Teacher Observation of Classroom Adaptation-Checklist: Measuring Children's Social, Emotional, and Behavioral Functioning

Bradshaw, C. P., & Kush, J. M. (2020). Teacher Observation of Classroom Adaptation-Checklist: Measuring Children's Social, Emotional, and Behavioral Functioning. *Children & Schools*, *42*(1), 29–40. https://doi.org/10.1093/cs/cdz022

Published in *Children & Schools*

Acknowledgement: This work was funded by grants from the U.S. Department of Education's Institute of Education Sciences (R305H150027; R324A07118). The authors would like to thank Philip Leaf and Katrina Debnam, Sheppard Pratt Health System, and the Maryland State Department of Education for their support of this project.

Abstract

There is a growing need for valid measures that can be administered efficiently in school settings to assess the impact of school-based preventive interventions. The current paper aimed to establish a balance among assessment efficiency, reliability, and the measurement properties of an instrument widely used to assess the impact of school-based programs, called the Teacher Observation of Classroom Adaptation-Checklist (TOCA-C; Bradshaw, Debnam, & Leaf, 2010; Koth, Bradshaw, & Leaf, 2009; Werthamer-Larsson, Kellam, & Wheeler, 1991). We leveraged item response theory (IRT) analyses to create a shortened, more focused checklist version of the TOCA-C which is both valid and efficient for large-scale use in schools to track students' behavioral, social-emotional, and family factors over the course of elementary school. The sample included 17,456 children in Kindergarten through grade 5 (47.7% female, 54.2% African American). IRT analyses resulted in the retention of 33 of the original 39 items comprising 7 subscales: 1) *Concentration Problems*, 2) *Aggressive/Disruptive Behavior*, 3) *Prosocial Behavior*, 4) *Emotion Regulation Problems*, 5) *Internalizing Problems*, 6) *Family Problems*, and 7) *Family Involvement*. IRT, item difficulty estimates, and confirmatory factor analyses revealed limited evidence of bias based on gender, race, or grade; together, the findings suggested that the 33-item TOCA-C is both a highly valid and reliable measure.

Teacher Observation of Classroom Adaptation-Checklist: Measuring Children's Social, Emotional, and Behavioral Functioning

Improving the social-emotional functioning of students and increasing positive family factors are common goals for public schools. Yet social workers and researchers alike continue to struggle with efficient but valid methods for assessing the ways in which social-emotional and behavioral aspects of children's functioning change and develop over time. In fact, in school-based research, teachers are often asked to rate entire classrooms of students simultaneously on several developmental dimensions over multiple time points, resulting in a significant burden on these key informants. As such, there is a growing need for valid measures that can be efficiently completed by teachers in regard to specific students. The current paper aimed to establish a balance among assessment efficiency, reliability, and the measurement properties of an instrument widely used to assess the impact of school-based prevention programs, called the Teacher Observation of Classroom Adaptation (TOCA; Werthamer-Larsson, Kellam, & Wheeler, 1991). We leveraged item response theory analyses to create a more focused version of TOCA, called the TOCA-C (checklist; Bradshaw, Debnam, & Leaf, 2010; Koth, Bradshaw, & Leaf, 2009), which is both valid and efficient for large-scale and targeted use in schools to track student social-emotional, behavioral, and family outcomes over the course of elementary school.

Background on the TOCA

The original Teacher Observation of Classroom Adaptation (TOCA) was created in the 1970s by the Woodland Research Center in Chicago, Illinois to serve as a measure of children's behavior (Koth et al., 2009). Originally administered to teachers through interviews by trained evaluators, the TOCA was conceptualized as a measure of students' social adaptation to the classroom and school settings, and used to assess the impact of school-based preventive interventions on students’ behavior. Subsequently, the original interview version was revised (referred to as the TOCA-Revised) to be a written, more efficient, and self-administered teacher self-report measure. Multiple studies have used various versions of the TOCA to evaluate prevention programs, such as the Fast Track Program (Conduct Problems Prevention Research Group, 2002), the Good Behavior Game (Petras, Chilcoat, Leaf, Ialongo, & Kellam, 2004; Schaeffer et al., 2006; Werthamer-Larsson et al., 1991), Positive Behavioral Interventions and Supports (PBIS; Bradshaw, Waasdorp, & Leaf, 2015), and the Incredible Years Program (Reinke, Herman, & Dong, 2018). Several studies also examined the convergent and predictive validity of the TOCA-R (e.g., Petras et al., 2004) and various aspects of its psychometric properties (Dong, Reinke, Herman, Bradshaw, & Murray, 2016; Koth et al., 2009). Other studies have examined specific subscales of the TOCA-R (i.e., 10-item aggressive-disruptive behaviors scale) in relation to measurement invariance, differential item functioning, and predictive validity (see Racz, King, Wu, Witkiewitz, McMahon, & CPPRG, 2013; Wu, King, Witkiewitz, Racz, & McMahon, 2012); in fact this specific subscale (originally referred to as “authority acceptance”) has received the most attention with regard to its psychometrics and predictive validity (also see Petras et al., 2004). More recently, the TOCA was augmented to include additional items to assess a broader set of social-emotional skills and family factors in order to increase its utility from both a research and clinical perspective (Bradshaw et al., 2010).

