# What Affects Teacher Ratings of Student Behaviors?
# The Potential Influence of Teachers' Perceptions of the School Environment and Experiences

**Elise T. Pas · Catherine P. Bradshaw**

**Abstract** Teachers serve as the natural raters of students within the school and classroom contexts. Yet teachers' ratings of their students may vary based on these contextual factors. The current study explored the extent to which teacher perceptions of the school environment predict their longitudinal ratings of student behaviors. Data for this study come from 702 teachers in 42 elementary schools. Teachers self-reported their perceptions of the school context at a single time point, and provided ratings of their students' behavior via the Teacher Observation of Classroom Adaption-Checklist (TOCA-C) across three school years. Latent profile analysis identified three latent classes of teachers based on their ratings of school organizational health, burnout, and efficacy. A regression framework demonstrated an association between the baseline profiles in relation to TOCA-C ratings of student behavior across 3 years. Teachers with more favorable perceptions of the environment had lower initial ratings of concentration problems, disruptive behavior, and internalizing symptoms, and higher ratings of prosocial behaviors and family involvement. They also showed slower growth in their ratings of emotion dysregulation and greater increases of their ratings of family involvement over time. This work is particularly important for determining the extent to which teacher ratings may be biased by teacher and contextual factors, and may have implications for the identification of teachers who may rate students poorly over time.

**Keywords** TOCA-C student ratings · Teacher perceptions · Context · Natural raters

E. T. Pas (✉) · C. P. Bradshaw
Johns Hopkins Center for the Prevention of Youth Violence, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, Baltimore, MD 21205, USA
e-mail: epas@jhsph.edu

Teachers serve as the natural raters of students within the school and classroom contexts (Kellam 1990) and are often tasked with both formally assessing student performance (e.g., grades, student behavior) and informally assessing students' day-to-day performance and ability to meet academic or social demands (Kellam 1994; Kellam et al. 1998; Kellam and Rebok 1992). Teachers also are a source of information when there are concerns regarding students' progress, both in regular school practice and often as indicators of child outcomes in prevention research trials. In fact, teacher ratings of child behavior are among the most widely collected assessment tools in schools (Merrell 2001). Yet, teacher ratings of their students may vary based on school and classroom contextual factors (Kellam 1994). This variability has rarely been studied, but is critical for informing our understanding of the extent to which individual teacher and contextual factors influence ratings of student behavior over time.

## Validity of Teacher Ratings

Several studies have explored the association between different types and sources of ratings (e.g., the relationship between student and teacher ratings of behavior) and have generally shown low to moderate correlations between different raters (Renk and Phares 2004; Wright and Torrey 2001) and low correlations with observations of student behavior (Skiba 1989). Teacher ratings of student behaviors generally have been shown to have greater convergent validity with student ratings and school outcomes than parent ratings (e.g., Achenbach et al. 1987; Dwyer et al. 2006; Renk and Phares 2004; Wright and Torrey 2001), although this is not always the case (cf. Stanger and Lewis 1993). Nevertheless, ratings of youths' behavioral and social emotional well-being are subject to variability based on who completes a rating because contexts, and therefore displayed behaviors, vary (Merrell 2001). In addition, different raters construct unique standards

and conceptualizations of behaviors and performance (e.g., social competence; Renk and Phares 2004), which also leads to variability in ratings. In fact, ratings of student behavior are inherently biased, as each rater has observed and processed events and behaviors in a unique way (Dobbs and Arnold 2009). In psychometric terms (Crocker and Algina 1986), one can conceptualize that ratings provided regarding student behavior are made up of *true* student behavior and that the *error* involved in ratings is comprised of contextual and rater factors (Mangelsdorf et al. 2000; Kellam 1994; Mascendaro et al. 2012), in addition to regularly expected measurement error. Therefore, the agreement between raters may in fact be reflective of the true behavior, whereas the non-overlapping information provided by multiple raters may in part be the error in the ratings, comprised of contextual and rater factors (Dobbs and Arnold 2009). This error or variability that the rater introduces to the rating process has not been explicitly measured in the extant literature on teacher ratings and may in part be the result of rater experiences and perceptions of context.

