Published in final edited form as:

Educ Res Eval. 2007 February 1; 13(1): 71. doi:10.1080/13803610601058215.

Multiple Teacher Ratings: An evaluation of measurement strategies

Clara G. Muschkin^{a,*}, Patrick S. Malone^a, and Conduct Problems Prevention Research Group^{b,1}

^aDuke University, Durham, North Carolina, USA;

^bCPPRG, USA

Abstract

This study addresses the questions that arise when collecting, describing, and analyzing information from multiple informants regarding attributes of individual students. Using data from the Fast Track study, we evaluate alternative measurement strategies for using multiple teacher ratings of student adjustment to middle school among a sample of 326 Grade-6 pupils. One goal of the study was to compare the advantages of three measurement strategies using multiple and single informants in terms of their correlation with contemporaneous measures of behavior and academic achievement. Comparisons of residual variance using an aggregated rating, the rating from an “optimal informant,” and a score selected at random from the response set, indicate that aggregation provides the highest criterion-related validity. As part of these analyses, we explore the significance of inter-rater concordance, measured in terms of the intraclass correlation coefficient (ICC). Results indicate that for some aggregated scores, reliability can significantly limit their interpretability. The second main goal of the study was to evaluate the effects of variation in the number of teacher ratings on residual variance estimates for aggregate measures in selected behavioral domains. We conclude that the advantages of using multiple ratings are significant with a larger number of informants.

Introduction

Studies of childhood and adolescence often focus on behavior that takes place in school setting, with teachers as vital informants of student behavior and academic performance. There is little doubt regarding the value of the teacher report, particularly in primary schools, where children typically spend at least six hours per day with the same teacher (and, in some cases, an assistant teacher). Middle school students, however, almost universally move among several classrooms for core academic subjects. The presence of multiple teachers with equivalent value as informants raises questions as how to best collect and analyze information from teacher reports. Research that is designed to include teacher reports must commit to a measurement strategy that involves collecting data either from a single teacher or from multiple teachers. This decision has irrevocable consequences for measurement validity and reliability, and the types of analyses that are possible with the resulting data. Furthermore, if the decision is made to collect multiple teacher reports, the researcher must then select among alternative methods for handling these data, as well as for handling patterns of missing data and unequal numbers of

*Corresponding author. Center for Child and Family Policy, Duke University, Box 90545, Durham, NC 27708-0545, USA. muschkin@duke.edu.

¹Members of the CPRG are, in alphabetical order, Karen L. Bierman (Pennsylvania State University), John D. Coie (Duke University), Kenneth A. Dodge (Duke University), E. Michael Foster (University of North Carolina at Chapel Hill), Mark T. Greenberg (Pennsylvania State University), John E. Lochman (University of Alabama), Robert J. McMahon (University of Washington), and Ellen E. Pinderhughes (Tufts University).

reports across subjects. In this article, we review the data collection strategies that are used in research involving secondary school teachers and evaluate alternative methods for using the data from multiple respondents in terms of data reduction and statistical analysis. Although this article focuses on multiple ratings by teachers, it is useful to first examine the wider context of methodological issues in behavioral research involving multiple data sources.

Single Versus Multiple Respondents

Across disciplines and substantive topics, research design involves the methodological decision of selecting appropriate data sources for the conceptual constructs of interest (Holmbeck, Li, Schurman, Friedman, & Coakley, 2002). Using multiple data sources is, by definition, resource intensive; therefore, it is important to establish the benefit of this data collection strategy over the use of single sources. A single-source design may in fact be the most appropriate in some research situations: for example, in research seeking to maximize the power to detect significant effects, a single-source study with a large sample may be the most effective use of resources (Gray & Steinberg, 1999). However, this strategy has significant limitations. Data from a single source may provide an incomplete or simplistic view of the characteristics that are being measured. Further, if the response set is from a single source, it is not possible to rule out common method variance interpretations of the findings. Thus, the relations among variables that are identified in the analyses may reflect characteristics of the informants rather than associations between true score variance (Holmbeck et al., 2002).

