>-3.97*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ed. level</td>
<td>2,169</td>
<td>0.051</td>
<td>0.070</td>
<td>0.70</td>
<td>-0.059</td>
<td>0.078</td>
<td>-0.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latino</td>
<td>2,169</td>
<td>-0.307</td>
<td>0.101</td>
<td>-3.04*</td>
<td>-0.439</td>
<td>0.113</td>
<td>-3.91*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>2,169</td>
<td>0.183</td>
<td>0.102</td>
<td>1.80</td>
<td>0.301</td>
<td>0.111</td>
<td>2.71*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>2,169</td>
<td>-0.348</td>
<td>0.161</td>
<td>-2.17*</td>
<td>-0.468</td>
<td>0.184</td>
<td>-2.54*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-NC variance</td>
<td>1.842 (SE = 0.068)</td>
<td>1.691 (SE = 0.085)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between-NC variance</td>
<td>0.025 (SE = 0.016)</td>
<td>0.021 (SE = 0.019)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Approx. = approximate; Coeff. = coefficient; Age sq. = child-age quadratic term; Ed. level = education level; NC = neighborhood cluster

* $p < .05.$
\[ H_0: \begin{pmatrix} \gamma_{AGG1} - \gamma_{DELI} \\ \gamma_{AGG2} - \gamma_{DELI2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \]

The result indicates that one should reject the null hypothesis, \( \chi^2(2, N = 2,177) = 152.52, p < .01 \), and that there is a difference in the patterns of age-related changes for the two subscales. We then tested whether individual contrasts \( (\gamma_{AGG1} - \gamma_{DELI}) \) and \( (\gamma_{AGG2} - \gamma_{DELI2}) \) for the linear and quadratic age effects were equal to zero. The results indicate that there are differential linear, \( \chi^2(1, N = 2,177) = 127.94, p < .01 \), and quadratic, \( \chi^2(1, N = 2,177) = 7.03, p < .01 \), effects for the two subscales.

Furthermore, there is a statistically significant gender gap in predicting the log odds of endorsement of delinquency but not aggression. The estimated ratio of odds for a female versus male is \( \exp(-0.287) = 0.751 \), with the girls being less likely than boys to receive an endorsement on delinquency. We tested and rejected the null hypothesis that there was no difference in the gender effects for aggression and delinquency. \( \chi^2(1, N = 2,177) = 11.98, p < .01 \).

Holding other variables constant, the children whose primary caregivers belonged to the Latino and the Other category (including Pacific Islanders, Asian and Native Americans, and others) were significantly less likely to receive an affirmative rating on aggression and delinquency than were those with White primary caregivers. Black primary caregivers reported higher levels of delinquency \( (\gamma_{DELI6} = 0.301, SE = 0.111) \) but not aggression \( (\gamma_{AGG6} = 0.183, SE = 0.102) \) than did White primary caregivers. To test whether ethnicity effects differed across subscales \( (\gamma_{AGG5}, \gamma_{AGG6}, \gamma_{AGG7}, \gamma_{DELI5}, \gamma_{DELI6}, \gamma_{DELI7} \) for delinquency in Equation 10, we performed another multivariate hypothesis test with the null hypothesis

\[ H_0: \begin{pmatrix} \gamma_{AGG5} - \gamma_{DELI5} \\ \gamma_{AGG6} - \gamma_{DELI6} \\ \gamma_{AGG7} - \gamma_{DELI7} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \]

The result achieves statistical significance at the .01 level, \( \chi^2(3, N = 2,177) = 14.02, p < .01 \), and suggests there are differential PC ethnicity effects for the two subscales. However, results of follow-up tests of whether the three individual contrasts were nonsignificant, making the specific nature of these differences uncertain.

In sum, the results indicate that there is a linear and quadratic age trend as well as a gender effect for delinquency but not aggression. The two subscales relate differently to age and sex of the child and to PC ethnicity. This invalidates the assumption of unidimensionality that requires the two subscales relate similarly to key covariates.

Table 6
Partial Results for the Conditional Model With Only Person- and NC-Level Covariates