Overview of the Current Study

Although there has been considerable research on earlier versions of the TOCA, and a preliminary study examined some of the psychometric properties this checklist version of the TOCA (Crowder, 2014), additional research is needed on the TOCA-Checklist version to more systematically document its psychometric properties. Specifically, we leveraged item response theory (IRT; Lord, 1953) analyses to create an efficient yet valid version of TOCA which covers a broader range of social, emotional, and behavioral outcomes and family factors. The overarching goal of this study was to provide evidence that the TOCA-C is both highly valid and efficient for large-scale use in schools to assess elementary school students' behavior. Such a measure has utility both for research and clinical practice, in relation to student outcomes commonly targeted through universal programs, as well as group and intensive interventions.

Method

Sample

The data come from 45 elementary schools in six public school systems within a mid-Atlantic state. Using teacher rating data from *J* = 907 teachers, a total of *N* = 17,456 students in Kindergarten through grade 5 were assessed (47.7% female, 54.2% African American, 7.3% Latinx). The majority of the teachers were White (73.5%), female (89.0%) and had 5 or more years of educational experience (47.5%) (see teacher and student demographics in Table 1).

Measure

As described above, the Teacher Observation of Classroom Adaptation-Checklist (TOCA-C; Bradshaw et al., 2010) is self-administered, written, checklist based on the TOCA-R (Koth et al., 2009; Werthamer-Larsson et al., 1991). The measure administered in this study consisted of 39 items on a six-point Likert scale ranging from 1 = "*Never*" to 6 = "*Almost Always*" (see items by scale in Table 2). Specifically, the instructions on the measure read "*In the last three weeks*, would you say the following statements were never, rarely, sometimes, often, very often, or almost always true of this child . . .". Each item is scored, such that a higher score indicates more of that construct; therefore, some items were reverse coded (see Table 2). Items should be rescaled then averaged, thereby resulting in a score from 1 to 6 for each subscale. Prior exploratory and confirmatory factor analyses were used to establish the following seven subscales (Crowder, 2014): 1) *Concentration Problems*, which assessed inattentive and off-task behavior (Koth et al., 2009); 2) *Internalizing Problems*, which assessed the extent to which the child feels nervous, fearful, sad, withdrawn, and worries (Achenbach, 1991); 3) *Aggressive/Disruptive Behaviors*, which assessed disobedient, disruptive, and aggressive behaviors (Koth et al., 2009); 4) *Prosocial Behavior,* which assessed positive social interactions (Koth et al., 2009); 5) *Emotional Regulation Problems*, which assessed the child's impulsivity, frustration, and how the child deals with anger and being upset (Achenbach, 1991); 6) *Family Problems*, which assessed caregivers' degree of stability in home life and academic support of their children (Malone, 2000); and 7) *Family Involvement*, which assessed the caregivers' involvement in their child's school and parent's comfort in their relationship with the teacher (Malone, 2000). A copy of the measure is available for free upon request by contacting Catherine Bradshaw at cbradsha@jhsph.edu.

Procedure

Homeroom teachers provided TOCA-C ratings for all students in Kindergarten through grade 5 in their classroom as a part of the baseline data collection for an evaluation of a school-based prevention program (see Bradshaw et al., 2015). An open cohort design was employed, whereby all students and staff who entered the participating schools were eligible for inclusion. Data were collected in fall 2007 or 2008 on 90.4% of eligible children. The Institutional Review Board approved the study, which included a waiver of active parental consent.

Results

Item Response Theory (IRT) Analyses

The IRT analyses resulted in the retention of 33 of the original 39 items. In an effort to identify any potential items exhibiting bias, differential item functioning (DIF; Holland & Wainer, 1993) analyses were conducted. DIF analyses control for group differences on the measured latent trait. Theoretically, after controlling for latent trait ability, an item without bias *should* perform the same for two individuals, regardless of their group membership (i.e., measurement invariance). Specifically, DIF was examined across three dichotomous areas: gender (with male as the reference group), race (White [reference] vs. non-White), and grade (lower [K-2nd] vs. upper [3rd-5th as reference]).

To test the null hypothesis that an item score is independent of group membership, conditional on the estimated latent trait score, the Mantel-Haenszel procedure was used to produce a chi-square statistic and an associated *p*-value for each item (Mantel & Haenszel, 1959; Potenza & Dorans, 1995). However, test statistics from hypothesis tests are known to be influenced by sample size, such that tests are overly sensitive when sample sizes are large, as is the case in this study (Uttaro & Millsap, 1994; Zwick, 2012). As a result, the Educational Testing Services (ETS; Zwick, 2012) classification levels were also reported. For polytomous items, ETS classification levels are: 1) AA – little to no DIF, 2) BB – moderate amounts of DIF, and 3) CC – large amounts of DIF. For BB and CC items, a '+' sign indicates the item favors the focal group, while a '-' sign indicates the item favors the reference group (for additional information, see Zwick, 2012). All DIF analyses were conducted using jMetrik 4.1.1 software (Meyer, 2014).

DIF results are reported in Table 4. While many items have chi-square statistics and associated *p*-values less than .05, suggesting the item *does* function differently between the two groups, these estimates are known to be influenced by sample size (Uttaro & Millsap, 1994). As a result, more emphasis is placed on the ETS DIF classification level scores. Importantly, all items on the *Concentration*, *Aggressive/Disruptive Behavior*, *Emotion Regulation Problems, Family Problems*, and *Family Involvement* subscales were shown to have little to no DIF for gender, race, and grade subgroups. In total, only three items were shown to have moderate amounts of DIF for gender; specifically, "*Shows empathy and compassion for others feelings*" favored females more than males, whereas "*Has many friends*" (both from the *Prosocial Behavior* subscale) and "*Worries*" (from the *Internalizing* subscale) favored males more than females. Overall, results from the DIF analyses provide strong evidence of measurement invariance, helping to establish the desired psychometric properties of the TOCA-C.