When using multiple respondents, the extent of disagreement among informants may be assessed and evaluated as a source of variance. Methodological issues of estimating and interpreting inter-rater agreement and the reliability of scale scores have received considerable attention in behavioral research, in substantive areas such as learning behaviors (Buchanan, McDermott, & Schaefer, 1998), special education (Cicchetti & Sparrow, 1981), developmental psychology (Fan & Chen, 2000; Mitchell, 1979), academic performance (Abedi & Baker, 1995), academic advising (Sun & Valiga, 1997), and educational testing (Brennan, 1992). Both the level of discordance and the extent to which this will influence the interpretation of the measures' effects depend on the outcomes that are analyzed and the overall variability among the subjects. If between-subject variability is low, then differences in scores due to scorer disagreement may have an important effect on outcomes; if there is a high level of between-subject variability, inter-rater agreement is not as salient an issue (Mitchell, 1979).

When it is expected that there will be a low correlation among multiple informants, or where resources are not available to replicate data collection, the researcher may opt for the "optimal informant" strategy. This method is based on the premise that some raters are more reliable than others for certain child and adolescent behaviors (Holmbeck et al., 2002; Loeber, Green, Lahey, & Stouthamer-Loeber, 1989). Identifying the person most likely to provide accurate information for a particular construct then becomes a part of the research design and should be a consideration in designing the data collection process.

Approaches to Using Data From Multiple Respondents

A research design using multiple respondents must define the methods that will be used to handle these data. Holmbeck et al. (2002) provided a succinct schematic presentation of the advantages and drawbacks of these methods, portions of which are included in the following summary.

1. One option is for the researcher to employ a selective strategy, using a single or optimal informant identified from the set of respondents. This approach would make sense when there is low agreement among the sources and the researcher elects not to use the disaggregated information in the analyses. The main problem with this

approach is the assumption that the selected information is the most accurate and the loss of the unique information that each respondent could provide.

2. A second strategy is to use all or a subset of multiple informant data with data reduction methods. One such method involves aggregating data across informants using a simple linear model, which “captures multiple perspectives while improving predictive power” (Holmbeck et al., 2002, p. 10). This method also has the advantage of applying to any number of sources, which can be differentially weighted based on the perspective that the study elects to emphasize. This method is useful when the measures across informants have a relatively high degree of association, indicating a level of shared variance across the informants (Holmbeck et al., 2002). However, the collapsing of measures across respondents involves a loss of information in that the unique contribution of each informant cannot be discerned. Problems with this method arise when data are missing for one or more informants or when there is systematic variation in the number of informants available for different research subjects. In either case, the variances are unequal across subjects, which can result in biased standard errors in regression-based analyses (linear regression, analysis of variance, structural equation modeling [SEM], etc.). Weighting strategies can be used to obtain a weighted least squares or a generalized least squares solution for linear regression (Greene, 2000, ch. 12), but their application may not be widespread and they do not necessarily generalize easily to more complex analytic techniques, such as SEM.
3. Some research designs permit evaluation of multiple indicators of a construct within a single model, using a latent variable model approach. As Holmbeck et al. (2002) noted, this approach captures systematic variance that is common across informants (in the latent construct); and treats across-informant variance as error. This approach has the advantage of simultaneously considering measures from multiple informants in the estimation of structural models, thus retaining the maximum amount of information. Latent variable models have the disadvantage of (often) requiring a larger sample for the estimation methods (e.g., maximum likelihood) and may be conceptually inappropriate to research questions that do not involve the interrelationship of latent constructs with multiple measures of behavior that are associated with these constructs.

Methodological Approaches in Existing Research

Research with a focus on middle and high schools that uses teacher reports on student behavior must, by default, select from a potential set of multiple respondents in all but the smallest of schools. Our review of recent articles on adolescent development that include teacher reports of behavior and performance at school yielded a variety of approaches to this issue. Our literature review focused primarily on studies that specifically address issues of classroom behavior, adaptation, and academic achievement/performance in middle and high school. Several studies included only teacher ratings obtained in elementary school as predictors of behavior in middle and late adolescence (Roeser, Eccles, & Freedman-Doan, 1999; Shiner, 2000). A number of studies using a middle and/or high school student sample opted to obtain a teacher report from a single teacher. In one case, the method for selecting the teacher was specified: a subset of teachers chosen by the school principal as representative of the school population (Wentzel, 1994). Other articles reported ratings from a single teacher without specifying the selection criteria (Adams, Ryan, Ketsetzis, & Keating, 2000; Cole, Martin, Peeke, Seroczynski, & Fier, 1999; Loeber et al., 1989; Pakaslahti & Keltikangas-Jarvinen, 2000; Youngstrom, Loeber, & Stouthamer-Loeber, 2000).