## Importance of Teacher Perceptions and Ratings of Students

Beyond the fact that teachers are tasked with rating students, thus highlighting the importance of accuracy and objectivity of these ratings, literature has also established that teachers' perceptions of their students have important implications for student success (Kellam 1990; Kellam et al. 1975). Kellam's life-course social fields theory suggests that teachers are natural raters of their students' behavior, which in turn influences how the teachers interact with those students and the interactions among students across time and development. At the most basic level, teachers' interactions with students are in part related to their perception of students. For example, a study of preschoolers found that teaching style varied by perceptions of student behavior, such that teachers used more commands with students who they perceived to display problem behaviors, controlling for the objectively observed student behaviors (Dobbs and Arnold 2009). In a more distal effect, the quality of student-teacher relationships has also been linked with student achievement (Hamre and Pianta 2005). Therefore, other than the measurement implications, understanding the variability in teacher ratings of students as a function of teacher experiences and perceptions of the school context may inform intervention literature about ways to enhance how teachers perceive their students, thus improving student-teacher relationships.

Related research on parent ratings of their children's behavior and adjustment indicates that parents' depression and/ or other mental health concerns influence their perceptions, and therefore, their ratings of their child's behavior and social-emotional well being (McDermut et al. 1997). For example, previous research has indicated that depressed parents are more likely to over report behavioral problems when compared to non-depressed parents (Berg-Neilsen et al. 2003; Webster-Stratton and Hammond 1988). However, there has been limited consideration of a similar phenomenon in school settings, whereby teachers who experience high levels of burnout or perceive their school environment to be less favorable may perceive their students' behavior less favorably. This issue requires greater consideration in the prevention literature, given that many prevention trials rely heavily on teacher ratings of student behaviors in assessing program outcomes (e.g., Conduct Problems Prevention Research Group 2002; Ialongo et al. 1999; 2001). Typically, little consideration is given to the teachers' own perceptions and experiences, when in fact these factors may greatly influence not only their interactions with their students (Jennings and Greenberg 2009) but their own ratings of the students. In addition, previous studies examining the link between parent mood and child ratings (e.g., Berg-Nielsen et al. 2003; Webster-Stratton and Hammond 1988) have only considered single time point measurement of children; additional knowledge regarding how raters assess children/students over time is also needed.

## The Current Study

The present study aimed to examine the extent to which teacher ratings of student behaviors varied over time as a function of teacher experience and perceptions of context, measured concurrently with the baseline assessment of student behaviors. Specifically, we examined teacher ratings of student behaviors across several domains, including concentration problems, emotion dysregulation, and aggressive behavior. Given the parent assessment literature (McDermut et al. 1997), which documents an association between stress and mental health with biased ratings of children's behavior, we were particularly interested in proxies for these factors in the classroom. For example, teachers' perception of burnout, low efficacy, and unsupportive work environment are likely functionally associated with their ratings of their students' classroom behavior. Utilizing a person-centered approach to identify profiles of teachers who share similarly negative perceptions of their environment, we explored whether these profiles were functionally associated with their ratings of student behaviors over time. In addition to informing our understanding of measurement issues, this work also has potential implications for identifying teachers in need of support, both to address their own personal issues and negative perceptions (Jennings and Greenberg 2009), and to enhance the quality of education supports provided to the students.

To explore these associations, we drew upon data from 702 teachers from 42 elementary schools who provided student data over a 3-year period. We used a person-centered approach called latent profile analysis (LPA) to group the teachers with similar profiles on their perceptions of the school environment as well as their own burnout and efficacy (at one time point) to account for heterogeneity in the patterns of these perceptions within the teachers. Then, the functional relationship between the profile assignments and their ratings of students' behavior problems was examined across three subsequent school years. This enabled us to determine whether teachers' experiences with the school or personal characteristics were associated with their appraisal of students' behavior over time, such that the perceptions of the school context preceded their student ratings.

We hypothesized that distinct latent profiles of teachers would emerge and that these profiles would be related to the teachers' ratings of student behavior across 3 years. Specifically, it was anticipated that teachers with more negative perceptions of their environment, higher rates of burnout, and lower efficacy would provide poorer and more consistent ratings of student behavior over time (Maslach and Jackson 1981; Skaalvik and Skaalvik 2007). The extent to which teacher perceptions of the environment and student behavior are associated is an important area to explore, as teachers are often called upon to rate students' performance in their daily functioning and are responsible for referring students who are in need of more assistance or have disciplinary concerns. Prevention researchers often use teacher ratings as outcomes (e.g., Petras et al. 2004) although little consideration has been given to the role the teachers' experiences may play in these ratings, contemporaneously or over time.

## Method

### Participants

Data come from a large-scale study of school reform models (Bradshaw et al. 2012). The current analyses used data from 702 general education teachers in 42 elementary schools across three school years. The majority of the teacher sample was female (i.e., 95.2 %), White (i.e., 81.8 %), and state-certified teachers (i.e., 97.3 %). Slightly less than half of the sample held a graduate degree (i.e., 46 %). On average, teachers had about 9 years (i.e., $M$ =9.02, SD=8.77) of teaching experience. Years of experience ranged from first-year teachers to veterans of 42 years, with a median experience of 6 years.