Andrich's (1978) rating scale model was fit to the data for each item within a specific subscale using jMetrik 4.1.1 software (Meyer, 2014). For item *j* with *h* = 0, …, *m* response categories, the probability that individual *i* will select category *u* can be written as:

$$P(U_{ij} = u|\theta_i) = \frac{\exp \sum_{v=0}^{u_{ij}} [\theta_i - (b_j + \tau_v)]}{\sum_{h=0}^{m_j} \exp \sum_{v=0}^{h} [\theta_i - (b_j + \tau_v)]}$$

in which θ_i represents the latent trait ability of individual *i*, b_j represents overall item difficulty, and τ_v represents the category threshold parameter. The threshold parameter was fixed for all items sharing the same rating scale, representing a special case of Masters' (1982) partial credit model, and all items comprising each subscale were simultaneously estimated.

With a goal of accurately assessing latent trait scores across the spectrum of the latent trait, it was important to examine estimates of item difficulty. Simply stated, item difficulty represents the average location of an item along the latent trait continuum, or the amount of the latent trait necessary for an individual to endorse an item. For example, an item with a positive difficulty estimate would indicate that *more* of the latent trait was necessary to endorse the item, whereas an item with a negative difficulty estimate would indicate *less* of the latent trait was needed to endorse the item. Although item difficulty theoretically ranges from -∞ to +∞, in practice, estimates typically fall between ±6. Andrich's (1978) rating scale model produced item difficulty estimates reflecting an assessment of multiple locations along the latent trait continuum for all TOCA-C subscales. Item difficulty estimates for all items are shown in Table 2. For example, within the *Aggressive/Disruptive Behavior* subscale, item 3 ("*Harms others*") has a difficulty estimate equal to 1.499, whereas item 4 ("*Gets angry when provoked by other children*") has a difficulty estimate equal to -1.168. This represents a wide range of assessment along the latent trait continuum, ensuring an accurate assessment of all students.

Item maps. To aid in the interpretation of scores, as well as refine the items used for each subscale, item mapping procedures were produced (Huynh, 1998). Item mapping places both the distribution of individual latent trait estimates and item response categories on the same latent continuum. Specifically, maximum information item category mapping was used to locate the item category at the place in which it contributes the most information toward estimating the latent trait (Huynh, 1998). Specifically, $I_{jk}\ (\theta)$ is the item information for item j with response category k regarding the latent trait (θ). To maximize the category information function, we specified $I_{jk}\ (\theta) = P_{jk}\ (\theta) * I_{j}\ (\theta)$, in which $P_{jk}\ (\theta)$ is the rating scale model for response category k for item j, and $I_{j}\ (\theta)$ is the item information function. Item maps for each subscale are shown in Figure 1. For example, visual inspection of the item maps for the *Disruptive Behavior* items suggests that vast majority of individuals have negative latent trait estimates, and that very few individuals have *large* estimated trait values of disruptive behavior problems (positive latent trait estimates). Placing items onto the same latent continuum, it is easy to see the variability in difficulty estimates across the items. As in the example above, the map illustrates how item 3 ("*Harms others*") has the greatest item difficulty estimate, suggesting that a greater amount of the latent trait would be required for an endorsement of this item. In contrast, item 4 ("*Gets angry when provoked by other children*") has the smallest item difficulty estimate. Together, the DIF and item mapping procedures illustrate the range of item assessment along the latent continuum for each subscale, as well as the distribution of student trait estimates.

Confirmatory Factor Analysis

Confirmatory factor analyses models were fit to the data for each subscale. Model fit indices for each model are shown in Table 2 (means and standard deviations by subscale are reported in Table 3). Model fit indices including the Tucker-Lewis index (TLI), comparative fit index (CFI), and root mean square error of approximation (RMSEA) were calculated for each model. Values greater than .95 are considered desirable for both the TLI and CFI statistics, whereas an RMSEA less than or equal to .06 is recommended (Browne & Cudeck, 1993; Hu & Bentler, 1999). All models were estimated in Mplus using the maximum likelihood with robust standard errors estimator (Muthén & Muthén, 2017). Standardized beta coefficients from confirmatory factor analyses were used to understand the strength of the relationships between items and the underlying latent factor being measured. By scaling to units of a standard deviation change of *Y* to a standard deviation change of *X*, standardized beta (β) coefficients were obtained using model-fitted variances (Bollen, 1989; Muthén & Muthén, 2017). Because each item was loaded onto a single underlying factor, the standardized beta coefficients can be interpreted as correlation coefficients. The TLI and CFI statistics for each model were above the recommended .95 cutoff; however, the RMSEA estimates were slightly larger than the recommended .06 cutoff. Standardized beta coefficients are shown in Table 2. The *Emotion Regulation Problems* subscale had the smallest average mean standardized beta coefficient, $\overline{\beta} = .812$, whereas the *Family Involvement* subscale had the largest average mean standardized beta coefficient, $\overline{\beta} =$

.903. Overall, the estimated standardized beta values reflected strong associations between the items and the underlying latent factors.