Studies that used multiple reports from teachers have used the options identified earlier for summarizing and analyzing these data. Rudolph, Lambert, Clark, and Kurlakowsky (2001)

used teacher ratings of academic engagement in models predicting emotional and academic difficulties in the transition to middle school. Teacher ratings were obtained prior to and following the move to middle school, with ratings averaged across subject areas. At the elementary school level, the ratings were provided by a single classroom teacher; in the middle school analyses, the scores were averaged ratings from different core subject teachers. In their analyses, using structural equation models, the two scales are considered equivalent despite the confounding of variation across raters and variation across subjects. Alvidrez and Weinstein (1999) examined the relationship of teacher perceptions and academic achievement, with teacher ratings at ages 4 and 11. Although only one teacher rating was used for most students, part of the sample received more than one rating. For these students, the ratings from separate teachers were averaged.

A latent variable modeling approach was used in two studies of academic self-concept (Byrne & Gavin, 1996; Marsh & Yeung, 1997). Each of these studies obtained ratings of multiple academic domains from at least two teachers for each student. The information from each teacher for each domain was incorporated as an indicator to a latent construct; a further level of disaggregation was achieved by specifying separate models by core subject area. This approach thus avoided the issue of missing data and uneven numbers of teacher reports.

As noted in this brief literature review, research using teacher reports on student behavior most often uses a single teacher informant, with varying criteria for selecting the teacher. Studies that use ratings from more than one teacher often encounter difficulties in resolving issues of data reduction and analysis. In some studies, multiple teacher measures are constructed post hoc, resulting in averages based on information from different sources or that are available for only a part of the research sample. Research studies using multiple teacher reports in latent variable models do not share these methodological pitfalls; however, the technique of subsetting the sample by subject (or other criteria) may not be feasible with smaller samples or for other conceptual frameworks.

Method

The analyses presented in this study are part of a methodological inquiry rather than an estimation of theoretically driven models. Our focus is on evaluating the relative advantages of alternative measurement strategies for using multiple teacher ratings of student behavior in middle school. The three measurement strategies that we examine are: aggregated ratings, optimal informant rating, and random selection of a rating from a teacher response set. The first goal of our analyses was to compare the strategies in terms of correlation with contemporaneous measures for several behavioral domains. Our second goal was to evaluate, for measures that aggregate teacher ratings, the effects of having different numbers of teachers providing ratings across subjects.

Measures

For our evaluation of measurement strategies, we used data from Fast Track, an ongoing longitudinal investigation of conduct problems and their development from early childhood through adolescence and young adulthood (Conduct Problems Prevention Research Group [CPPRG], 1992, 2002). The Fast Track study has used a wide variety of measures in its annual data collection, including child, parent, and teacher interviews, as well as archival records. Teacher ratings have been of particular importance to this study, as kindergarten teachers provided the first of multiple sample screenings, using items from the Teacher Observation of Classroom Adaptation–Revised to rate children on levels of disruptive behavior (Werthamer-Larsson, Kellam, & Wheeler, 1991).

As respondents in the Fast Track study entered middle school, school-related behaviors were measured through administration of instruments to teachers in each of the student's core subjects. Therefore, some teacher measures administered in Grades 6, 7, and 8 contain multiple ratings of the target behaviors. Our focus is on the Teacher Rating of Student Adjustment (TRSA), administered for the first time to multiple teachers in Grade 6. This measure was developed by the Fast Track project to assess dimensions of success in adjusting to middle and high school. The measure targets teachers' perceptions of the student's academic performance, academic motivation, social skills, adult relationships, conduct, and personal maturity. The teacher ratings are coded on a 5-point scale (Muschkin., 2003). The decision by Fast Track to collect ratings from multiple teachers was based on practical and substantive considerations. Because the participants were located within a subset of schools within only four sites, the costs of obtaining multiple teacher surveys for each student were relatively low. Further, one of the goals for this stage of the study was to evaluate the benefits of using multiple respondents, and to identify the most useful methods for using these data. This investigation has as its aim to inform the use of this measure within the Fast Track project, as well as to generalize to other research regarding alternative methods for handling multiple ratings.