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Approx. df</th>
<th>Coeff.</th>
<th>SE</th>
<th>t</th>
<th>Coeff.</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor of Level 1 Intercept</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>77</td>
<td>-1.093</td>
<td>0.055</td>
<td>-19.72*</td>
<td>-3.319</td>
<td>0.069</td>
<td>-48.28*</td>
</tr>
<tr>
<td>Female</td>
<td>2,169</td>
<td>-0.095</td>
<td>0.064</td>
<td>-1.49</td>
<td>-0.285</td>
<td>0.072</td>
<td>-3.95*</td>
</tr>
<tr>
<td>Latino</td>
<td>2,169</td>
<td>0.359</td>
<td>0.103</td>
<td>-3.50*</td>
<td>-0.510</td>
<td>0.115</td>
<td>-4.45*</td>
</tr>
<tr>
<td>Black</td>
<td>2,169</td>
<td>0.058</td>
<td>0.113</td>
<td>0.51</td>
<td>0.134</td>
<td>0.124</td>
<td>1.08</td>
</tr>
<tr>
<td>Other</td>
<td>2,169</td>
<td>-0.421</td>
<td>0.162</td>
<td>-2.59*</td>
<td>-0.567</td>
<td>0.186</td>
<td>-3.64*</td>
</tr>
<tr>
<td>Ed. Level</td>
<td>2,169</td>
<td>0.080</td>
<td>0.071</td>
<td>1.13</td>
<td>-0.020</td>
<td>0.081</td>
<td>-0.26</td>
</tr>
<tr>
<td>Conc. dis.</td>
<td>77</td>
<td>0.276</td>
<td>0.079</td>
<td>3.47*</td>
<td>0.357</td>
<td>0.088</td>
<td>4.04*</td>
</tr>
<tr>
<td>Predictor of Age Linear Slope</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>2,169</td>
<td>-0.021</td>
<td>0.013</td>
<td>-1.59</td>
<td>0.109</td>
<td>0.015</td>
<td>7.17*</td>
</tr>
<tr>
<td>Conc. dis.</td>
<td>2,169</td>
<td>-0.021</td>
<td>0.019</td>
<td>-1.12</td>
<td>-0.023</td>
<td>0.021</td>
<td>-1.10</td>
</tr>
<tr>
<td>Predictor of Age Quadratic Slope</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>2,169</td>
<td>0.003</td>
<td>0.007</td>
<td>0.43</td>
<td>0.020</td>
<td>0.008</td>
<td>2.53*</td>
</tr>
<tr>
<td>Conc. dis.</td>
<td>2,169</td>
<td>-0.023</td>
<td>0.010</td>
<td>-2.39*</td>
<td>-0.030</td>
<td>0.011</td>
<td>-2.70*</td>
</tr>
<tr>
<td>Within-NC Variance</td>
<td></td>
<td>1.834 (SE = 0.068)</td>
<td>1.677 (SE = 0.085)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between-NC Variance</td>
<td></td>
<td>0.017 (SE = 0.015)</td>
<td>0.012 (SE = 0.017)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The estimates of the item difficulties are not shown here in the interest of simplifying the table. Approx. = approximate; Coeff. = coefficient; Ed. = education; Conc. dis. = concentrated disadvantage; NC = neighborhood cluster.

* p < .05.
Neighborhood-level covariates. After assessing the effects of person-level covariates, we entered a neighborhood characteristic into the model to study if and how the neighborhood characteristic relates to the subscales. The Level 1 and Level 2 models stay the same as given by Equations 1 and 9. We enter the neighborhood characteristic into the Level 3 model. Once again, there are two major sets of equations in the model specification. First, for the adjusted true scores, \( \beta_{AGGk} \) and \( \beta_{DELk} \), the following model is formulated: \( \beta_{AGGk} = \gamma_{AGG} + \gamma_{AGG01}(Conc.\ dis.)_{1k} + v_{AGGk} \) and \( \beta_{DELk} = \gamma_{DEL} + \gamma_{DEL01}(Conc.\ dis.)_{1k} + v_{DELk} \). We centered the level of concentrated disadvantage (Conc. dis.) around its mean, and thus \( \gamma_{AGG} \) and \( \gamma_{DEL} \) are the expected mean levels of the two latent scores on aggression and delinquency in Chicago neighborhoods, and \( \gamma_{AGG01} \) and \( \gamma_{DEL01} \) are structural coefficients relating level of concentrated disadvantage to aggression and delinquency. The residual random effects \( v_{AGGk} \) and \( v_{DELk} \) are assumed bivariate normally distributed with zero means and with residual NC-level variances \( \omega_{AGG} \) and \( \omega_{DEL} \) and covariance \( \omega_{AGGDEL} \).