Reliability Analyses

For the final seven subscales, we computed both Cronbach's alpha (α) and omega estimates as indicators of reliability. Alpha reliability estimates may underestimate the true reliability of the constant item variances of the true scores assumption is violated. Omega estimates can be used to correct this underestimation. While both indicators are presented, more emphasis is placed on alpha as these are more conservative estimates of reliability (Dunn, Baguley, & Brunsden, 2014). In general, the subscale alphas were high, ranging from .819 to .931 (see Table 2). Variance component (fully unconditional) multilevel models were fit to the data to explore the extent to which variation in each subscale was at the classroom and school level. In total, three separate variance component multilevel models were fit to the data for each subscale: 1) a 2-Level model in which students are nested within classrooms, 2) a 2-Level model in which students are nested within schools (ignoring clustering within classrooms), and 3) a 3-Level model in which students are nested within classrooms, and classrooms are nested within schools. Intraclass correlation coefficients (ICC) were calculated for each model using restricted maximum likelihood estimation in Stata (StataCorp, 2017a, 2017b) (see Table 2 for ICCs); these estimates may be helpful in conducting power analyses for future studies.

Discussion

In an effort to improve both the usability and psychometric properties of the TOCA-C scale, IRT analyses were used to develop a shortened, more targeted version of the measure. The IRT analyses resulted in the retention of 33 of the original 39 items. Item difficulty estimates reflected an assessment of multiple locations along the latent trait continuum for all TOCA-C subscales, allowing for a more efficient version which covered the full range of items across the latent trait. Visual inspection of item maps revealed similar findings, in that estimated locations of individual items along the latent trait continuum were shown against estimates of person ability scores (see Figure 1). Standardized beta coefficients from confirmatory factor analyses indicated strong, positive relationships for nearly all items (see Table 2). Mean standardized beta estimates for each subscale ranged from .489 for the *Emotion Regulation Problems* subscale to .781 for the *Family Involvement* subscale. Lastly, in an effort to identify any potential items exhibiting bias, DIF analyses were conducted; they revealed no measurement differences based on gender, race, or grade for 96 of the 99 parameters estimated, once ETS sample size adjustments were considered.

Although a prior study of the TOCA-R used IRT analyses (see Wu et al., 2012), that study focused on a single 10-item subscale (which they referred to as authority acceptance, and we refer to here as *Aggressive/Disruptive Behavior*) and only included kindergarteners (*N*=8,820) from the Fast Track study. The results of that study provided some evidence of DIF by gender, whereas we did not. Specifically, Wu et al. (2012) found differences on the overt behaviors within this specific scale, favoring males, whereas the nonphysical behaviors favored females; however, they found no consistent evidence of DIF by race/urban status for this scale. As such, the current study provides some convergent evidence of the validity of this particular subscale, but also contributes new information on six other subscales not previously examined using IRT. A related study on a subsample of higher risk students from the same Fast Track study found that the IRT-scaled version of this subscale of the TOCA was a better predictor of subsequent mental health outcomes through high school than a simple summed score (Racz et al., 2013). Additional research is needed to further replicate the current IRT findings with other samples; unfortunately, we currently lack data on all of these subscales from another sample in order to replicate the findings. However, a unique aspect of IRT analyses with regard to model-data fit is that item parameter estimates do not depend on the sample used for analyses, while person ability estimates are invariant across different samples of items (de Ayala, 2009). As our analyses indicated good model-data fit, researchers and practitioners should feel confident in using the reduced scale presented in this study. This finding, together with the large sample used to conduct the current analyses, leads us to conclude that the findings are stable enough to formulate conclusions supporting the use of the reduced 33-item scale based on these data.

Implications for Practice

The results suggested that the 33-item TOCA-C is efficient, valid, and reliable for use in elementary school settings, and thus is a potentially useful tool for a range of purposes. For example, there is some interest in using the TOCA-C as a screener to identify students in need of services. Although analyses examining the predictive validity and sensitivity/specificity of the current version of the TOCA-C are beyond the scope of the current study, as are efforts to identify specific cut points or thresholds of concern, the current findings, together with prior work on specific subscales of the TOCA (i.e., aggressive-disruptive behavior) do suggest some promise of this measure as a screener (see Petras et al., 2004; Racz et al., 2013). The current findings may also inform social workers' and other clinicians' use of the TOCA-C to identify individual students in need of services as well as evaluate or track progress over multiple administrations of the TOCA-C. Moreover, various versions of the TOCA have been frequently used to monitor the impact of programs and services longitudinally (typically fall to spring within a year, and across multiple years) (e.g., Petras et al., 2004; Schaeffer et al., 2004) highlighting its potential as a progress monitoring tool. The TOCA-C can be used to assess individual students, or to sample students from a classroom and average up to the classroom level. As noted above, the TOCA-C has typically been used to assess the impact of or need for behavioral and social-emotional preventive interventions, mental health programs, or other tiered interventions. Although additional research is needed to examine the current version of the TOCA-C with regard to predictive validity, the current findings suggest the utility of the TOCA-C for a range of uses (e.g., screener, progress monitoring, research) by social workers and other clinicians.

References

Achenbach, T. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile.* Burlington, VT: University of Vermont Department of Psychiatry.

Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Bradshaw, C.P., Debnam, K.J., & Leaf, P.J. (2010). *Teacher Observation of Classroom Adaptation-Checklist (TOCA-C, Version 2)*. Unpublished measure. Johns Hopkins University. Available for free upon request by contacting Catherine Bradshaw at cbradsha@jhsph.edu.