Participants

The analyses presented here are based on the data for 326 Grade-6 pupils, who were part of Fast Track's normative sample, in the seventh annual interview of this longitudinal study (see Table 1 for descriptive statistics). These participants were recruited at the same time as the high risk control and intervention samples, through a multistaged screening of kindergarteners at four sites across the USA. Children in the normative sample displayed the full range of problem behaviors in their schools; however, they did not receive any of the intervention services included in the Fast Track project. The full normative sample consists of 387 participants, 61 of whom were excluded from the present analysis as a result of missing data on the TRSA measure.

Analytical procedure

In our analyses, we propose three alternative strategies for representing the information available through the TRSA multiple ratings. Our primary strategy for comparing the measures is through the correlation of each of the sets of scores with contemporaneous measures selected from other Fast Track instruments. Criterion-related validation is possible for the measure of academic performance. The three sets of measures can be comparatively validated with regard to academic performance measures drawn from school records: grade point average and/or scores on end-of-grade-achievement tests. The use of school archival record information as a criterion for evaluating the measurement quality of a scoring rubric has been demonstrated in prior research (Abedi, 1991; Abedi & Baker, 1995). External construct validation of the sets of measures for the other behavioral domains is accomplished through correlation with items and scale scores from the Parent Rating of Student Adjustment. The latter instrument taps into the parent's perspective on the child's adjustment to school within the behavioral domains included in the TRSA.

As a first data reduction technique, we aggregated the multiple ratings by creating scale scores that average ratings across teachers for each of the behavior domains (academic performance, academic motivation, social skills, relationships with adults, conduct, and personal maturity). In general, summed scale scores exhibit high levels of reliability and intercorrelation; however, there are issues of concern in using them for analyses.

One of the issues involved in the aggregation of multiple ratings is the potential lack of concordance among the teacher ratings for each student. A low level of agreement across raters may indicate that there is a significant source of unexplained variance, limiting the

interpretability of the analyses. As a first step in our evaluation of measurement strategies, we evaluated inter-rater reliability using the intraclass correlation coefficient (ICC). The ICC is a useful estimate of reliability when the data include varying numbers of raters for each target (“class”; Mitchell, 1979). In this case, the “class” refers to an individual student in the sample; the ICC is the proportion of variance in the item responses accounted for by the student’s identity. The intraclass correlation values presented in Table 2 indicate acceptable reliability for four of the six behavioral domains contained in the TRSA; the domains *social skills* and *adult relationships* exhibited low reliability levels. We selected the four domains with acceptable reliability for the comparative analyses, in order to eliminate the potential confounding factor of low inter-rater agreement in comparisons of measurement strategies.

An additional issue stems from the condition that in the Fast Track data average scores are based on differing numbers of teacher ratings across students, with many respondents receiving only one teacher rating and some receiving as many as five. Therefore, the scale scores have different statistical properties that may limit their comparability. Specifically, the smaller the number of teachers reporting on a student, the more likely the mean rating is to diverge from the student’s “true score.” One consequence is that extreme ratings are more likely to be obtained from single-teacher reports than multiple-teacher averages. The problem of differing variances across averaged scores is ubiquitous in research using multiple raters of observed behavior when variation in the number of raters either reflects true conditions or is due to missing data. The problem of heteroscedasticity due to different numbers of raters across subjects has received little acknowledgement in existing research applications, although the issue is acknowledged in methodological tests (e.g., Greene, 2000, ch. 12).

The TRSA instrument provides a unique opportunity to evaluate the “optimal informant” measurement strategy. Along with the six items tapping behavioral domains, the instrument contains an item that asks each teacher to rate how well he or she knows the student. This 5-point rating of how well each teacher claims to know the child was then used to identify the optimal informant within the set of available teacher raters. In cases where two or more of the teachers rating a particular student provided the same (and highest) score on “how well do you know the child”, one of the ratings was selected randomly from the “tied” set.

Finally, we used the multiple-teacher ratings contained in the Fast Track data set to simulate a single-teacher rating strategy. This method involves a random selection of one teacher from the available set of teachers providing a rating for each child. This strategy is included to provide a baseline for comparing the relationships of the alternative TRSA scoring rubrics to criterion variables.