As in the last model, we postulate the effects of the sex of the child and the ethnicity and the educational level to be the same across the NCs. That is, \( \theta_{AGGqk} = \gamma_{AGGq} \) and \( \theta_{DELqk} = \gamma_{DELq} \) for \( q = 3 \) to \( 7 \), for all \( k \). For the coefficients capturing the linear and quadratic age effects \( \theta_{AGGqk} \) and \( \theta_{DELqk} \) for \( q = 1, 2 \), we model them as a function of the level of concentrated disadvantage of a neighborhood: \( \theta_{AGGqk} = \gamma_{AGGq} + \gamma_{AGGq1}(Conc.\ dis.)_{1k} \) and \( \theta_{DELqk} = \gamma_{DELq} + \gamma_{DELq1}(Conc.\ dis.)_{1k} \), where \( q = 1, 2 \). \( \gamma_{AGGq} \) and \( \gamma_{DELq} \) are now the expected age effects for aggression and delinquency for neighborhoods with an average level of concentrated disadvantage. \( \gamma_{AGGq1} \) and \( \gamma_{DELq1} \) capture the interaction effects between age effects and level of concentrated disadvantage on the log-odds of endorsement of aggression and delinquency.\(^{12}\)

Results for the conditional model with person- and NC-level covariates. As indicated by Table 6, the partial PC ethnicity effect (Black vs. White) ceased to be statistically significant for delinquency after controlling for level of concentrated disadvantage. The results suggest that the Black–White gap in the log-odds of endorsement of delinquency at the individual level can be explained by the neighborhood characteristic of level of concentrated disadvantage. After controlling for the level of concentrated disadvantage, the gender effects for the two subscales remain statistically different, \( \chi^2(1, N = 2,177) = 11.97, p < .01 \). The three PC ethnicity effects are also statistically different for aggression and delinquency, \( \chi^2(3, N = 2,177) = 11.02, p = .01 \).

For both aggression and delinquency, the results indicate important associations with neighborhood concentrated disadvantage. For aggression, the neighborhood intercept is significantly elevated in neighborhoods characterized by high concentrated disadvantage (\( \gamma_{AGG10} = 0.276, SE = 0.079 \)). There is no association between concentrated disadvantage and the linear age slope (\( \gamma_{AGG11} = -0.021, SE = 0.013 \)), but there is a significant negative association between concentrated disadvantage and the quadratic age slope (\( \gamma_{AGG21} = -0.023, SE = 0.010 \)). For delinquency, the results are similar, but the estimated coefficients are a bit larger (\( \gamma_{DEL10} = 0.357, SE = 0.088 \); \( \gamma_{DEL11} = -0.023, SE = 0.021 \); and \( \gamma_{DEL21} = -0.030, SE = 0.011 \)). The results are difficult to interpret without visualization. Therefore, we plotted the probability of endorsement of the two subscales as a function of the two covariates (see Figures 3 and 4). The two plots show the predicted endorsement of aggression and

\[
\eta_{ijkt} = D_{AGGijkt}(\gamma_{AGG} + \sum_{q=1}^{7} \gamma_{AGGq}(P)_{ijkt} + \gamma_{AGG01}(Conc.\ dis.)_{kt} + \sum_{q=1}^{2} \gamma_{AGGq1}(P)_{ijkt}(Conc.\ dis.)_{kt} + \sum_{m=1}^{10} \alpha_{mk}X_{mkij} + \mu_{AGG} + v_{AGG} + D_{DELijkt}(\gamma_{DEL} + \gamma_{DEL01}(Conc.\ dis.)_{kt} + \gamma_{DEL11}(Conc.\ dis.)_{kt} + \gamma_{DEL12}(Conc.\ dis.)_{kt} + \sum_{m=1}^{12} \delta_{mk}Z_{mkij} + \mu_{DEL} + v_{DEL}),
\]

where \( P \) stands for a person-level variable, and the product of \( P \) with \( Conc.\ dis. \) stands for an interaction term. Equation 11 indicates that the log-odds of the probability of the endorsement of the behavior item \( i \) for person \( j \) of neighborhood \( k \) depends on person- and neighborhood-level characteristics, the cross-level interactions between characteristics at the person and neighborhood levels, and unique child and neighborhood effects.
delinquency for a typical child at ages 9, 12, and 15 in neighborhoods that are two standard deviations above and below the mean level of concentrated disadvantage. For aggression, Figure 3 shows that the effects associated with concentrated disadvantage are quite large at age 12 and much smaller at ages 9 and 15. For delinquency, Figure 4 indicates a similar effect of concentrated disadvantage, with most pronounced effects at age 12. The key difference between the two graphs is the differential age effect previously discussed: There is little evidence of an age trend for aggression but a pronounced positive age trend for delinquency.

Finally, we evaluated the hypothesis that neighborhood effects were more pronounced for delinquency than for aggression. We tested whether the effects of concentrated disadvantage on the intercept, linear slope, and quadratic age slope were different for aggression and delinquency. The null hypothesis was

$$H_0: \begin{pmatrix} \gamma_{AGG10} - \gamma_{DEL10} \\ \gamma_{AGG11} - \gamma_{DEL11} \\ \gamma_{AGG21} - \gamma_{DEL21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$ 

The result shows that there is no differential association for the two subscales, $\chi^2(3, N = 79) = 1.45$, $p > .50$. Thus, we retained the null hypothesis of no differential association and concluded that the effects of concentrated disadvantage do not seem to differ across subscales as we had hypothesized.