Bradshaw, C.P., Waasdorp, T.E., & Leaf, P.J. (2015). Examining variation in the impact of School-wide Positive Behavioral Interventions and Supports: Findings from a randomized controlled effectiveness trial. *Journal of Educational Psychology, 107*(2), 546-557.

Browne, M.W. & Cudeck, R. (1993) Alternative ways of assessing model fit. In: K.A. Bollen, & J.S., Long (Eds.) *Testing Structural Equation Models*. Newbury Park, Sage, 136-162.

Conduct Problems Prevention Research Group. (2002). Evaluation of the first 3 years of the Fast Track prevention trial with children at high risk for adolescent conduct problems. *Journal of Abnormal Child Psychology, 30*, 19–36.

Crowder, S. C. (2014). *An analysis of the psychometric properties of the PBISPlus Teacher Observation of Classroom Adaption-Checklist* (Unpublished doctoral dissertation). Morgan State University, MD, U.S.

de Ayala, R. J. (2009). *Theory and practice of item response theory*. New York: Routledge.

Dong, N., Reinke, W.M., Herman, K.C., Bradshaw, C.P., & Murray, D.W. (2016). Meaningful effect sizes, intraclass correlations, and proportions of variance explained by covariates for planning two- and three-level cluster randomized trials of social and behavioral outcomes. *Evaluation Review, 40*(4), 334-377. DOI: 10.1177/0193841X16671283

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105,* 399-412. doi:10.1111/bjop.12046

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*, *23*, 35–56.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2009). Teacher observation of classroom adaptation—checklist: Development and factor structure. *Measurement and Evaluation in Counseling and Development, 42*(1), 15-30.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*(4), 517-548. doi: 10.1177/001316445301300401

Malone, P.S. (2000) Parent and teacher involvement measure-teacher: Year-6 update (Technical Report) (online). Retrieved from http://www.fasttrackproject.org/techrept/p/ptt/

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Meyer, J. P. (2014). *Applied measurement with jMetrik*. New York: Routledge.

Muthén, L. K., Muthén, B. O. (2017). *Mplus User's Guide. Eighth Edition.* Los Angeles, CA: Muthén & Muthén.

Petras, H., Chilcoat, H. D., Leaf, P. J., Ialongo, N. S., & Kellam, S. G. (2004). Utility of TOCA-R scores during the elementary school years in identifying later violence among adolescent males. *Journal of the American Academy of Child and Adolescent Psychiatry, 43*(1), 88–96.

Potenza, M.T., & Dorans, N.J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Racz, S., King, K., Wu, J., Witkiewitz, K., J McMahon, R., & Conduct Problems Prevention Research Group (2013). The predictive utility of a brief kindergarten screening measure of child behavior problems. *Journal of Consulting and Clinical Psychology, 81*(4), 588-599. doi:10.1037/a0032366

Reinke, W. M., Herman, K. C., & Dong, N. (2018). The Incredible Years Teacher Classroom Management Program: Outcomes from a group randomized trial. *Prevention Science, 19*(8), 1043-1054. doi: https://doi.org/10.1007/s11121-018-0932-3

Schaeffer, C. M., Petras, H., Ialongo, N., Masyn, K. E., Hubbard, S., Poduska, J., & Kellam, S. (2006). A comparison of girls' and boys' aggressive-disruptive behavior trajectories across elementary school: Prediction to young adult antisocial outcomes. *Journal of Consulting and Clinical Psychology*, *74,* 500–510.

StataCorp (2017a). *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC

StataCorp (2017b). *Stata 15 Base Reference Manual*. College Station, TX: Stata Press.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel produce in the detection of differential item functioning. *Applied Psychological Measurement, 18*(1), 15-25.

Wu, J., King, K. M., Witkiewitz, K., Racz, S. J., & McMahon, R. J. (2012). Item analysis and differential item functioning of a brief conduct problem screen. *Psychological Assessment, 24*(2), 444-454.

Werthamer-Larsson, L., Kellam, S. G., & Wheeler, L. (1991). Effect of first-grade classroom environment on child shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology, 19*, 585–602.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Report-12-08). Educational Testing Services. www.ets.org/Media/Research/pdf/RR-12-08.pdf

Table 1 *Teacher and Student Demographic Characteristics*

Teacher Characteristics (*j* = 907)		Frequency	Percentage
Teacher Gender			
	Females	807	89.0
	Males	40	4.4
Teacher Ethnicity			
	White	667	73.5
	Black/African American	160	16.5
	Asian/Pacific Islander	16	1.9
	American Indian/Alaskan Native	6	0.7
	Hispanic/Latinx	1	0.1
	Other	4	0.4
Teacher Years of Educational Experience			
	Less than 2 years	109	12.0
	2 - 4.9 years	210	23.2
	5 or more years	431	47.5
Student Characteristics (*n*= 17,456)		Frequency	Percentage
Student Gender			
	Females	8,297	47.5
	Males	9,082	52.0
Student Ethnicity			
	Black/African American	9,209	52.8
	White	5,882	33.7
	Hispanic/Latinx	1,235	7.1
	Asian/Pacific Islander	550	3.2
	American Indian/Alaskan Native	88	0.5
	Other	33	0.2
Grade			
	Kindergarten (K)	2,817	16.1
	1st	2,716	15.6
	2nd	2,757	15.8
	3rd	2,991	17.1
	4th	2,936	16.8
	5th	2,949	16.9

Note. Does not total to 100% due to missingness.