Results

Evaluating Alternative Measurement Strategies

Analyses involved a series of structural equation models designed to test differences in the squared multiple correlations (as an indicator of reliability) between the teacher ratings and the selected correlates as a function of the teacher rating strategy. The full model included the five items from the school records form and the seven items from the parent report of student adjustment as predictors, each predicting the 12 measures from the TRSA (four domains × three strategies). (The “prediction” is purely in a statistical sense, to establish concordance between the TRSA measures and the other measures of the constructs. No specific causal hypotheses are intended.) The TRSA measures were allowed to correlate freely, resulting in a saturated model.

This model was estimated for the sample of 189 students who had TRSA reports from at least two teachers (students with only one teacher reporting would necessarily have the same results

from all three strategies and their inclusion would have made it more difficult to detect differences among strategies). All variables were standardized to a zero mean and unit standard deviation for this sample. This made the residual variances for the TRSA measures comparable, defining them as the coefficients of nondetermination.

The model with residual variance estimates is presented in Figure 1. For all four domains the aggregates resulted in the smallest residual variance estimate, as expected. We first tested the overall effect of the strategies on the residual variances. Residual variances for each domain were constrained to be equal across the three strategies. This resulted in a significant decrement in fit, likelihood ratio (L.R.) $\chi^2(8, n = 189) = 37.68, p < .001$. Somewhat surprisingly, there appeared to be no consistent difference between the random-informant and optimal-informant strategies; this conclusion was supported by a test of the equality of variance constraints between those two strategies, L.R. $\chi^2(4, n = 189) = 4.29, p > .30$.

Variance With Different Numbers of Informants

The second stage of analysis focused on the properties of the aggregate strategy across students with different numbers of teachers providing ratings. Our primary intent was to explore the effects on the residual variance estimates, as an indicator of heteroscedasticity. The sample for this analysis was the 319 students with one or more teacher ratings: 130 with one rating, 70 with two ratings, and 119 with three or more (the sample with four or more ratings was too small to support the analysis approach if considered a separate group). All variables were again standardized in the analysis sample to facilitate interpretation of residual variance estimates. The full model was a multigroup SEM with the 12 predictors used above, as well as effect-coded variables for study site, predicting the aggregate TRSA measures in the four domains. Study site was included due to possible procedural differences that might have been related to the teacher recruitment process. In the full model, all paths from the predictors to the TRSA scores were constrained to be equal across groups, where groups were defined by the number of raters. Intercepts on the TRSA variables were also constrained to be equal across groups. Thus, the regression equations predicting the TRSA variables were equivalent, but the residual variances were free to vary.

This model fitted the data adequately, but not especially well: $\chi^2(128, n = 319) = 209.13, p < .001$, CFI = .95, TLI = .92, est. RMSEA = .077, 90% CI: .058, .096. The model with residual variance estimates appears in Figure 2. The lack of fit may have been due in part to undocumented systematic differences in the recruitment of teachers, a point to which we will return below. For the purposes of exploration, we took this model as true and fitted a constrained model with the residual variances constrained to be equal across groups, which resulted in a significant impairment of fit, L.R. $\chi^2(8) = 39.14, p < .001$. A specific test of the equality between the two-rater and three-rater groups also showed a decrease in fit, L.R. $\chi^2(4) = 12.41, p < .02$, but the equality constraints between the one-rater and two-rater groups did not, L.R. $\chi^2(4) = 6.56, p > .15$.

Discussion

Summary of the Findings

The Fast Track study provides a unique opportunity for evaluating divergent strategies used to represent and analyze multiple ratings of behavior. Capitalizing on the detailed information obtained from multiple teachers for each study subject, as well as the availability of contemporaneous related measures, we have evaluated the relative merits of measurement strategies used to represent teacher ratings of student adjustment to middle school. Although this research focuses on teacher ratings in a secondary school setting, the findings can be

generalized to other areas of behavioral research in terms of alternative methods for reduction and analysis of data from multiple sources.