### Procedure

All general educators in the participating schools were asked to complete ratings of student behavior for students in their classroom for 3 years (i.e., for three sets of students), as well as self-report ratings regarding the school environment. Specifically, data were collected during the fall of the first year of the trial and the three subsequent springs. For the purposes of this study, teacher perception data at baseline and teacher ratings of student behavior across all four time points (over 3 years) were analyzed. Surveys were sent to each school containing individually-addressed survey packets for all staff and distributed to their school mailboxes. To ensure confidentiality, staff returned the teacher self-report surveys directly to the researchers via a self-addressed, stamped envelope. Each staff questionnaire packet included a small incentive for participation (e.g., a disposable ballpoint pen). Although staff participation in both the self-report assessment and ratings of student behavior was voluntary, the participation rates were high for a study of this scale and scope. Specifically, 76 % of eligible general educators submitted the self-report measures, and 94 % submitted ratings of student behaviors. All of the study procedures were approved by the investigators' Institutional Review Board. Teachers provided written consent for participation, and students' participation was covered through a passive parental consent process.

### Measures

*Teacher Ratings of Student Behavior* Teachers provided ratings of each student's classroom behaviors at four time points, using an augmented version of the Teacher Observation of Classroom Adaptation—Checklist (TOCA-C; Koth et al. 2009), which includes six subscales. The TOCA-C is a checklist version based on the original TOCA-R used in several prevention trials (e.g., Ialongo et al. 1999; Petras et al. 2004; Werthamer-Larsson et al. 1991). The TOCA-C included a seven-item subscale assessing *concentration problems* (e.g., pays attention, is easily distracted; Cronbach's alpha [$\alpha$]= 0.95); a nine-item measure of *aggressive-disruptive behavior* (e.g., breaks rules, harms property, fights; $\alpha$ =0.90); a 5-item measure of *emotion dysregulation* (e.g., changes moods quickly, is impulsive; $\alpha$ =0.87); a five-item measure of *internalizing symptoms* (e.g., worries, is sad; $\alpha$ =0.87); a five-item measure of *prosocial behavior* (e.g., is friendly, is liked by classmates; $\alpha$ =0.89); and a five-item measure of *family involvement* (e.g., this child's parent(s) attend parent–teacher conferences; $\alpha$ =0.94). Teachers rated each student in their classroom on the TOCA-C using the six-point scale from "never" to "almost always." The concentration problems, disruptive behavior, emotion dysregulation, and internalizing symptoms subscales were scored such that higher scores reflected greater problems, whereas the prosocial behavior and family involvement subscales were scored such that higher scores reflected positive behaviors. Confirmatory

factor analyses on these data confirmed the hypothesized factor structure of these six scales.

Prior research documents the validity of the TOCA-C (for a review, see Koth et al. 2009), including four-month test–retest correlations of 0.75 or higher and alpha coefficients above 0.80. Furthermore, previous studies in other samples using the aggressive-disruptive behavior and concentration subscales from the original TOCA have demonstrated the predictive validity of elementary school scales on violence in adolescence and young adulthood (e.g., Petras et al. 2004). In the current study, since we were interested in teachers' ratings of students over multiple years, we averaged the student ratings on each subscale and aggregated to the teacher level, thereby assigning each teacher a set of six average TOCA-C scores across four time points. These six average scores served as the outcome variables of interest in the current study.