Table 2

Confirmatory Factor Analyses and Item Difficulty Parameter Estimates for the 33-item TOCA-C

TOCA-C Subscale	$\beta_{\text{(standardized)}}$	Standard Error	Item Difficulty (b)	Standard Error
Concentration Problems				
Concentrates [r]	.939*	0.002	-0.066	0.013
Stays on task [r]	.941*	0.002	-0.206	0.013
Is easily distracted	.760*	0.005	-0.735	0.012
Completes assignments [r]	.865*	0.003	0.754	0.013
Learns up to ability [r]	.863*	0.004	0.254	0.013
Subscale Fit: CFI = 0.974, TLI = 0.949, RMSEA (95% CI) = 0.190 (0.185, 0.196)				
Aggressive/Disruptive Behavior				
Breaks rules	.791*	0.005	-0.946	0.012
Doesn't get along with others	.795*	0.005	-0.507	0.012
Harms others	.820*	0.006	1.499	0.017
Gets angry when provoked by other children	.801*	0.005	-1.168	0.011
Fights	.857*	0.005	1.044	0.016
Teases classmates	8.22*	0.005	0.079	0.014
Subscale Fit: CFI = 0.988, TLI = 0.980, RMSEA (95% CI) = 0.065 (0.061, 0.069)				
Prosocial Behavior				
Is friendly	.836*	0.004	-0.380	0.013
Shows empathy and compassion for others feelings	.816*	0.004	0.901	0.012
Is rejected by classmates [r]	.752*	0.006	-1.038	0.014
Has many friends	.830*	0.004	0.518	0.012
Subscale Fit: CFI = 0.979, TLI = 0.936, RMSEA (95% CI) = 0.160 (0.151, 0.169)				
Emotion Regulation Problems				
Stops and calms down when angry or upset [r]	.691*	0.007	-0.592	0.010
Changes moods quickly	.830*	0.005	0.393	0.012
Impulsive	.724*	0.007	0.130	0.011
Easily frustrated	.889*	0.003	0.020	0.011
Easily upset	.926*	0.004	0.049	0.011
Subscale Fit: CFI = 0.975, TLI = 0.950, RMSEA (95% CI) = 0.117 (0.111, 0.122)				
Internalizing Problems				
Nervous	.804*	0.006	-0.261	0.014
Withdrawn	.787*	0.006	-0.009	0.014
Fearful	.880*	0.004	0.671	0.016
Sad	.828*	0.005	0.043	0.015
Worries	.837*	0.005	-0.445	0.013
Subscale Fit: CFI = 0.962, TLI = 0.923, RMSEA (95% CI) = 0.127 (0.122, 0.133)				

(Table 2 continued)

TOCA-C Subscale	β(standardized)	Standard Error	Item Difficulty (b)	Standard Error
Family Problems				
Has a stable family life ^{r}	.906*	0.005	-0.486	0.011
Family problems negatively affect child's behavior in school	.768*	0.009	0.436	0.013
Family sends child to school ready to learn ^{r}	.827*	0.006	0.050	0.012
Subscale Fit: CFI = >0.999, TLI > 0.999, RMSEA (95% CI) <0.001 (0.000, 0.001)				
Family Involvement				
This child's guardian/ parent(s) attend parent-teacher conference	.887*	0.005	0.240	0.012
I have a good relationship with the child's parent	.935*	0.003	-0.353	0.012
I am able to contact the parent of this child if I need to talk about his/her progress or problems	.942*	0.003	-0.732	0.012
Parent is involved in and supportive of child's education	.943*	0.003	-0.480	0.012
Parent attends school functions such as open houses, book fair, and PTA meetings	.810*	0.008	1.325	0.013
Subscale Fit: CFI = 0.991, TLI = 0.981, RMSEA (95% CI) = 0.084 (0.079, 0.090)				

Subscale	Cronbach's Alpha	Omega	Number of Items	ICC for Students within Classrooms (2-Level)	ICC for Students within Schools (2-Level)	Students within Classrooms within Schools (3-Level)	
						Classroom Level	School Level
Concentration Problems	.925	.926	5	0.051	0.022	0.054	0.015
Aggressive/Disruptive Behavior	.867	.872	6	0.108	0.059	0.130	0.041
Prosocial Behavior	.836	.848	4	0.140	0.046	0.152	0.032
Emotion Regulation Problems	.870	.870	5	0.115	0.042	0.139	0.028
Internalizing Problems	.863	.864	5	0.167	0.053	0.204	0.027
Family Problems	.819	.827	3	0.190	0.032	0.162	0.022
Family Involvement	.931	.931	5	0.183	0.041	0.159	0.021

Note. * *p* < .001. Items marked with ^{r} indicate a reverse coding, such that a higher value is indicative of more of the latent trait. Also note that 6 of the original items from the full set of 39 items were dropped because of lower loadings and/or overlap with other subscale items; as such, the loading reflected in this table are the final 33-item model, without the 6 dropped items. Specifically, "Pays attention" and "Works hard" were dropped from the Concentration Problems subscale; "Yells at others", "Lies", and "Harms property" were dropped from the Aggressive/Disruptive Behavior subscale; and "Is liked by classmates" was dropped from the Prosocial Behavior subscale. CFI= Comparative Fit Index; TLI= Tucker-Lewis Index; RMSEA= Root Mean Square Error of Approximation; 95% CI = 95% Confidence Interval around the RMSEA. *Note*. Concentration Problems, Aggressive/Disruptive Behavior, Emotion Regulation Problems, Internalizing Problems, and Family Problems were coded (1 to 6), such that higher values were indicative of a less desirable trait.