Most researchers investigating social behavior do not have the luxury of selecting between single and multiple data sources at the analysis stage of a project; these parameters are usually set during the design stage, prior to data collection. Collection of data from multiple sources is significantly more resource intensive than obtaining information from a single source. One of the goals of this research was to determine, using the example of this Fast Track measure, the benefits that may be derived from the use of multiple ratings, as compared with two strategies using single informants. Our analyses use structural equation models to compare three strategies for representing teacher ratings, in terms of their correlation with contemporaneous Fast Track behavioral measures. Our comparison of residual variance estimates using an aggregated summary score, an optimal informant score, and a score selected at random from the ratings set indicates, not surprisingly, that aggregation produced the score most closely associated with the criterion information and with correlated constructs. In our comparisons of the two single-informant strategies, we found no consistent advantages to the optimal-informant strategy, as compared with the randomly selected informant, however, our sample size for that analysis was not exceptionally large.

Despite the advantages noted in terms of measurement validity, the use of information from multiple informants is not necessarily the optimal strategy. Use of multiple ratings raises issues that were evident in our analyses. One such issue involves the degree of concordance among raters. In this study, we used the intraclass correlation coefficient as an indicator of reliability levels for the behavioral domains rated by teachers in the TRSA measure. Our analyses illustrated the importance of evaluating inter-rater reliability when using aggregated scores for multiple ratings. Of the six domains included in the TRSA, only four yielded intraclass correlations at a reliability level that permits meaningful interpretation of aggregated scores. The other two domains were not included in our subsequent comparative analyses; we made the decision to exclude them in order to avoid confounding our comparisons with an additional source of unexplained variance in the measures. In research that focuses on behavioral domains that typically elicit heterogeneous response sets among multiple raters, the analyst may opt to select a single respondent's rating. In such cases, the decision to use an optimal informant or a randomly selected one is likely to be an a priori research design feature, based on the resources available for data collection and the conceptual justification for a particular measurement strategy.

Our second set of analyses, focusing on the properties of the aggregate measurement strategy, evaluate the effects of cross-subject heterogeneity in the number of teachers providing ratings on residual variance estimates for the aggregate measures in the four behavioral domains. SEM models were used to evaluate residual variance estimates across groups defined by the number of teachers providing ratings for each student. Once again, our results were somewhat surprising, in that the residual variances for the two-rater group were not significantly decreased relative to the group of subjects receiving a single rating. There was a significant improvement, however, in the three-rater group over the two-rater group.

Limitations of the Study and Conclusions

In summary, our results provide some evidence that there are some advantages to the use of multiple informants in behavioral ratings and that these advantages are enhanced with a larger number of informants. The strength of the evidence is tempered; however, by the extent to which the findings may reflect particular characteristics of the data used for our analyses. One caveat mentioned earlier is the possible source of systematic measurement error through differences in teacher recruitment processes at the four Fast Track project sites. Variation in the number of raters also may result from other non-random patterns of missing data (in the

ideal case we would have had a large sample of students with at least three raters, which could have been randomly assigned to conditions of using fewer reporters). When using data with considerable variation in the number of informants providing behavioral ratings, the researcher may choose to avoid problems of heteroscedasticity by selecting a single respondent according to the strategy that is most appropriate for the conceptual constructs that are being analyzed. The “optimal-informant” strategy did not yield any advantages in our study, perhaps due to the criterion that we used for identifying this respondent. The teacher’s self-report of how well he or she knows the student may actually reflect how well the student behaves in that class. Thus, the criterion variable may be endogenous and not appropriate for selecting from among the available teacher ratings.

When the use of multiple ratings is the strategy of choice, the most valid and reliable measures are obtained by maximizing the number of raters per subject (our results suggest at least three, at least for this content area) and by ensuring a “critical mass” of subjects with this number of ratings. For research on teacher ratings of student behaviour, the benefits of obtaining multiple teacher ratings are likely to outweigh the relatively low costs of administering the measure to several teachers. Clearly, across educational settings, as well as other contexts, these methodological decisions will reflect the focus as well as the constraints of the research design.

Acknowledgments

Fast Track has been funded by the National Institute of Mental Health (NIMH) through grants R18 MH48043, R18 MH50951, R18 MH50952, R18 MH50953, and R01 MH62988. The Center for Substance Abuse Prevention and the National Institute on Drug Abuse have also provided support through a memorandum of agreement with the NIMH. Department of Education grant S184U30002 and NIMH grants K05MH00797 and K05MH01027 also supported the study. We are grateful for the ongoing dedication of the Fast Track participants, investigators, and research staff.