Discussion

In studying children’s problem behaviors, researchers typically combine item responses into a scale score for each child and study associations between the resulting scale scores and covariates measured at the level of the child and the social setting. This approach assumes that the scale score acting as the dependent variable is unidimensional with respect to some meaningful metric. If such a scale is unidimensional, potential subscales should not only be highly intercorrelated; they should also relate similarly to covariates. Analytic models are thus required to define a meaningful metric and to model associations with covariates. The model should be multivariate to allow estimation of latent correlations between subscales and the testing of differential associations of the subscales with covariates.

Our general strategy for these analytic tasks is to embed an item response model within a hierarchical model. The item response model allows researchers to construct meaningful interval scales for latent variables. The hierarchical structure of the model enables one to assess the associations between subscales and covariates measured at different ecological levels. This approach allows researchers to treat the subscales as multivariate outcomes at higher levels.

We have illustrated this approach by studying externalizing behaviors of a large sample of children in Chicago. We have tested the hypothesis that externalizing behaviors represent a single dimension against the specific alternative hypothesis that the externalizing items actually tap two distinct dimensions, aggression and delinquency (Achenbach, 1991a).

The results of our analysis suggest that the two subscales, Aggression and Delinquency, were highly intercorrelated at both the person and neighborhood levels, after controlling for measurement error. Although this finding supports a conclusion of unidimensionality, we found that the two subscales related substantially differently to key covariates. Thus, we concluded that it would be inappropriate to combine the Aggression and Delinquency subscales and to treat the composite as a single externalizing dimension in our analyses.

More specifically, our results indicate a significant
positive age trend for delinquency but not for aggression. The multivariate nature of the model enabled us to confirm that the magnitude of the age effect differed significantly across the two subscales. We also found significantly higher levels of Delinquency for boys than for girls, with no significant gender gap for Aggression. Again using a multivariate test, we found that the magnitude of the gender gap was significantly larger for Delinquency than for Aggression. Moreover, Delinquency and Aggression exhibited significantly different associations with ethnicity. Each of these differential effects undermines confidence in viewing externalizing behaviors as unidimensional. To combine Aggression and Delinquency items into such a single scale would lead to erroneous inferences regarding age-related effects, gender differences, and ethnic disparities.

However, contrary to our hypothesis, neighborhood concentrated disadvantage related similarly to aggression and delinquency. High levels of concentrated disadvantage were associated with elevated levels of delinquency and aggression at age 12, with smaller differences at ages 9 and 15.

These results depend strongly on the assumption that the two constituent scales—Aggression and Delinquency—are themselves unidimensional. There are two general strategies for testing this key assumption.

The first is theory driven. One might test the hypothesis that aggression is unidimensional against a theoretically specified alternative. For example, Crick and Grettener (1995) have proposed that aggression has two distinct components: relational and overt. Exactly the same strategy used in the current study could be used to test this hypothesis. Thus, the items might be classified into relational and overt subsets and dimensionality assessed by (a) estimating the correlation between latent relational and overt components at the person and neighborhood levels and (b) studying differential associations with covariates such as age, gender, and ethnicity.

The second strategy is more exploratory. A researcher might study differential item functioning within our two scales. The question here is whether item severities are different for older versus younger children, for boys versus girls, or for members of different ethnic groups. This kind of analysis would involve an expansion of our Level 2 model (Equation 2) to view item severities as outcomes predicted by person-level covariates. An alternative exploratory strategy would involve item factor analyses within levels of age, gender, ethnicity, or other covariates.

One would investigate the dependence of factor loadings on the values of the covariates. A large, normative sample would be best for such an approach in order to support stable, generalizable results.

The models used here are based on simple logistic models (Rasch, 1980) with dichotomous responses. They can be extended to models with polytomous responses. For an example of a general multidimensional Rasch-type item response model that synthesizes many different existing Rasch models, see Adams, Wilson, and Wang's (1997) multidimensional random coefficients multinomial logit model. Three-level hierarchical models for ordinal data are presented by Gibbons and Hedeker (1997).

The general strategy can be readily extended to longitudinal data for future waves of PDCN, which allows researchers to assess and compare the growth trajectories of aggression and delinquency for the children in the NC. This can be achieved by incorporating another level for occasions to the existing three-level model. The resulting model would be a four-level model with item responses within the time points, time points within persons, and persons within neighborhoods. Thus, this general strategy, together with the longitudinal research design of PDCN, will enable researchers to formulate a model to evaluate specific hypotheses about the combined influences of attributes of the developing child and characteristics of neighborhoods on trajectories of change across multiple domains, as posited by Bronfenbrenner (1979, 1996) and Bronfenbrenner and Ceci (1994).

References