Table 3. *Means and Standard Deviations in Item Responses by Grade and Gender*

Final TOCA-C Subscale	Grade	Male		Female	
		Mean	SD	Mean	SD
Concentration Problems	K	4.08	0.80	4.34	0.76
	1st	3.95	0.80	4.26	0.74
	2nd	3.94	0.78	4.33	0.74
	3rd	3.92	0.82	4.56	0.76
	4th	3.86	0.78	4.26	0.73
	5th	3.91	0.81	4.25	0.79
Aggressive/Disruptive Behavior	K	2.11	0.84	1.76	0.68
	1st	2.01	0.85	1.70	0.71
	2nd	2.04	0.83	1.68	0.69
	3rd	2.07	0.90	1.69	0.71
	4th	2.16	0.87	1.76	0.71
	5th	2.10	0.85	1.76	0.73
Prosocial Behavior	K	4.17	0.51	4.31	0.48
	1st	4.20	0.53	4.36	0.49
	2nd	4.19	0.55	4.38	0.50
	3rd	4.16	0.55	4.37	0.49
	4th	4.08	0.57	4.33	0.51
	5th	4.08	0.57	4.31	0.51
Emotion Regulation Problems	K	2.64	0.71	2.44	0.61
	1st	2.66	0.74	2.42	0.60
	2nd	2.70	0.76	2.42	0.59
	3rd	2.68	0.76	2.45	0.59
	4th	2.73	0.76	2.48	0.62
	5th	2.68	0.70	2.49	0.61
Internalizing Problems	K	1.83	0.79	1.80	0.77
	1st	1.84	0.80	1.73	0.75
	2nd	1.83	0.80	1.74	0.80
	3rd	1.83	0.79	1.74	0.77
	4th	1.90	0.77	1.80	0.74
	5th	1.95	0.81	1.84	0.77
Family Problems	K	2.92	0.50	2.86	0.49
	1st	2.94	0.50	2.83	0.48
	2nd	2.94	0.56	2.82	0.48
	3rd	2.92	0.51	2.84	0.46
	4th	2.95	0.53	2.87	0.50
	5th	2.96	0.47	2.85	0.48
Family Involvement	K	4.77	1.19	4.85	1.14
	1st	4.51	1.30	4.64	1.29
	2nd	4.59	1.32	4.66	1.33
	3rd	4.46	1.35	4.53	1.34
	4th	4.36	1.35	4.55	1.31
	5th	4.40	1.32	4.47	1.32

Note. All individual items were coded on a scale from 1 to 6. Items within a subscale were then averaged to create a mean subscale score, again ranging from 1 to 6, such that higher values were more indicative of the trait.