References

- Abedi J. Predicting graduate academic success from undergraduate academic performance: A canonical correlation study. *Educational and Psychological Measurement* 1991;51(1):151–160.
- Abedi J, Baker EL. A latent-variable modeling approach to assessing interrater reliability, topic generalizability, and validity of a content assessment scoring rubric. *Educational and Psychological Measurement* 1995;55(5):701–705.
- Adams GR, Ryan BA, Ketsetzis M, Keating L. Rule compliance and peer sociability: A study of family process, school-focused interactions, and children’s classroom behavior. *Journal of Family Psychology* 2000;14(2):237–250. [PubMed: 10870292]
- Alvidrez J, Weinstein RS. Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology* 1999;91(4):731 – 746.
- Brennan RL. Generalizability theory. *Educational Measurement: Issues and Practices* 1992;11(4):27–34.
- Buchanan HH, McDermott PA, Schaefer BA. Agreement among classroom observers of children’s stylistic learning behaviors. *Psychology in the Schools* 1998;35(4):355–361.
- Byrne BM, Gavin DAW. The Shavelson model revisited: Testing for the structure of academic self-concept across pre-, early-, and late adolescents. *Journal of Educational Psychology* 1996;88(2):215–228.
- Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *Journal of Mental Deficiency* 1981;86(2):127–137.
- Cole DA, Martin JM, Peeke LA, Seroczynski AD, Fier J. Children’s over- and underestimation of academic competence: A longitudinal study of gender differences, depression, and anxiety. *Child Development* 1999;70(2):459–473. [PubMed: 10218266]
- Conduct Problems Prevention Research Group (CPPRG). A developmental and clinical model for the prevention of conduct disorders: The FAST Track Program. *Development and Psychopathology* 1992;4:509–527.

- Conduct Problems Prevention Research Group (CPPRG). Using the Fast Track randomized prevention trial to test the early-starter model of the development of serious conduct problems. *Development and Psychopathology* 2002;14:925–943. [PubMed: 12549710]
- Fan X, Chen M. Published studies of interrater reliability often overestimate reliability: Computing the correct coefficient. *Educational and Psychological Measurement* 2000;60(4):532–542.
- Gray MR, Steinberg L. Unpacking authoritative parenting: Reassessing a multidimensional construct. *Journal of Marriage and the Family* 1999;61:574–587.
- Greene, WH. *Econometric analysis*. 4. Upper Saddle River, NJ: Prentice Hall; 2000.
- Holmbeck GN, Li ST, Schurman JV, Friedman D, Coakley RM. Collecting and managing multisource and multimethod data in studies of pediatric populations. *Journal of Pediatric Psychology* 2002;27(1):5–18. [PubMed: 11726675]
- Loeber R, Green SM, Lahey BB, Stouthamer-Loeber M. Optimal informants on childhood disruptive behaviors. *Development and Psychopathology* 1989;1(4):317–337.
- Marsh HW, Yeung AS. Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. *Journal of Educational Psychology* 1997;89(1):41–54.
- Mitchell SK. Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin* 1979;86(2):376–390.
- Muschkin, CG. Teacher rating of student adjustment (Fast Track Project technical report). 2003. Retrieved February 15, 2005, from <http://www.fasttrackproject.org/>
- Pakaslahti L, Keltikangas-Jarvinen L. Comparison of peer, teacher and self-assessments on adolescent direct and indirect aggression. *Educational Psychology* 2000;20(2):177–191.
- Roeser RW, Eccles JS, Freedman-Doan C. Academic functioning and mental health in adolescence: Patterns, progressions, and routes from childhood. *Journal of Adolescent Research* 1999;14(2):135–174.
- Rudolph KD, Lambert SF, Clark AG, Kurlakowsky KD. Negotiating the transition to middle school: The role of self-regulatory processes. *Child Development* 2001;72(3):929–946. [PubMed: 11405592]
- Shiner RL. Linking childhood personality with adaptation: Evidence for continuity and change across time into late adolescence. *Journal of Personality and Social Psychology* 2000;78(2):310–325. [PubMed: 10707337]
- Sun A, Valiga MJ. Using generalizability theory to assess the reliability of student ratings of academic advising. *Journal of Experimental Education* 1997;65(4):367–380.
- Wentzel KR. Relations of social goal pursuit to social acceptance, classroom behavior, and perceived social support. *Journal of Educational Psychology* 1994;86(2):173–182.
- Werthamer-Larsson L, Kellam S, Wheeler L. Effects of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology* 1991;19:582–602.
- Youngstrom E, Loeber R, Stouthamer-Loeber M. Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology* 2000;68(6):1038–1050. [PubMed: 11142538]