Table 4. *Differential Item Functioning by Subgroup*

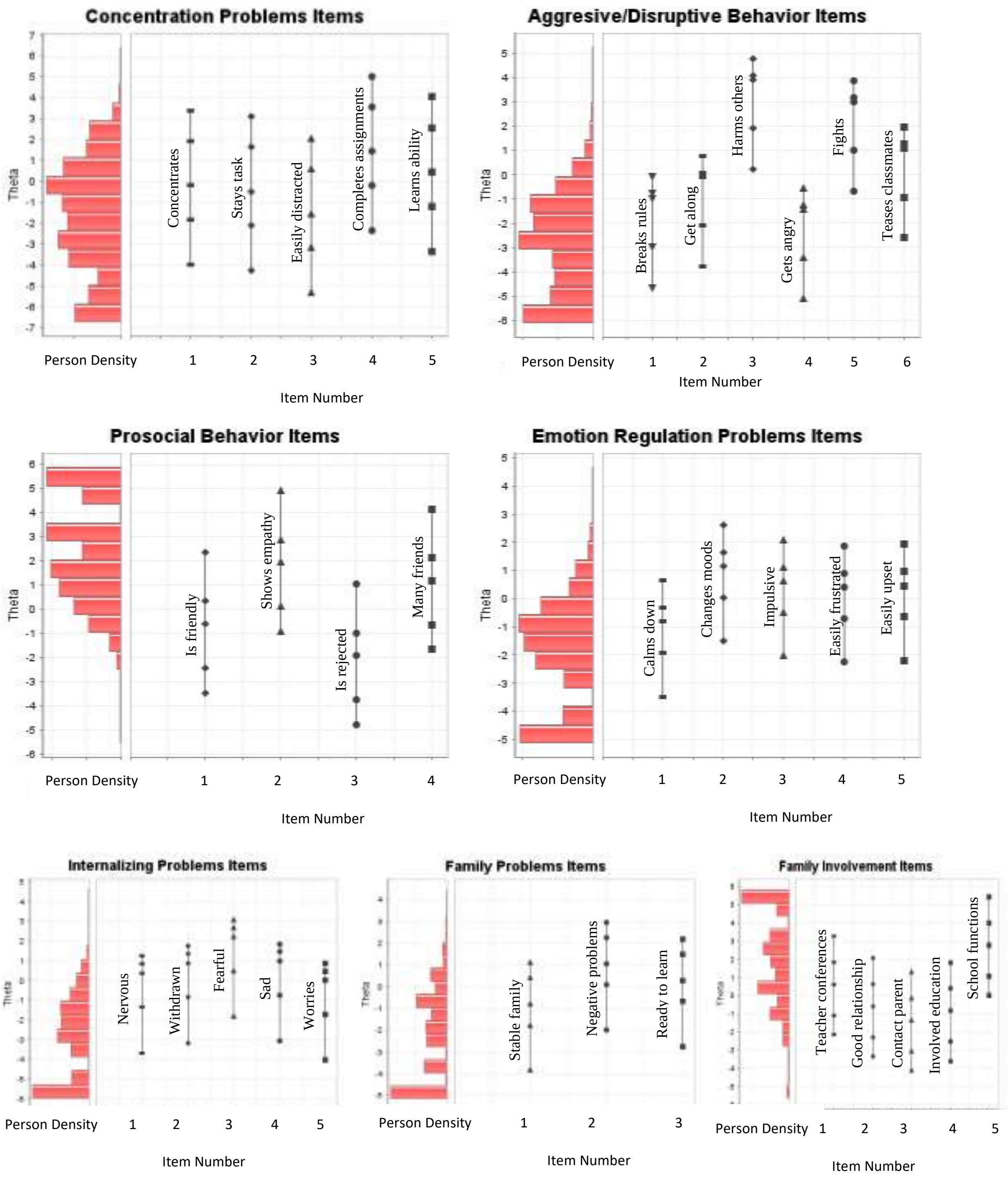
Item	**Gender**			**Race**			**Grade**		
Concentration Problems	χ^2	*p*-value	Class	χ^2	*p*-value	Class	χ^2	*p*-value	Class
Concentrates	6.37	.01	AA	0.09	.77	AA	10.64	<.01	AA
Stays on task	17.44	<.01	AA	0.02	.89	AA	0.02	.88	AA
Is easily distracted	16.67	<.01	AA	87.23	<.01	AA	17.16	<.01	AA
Completes assignments	2.38	.012	AA	57.05	<.01	AA	0.11	.74	AA
Learns up to ability	78.29	<.01	AA	19.24	. <.01	AA	52.98	<.01	AA
Aggressive/Disruptive Behavior									
Breaks rules	203.12	<.01	AA	0.72	.39	AA	55.21	<.01	AA
Doesn't get along with others	133.37	<.01	AA	0.81	.37	AA	31.77	<.01	AA
Harms others	0.83	.36	AA	0.11	.74	AA	58.15	<.01	AA
Gets angry when provoked by other children	5.27	.02	AA	9.55	<.01	AA	47.77	<.01	AA
Fights	1.12	.29	AA	8.14	<.01	AA	15.53	<.01	AA
Teases Classmates	30.61	<.01	AA	32.79	<.01	AA	12.53	<.01	AA
Prosocial Behavior									
Is friendly	186.39	<.01	AA	105.56	<.01	AA	57.27	<.01	AA
Shows empathy and compassion for others feelings	560.86	<.01	BB+	64.00	<.01	AA	39.53	<.01	AA
Is rejected by classmates	305.94	<.01	AA	109.03	<.01	AA	54.22	<.01	AA
Has many friends	412.62	<.01	BB-	64.70	<.01	AA	43.72	<.01	AA
Emotion Regulation Problems									
Stops and calms down when angry or upset	5.48	.02	AA	94.70	<.01	AA	<.01	.96	AA
Changes moods quickly	109.38	<.01	AA	1.27	.26	AA	22.19	<.01	AA
Impulsive	152.12	<.01	AA	38.54	<.01	AA	58.53	<.01	AA
Easily frustrated	0.06	.81	AA	11.03	<.01	AA	47.75	<.01	AA
Easily upset	93.00	<.01	AA	9.29	<.01	AA	1.31	.25	AA

(Appendix A continued)

Item	**Gender**			**Race**			**Grade**		
Internalizing Problems	χ^2	*p*-value	Class	χ^2	*p*-value	Class	χ^2	*p*-value	Class
Nervous	128.09	<.01	AA	169.57	<.01	AA	0.59	.44	AA
Withdrawn	12.96	<.01	AA	3.44	.06	AA	12.81	<.01	AA
Fearful	129.05	<.01	AA	5.61	.02	AA	19.09	<.01	AA
Sad	112.23	<.01	AA	0.09	.77	AA	1.98	.16	AA
Worries	477.11	<.01	BB-	85.48	<.01	AA	1.65	.20	AA
Family Problems									
Has a stable family life	47.86	<.01	AA	69.06	<.01	AA	25.17	<.01	AA
Family problems negatively affect child's behavior in school	86.14	<.01	AA	17.36	<.01	AA	2.97	.09	AA
Family send child to school ready to learn	131.40	<.01	AA	87.75	<.01	AA	26.47	<.01	AA
Family Involvement									
This child's parent(s) attend parent-teacher conferences	14.31	<.01	AA	46.79	<.01	AA	2.36	.12	AA
I have a good relationship with the child's parent	4.28	.04	AA	42.97	<.01	AA	20.17	<.01	AA
I am able to contact the parent of this child if I need to talk about his/her progress or problems	0.41	.52	AA	0.68	.41	AA	0.11	.75	AA
Parent is involved in and supportive of child's education	5.40	.02	AA	3.18	.07	AA	0.67	.41	AA
Parent attends school functions such as open houses, book fairs, and PTA meetings	15.63	<.01	AA	137.57	<.01	AA	14.94	<.01	AA

Note. AA, BB, and CC class values refer to the ETS classification system described in the Results.

Figure 1. *Item Maps by Subscale*



Note. An abbreviated item label is provided on the figure.