Figure 1.
Standardized residual variance estimates and standard errors with different strategies ($n = 189$)

School Performance Records / Parent Rating of Adjustment

Figure 2.
Standardized residual variance estimates and standard errors for aggregation strategy with differing numbers of raters ($n = 319$)

Table 1

Correlation coefficients, means, and standard deviations

Variable	n	M	SD	1	2	3	4	5	6	7	8	9	10	11	12
Teacher ratings, three measurement schemes															
1. Academic performance, average	319	2.80	1.10												
2. Academic motivation, average	319	2.92	1.18	.76											
3. Conduct, average	319	3.16	1.24	.58	.71										
4. Maturity, average	319	2.89	1.22	.68	.82	.75									
5. Academic performance, knows child best	319	2.86	1.18	.94	.72	.55	.65								
6. Academic Motivation, knows child best	319	2.93	1.29	.71	.94	.67	.76	.70							
7. Conduct, knows child best	319	3.16	1.34	.58	.69	.95	.74	.57	.70						
8. Maturity, knows child best	319	2.90	1.35	.64	.76	.72	.94	.65	.77	.76					
9. Academic performance, random selection	319	2.81	1.17	.95	.73	.57	.66	.89	.68	.57	.62				
10. Academic motivation, random selection	319	2.88	1.26	.73	.95	.68	.78	.68	.88	.66	.72	.73			
11. Conduct, random selection	319	3.14	1.31	.55	.67	.95	.72	.52	.63	.90	.69	.67	.68		
12. Maturity, random selection	319	2.88	1.33	.65	.78	.70	.93	.61	.72	.67	.87	.55	.81	.71	
Outcomes—Academic performance, Grade 7															
13. Reading grade	273	8.27	3.40	.68	.67	.56	.60	.64	.64	.57	.57	.63	.63	.53	.57
14. Mathematics grade	281	7.35	3.76	.64	.60	.46	.57	.60	.57	.46	.55	.61	.59	.44	.55
15. Language arts grade	282	8.21	3.48	.65	.65	.54	.63	.62	.61	.56	.59	.62	.62	.51	.58
16. Social studies grade	279	7.83	3.55	.65	.67	.53	.65	.60	.62	.53	.62	.61	.64	.50	.61
17. Science grade	270	8.25	3.71	.67	.63	.52	.61	.62	.59	.53	.57	.64	.61	.51	.58
Outcomes—Parent report															
18. Got along with friends	302	3.86	.97	.17	.28	.35	.29	.16	.26	.30	.25	.14	.28	.34	.29
19. Stayed out of trouble	302	3.57	1.26	.21	.40	.32	.40	.21	.38	.51	.39	.21	.38	.49	.36
20. Peers tried to negatively influence child	302	2.56	1.16	.11	.20	.27	.22	.10	-.20	-.26	-.21	-.10	-.90	-.24	-.21
21. Got into trouble	302	2.15	1.17	.25	.41	.54	.43	-.24	-.40	-.52	-.40	-.24	-.39	-.48	-.40
22. Did not do well in academics	302	2.78	1.35	.35	.41	.36	.43	-.33	-.37	-.36	-.38	-.32	-.38	-.33	-.41
23. Did not have many friends	302	2.08	.93	.17	.22	.30	.28	-.18	-.20	-.27	-.25	-.15	-.20	-.26	-.23
24. Teachers unduly strict with child	302	2.15	1.17	.37	.48	.60	.51	-.37	-.48	-.58	-.50	-.34	-.46	-.56	-.47

At $n = 270$ correlations of magnitude of .12 are significant, $p < .05$. Correlations among criterion variables are not shown.

Table 2

Intraclass correlation coefficients

Behavioral domain	Intraclass correlation coefficient
Academic performance	.61
Academic motivation	.62
Social skills	.31
Relationships with adults	.36
Conduct	.65
Personal maturity	.55

n = 532 ratings of 189 students.