*Teacher Beliefs and Perceptions* Teachers' perceptions of the school environment were provided using the 31-item Organizational Health Inventory (Hoy and Feldman 1987), which included scales of *collegial leadership* (ten items, e.g., the principal conducts meaningful evaluations, the principal treats all faculty as his or her equal, $\alpha = 0.94$), *teacher affiliation* (nine items, e.g., there is a feeling of trust and confidence among the staff, teachers identify with the school; $\alpha = 0.88$), *academic emphasis* (five items, e.g., students respect others who get good grades, students neglect to do homework; $\alpha = 0.68$), and *resource influence* (seven items, e.g., extra materials are available if requested, $\alpha = 0.85$). Teachers also completed five additional subscales assessing different beliefs and perceptions related to their work as educators. A *school organizational climate* scale (Glisson and Green 2006) was adapted for use in this study; it included seven items (e.g., this school rewards experience, dedication, and hard work; $\alpha = 0.82$). Two scales pertained to the perceived support by the schools' administration related to universal, *school-wide supports* to students (three items; e.g., the principal allocates time and resources to support the school-wide prevention programs; $\alpha = 0.94$) and use of advanced, *selected and indicated* preventive interventions (six items; e.g., the principal allocates time and resources to implement individual intervention plans; $\alpha = 0.86$); these two scales were developed based on the work of Kam et al. (2003; also see Debnam et al. 2012). The five-item measure of *teacher efficacy* (Hoy and Woolfolk 1993) addressed perceptions of one's efficacy in handling student behavior problems (e.g., I can effectively work with deviant or disruptive students, I can manage almost any student behavior problem; $\alpha = 0.84$). The *burnout* scale was comprised of four items from the emotional exhaustion component of the Maslach Burnout Inventory (Maslach et al. 1996) and included items such as "I feel emotionally drained from my work" and "I feel like I am at the end of my rope" ($\alpha = 0.90$). Item responses on all nine of the staff-report

measures were based on a four-point Likert-type scale, which ranged from *rarely occurs* to *frequently occurs*. In all cases, scale scores were created by taking the average of responses across items. Although these data were collected across four time points as a part of the larger study, only the baseline data were used in the current paper, as they serve as predictor variables (i.e., the teacher perceptions) that precede the student outcome data (i.e., ratings of student behaviors).

*Teacher Demographics* Data regarding the individual teachers' demographic information were assessed through a brief self-report measure. The demographic measure included their *gender* (coded 1=male, 0=female), *race/ethnicity* (coded White=1, not White=0), the number of *years of teaching experience*, *education level* (coded 1=received a graduate degree, 0=no graduate degree), and *teacher certification* (coded 1=received a regular or standard state certification, 0=is not state-certified).

Analyses

*Latent Profile Analysis of Teacher Ratings* LPA (e.g., Hagenaars and McCutcheon 2002; Muthén and Muthén 1998–2010) was conducted using the M*plus* 6.1 software and was used to create latent profiles of teachers based on the teachers' ratings of organizational health, burnout, and efficacy. LPA is a latent class approach for continuous data, in which non-observable profiles are generated based on a set of continuous manifest indicators. The LPA classifies individuals based on similar response patterns across multiple indicators. In order to determine the accurate number of profiles, three fit indices and two tests were examined: the Akaike Information Criteria (AIC), Bayesian Information Criterion (BIC; Schwartz 1978), sample size adjusted BIC, Lo–Mendell–Rubin likelihood ratio test (LMR; Lo et al. 2001), and Vuong-LMR likelihood ratio test (Muthén and Muthén 1998–2010). Models were built iteratively, such that one profile was added at a time, to determine improved fit based on decreasing AIC, BIC, and adjusted BIC and the statistically significant LMR and Vuong-LMR. A non-significant LMR and Vuong-LMR indicates that the addition of a class does not result in a significant improvement in model fit. Other quantitative output, such as the entropy and posterior probabilities of the latent class assignment, were also considered, where an entropy score closest to 1.00 and posterior probabilities greater than 0.70 were favored (Nagin 2005; Ramaswamy et al. 1993). Finally, whether the profiles were of an adequate size and conceptually and theoretically meaningful was taken into account (Muthén 2004; Nylund et al. 2007). Clustering at the school level was modeled through the use of the Huber–White sandwich estimator to adjust the standard errors (Muthén and Muthén 1998–2010).

*Teacher-Level Ratings of Student Behaviors Across Time* The classroom-level average TOCA-C subscale ratings (i.e., on concentration problems, disruptive behavior, emotion dysregulation, internalizing, prosocial behavior, and family involvement) were analyzed using M*plus* to determine whether there was variability in the teachers' TOCA-C ratings over the four time points, across three consecutive school years. It is important to keep in mind that these analyses were conducted at the classroom level, and thus are groups of students in the classroom rather than individual students, as the students change classrooms each year. Each longitudinal TOCA-C outcome was regressed on the latent profiles of teacher perceptions and beliefs as well as on demographic characteristic variables.

*Missing Data* The target sample included 1,175 teachers who *ever* worked in a sample school and self-identified as a general educator and/or submitted student data (i.e., only general educators were asked to complete ratings of student behavior). In this sample, 71 teachers reported that they were general educators but were dropped from the analyses because they never provided student data (i.e., either they inaccurately reported being a general educator or were among the few teachers who did provide student data). Of this target sample, 862 teachers were present in the first school year and therefore were able to provide data on LPA indicators in the fall assessment of year 1 (i.e., the 313 teachers only present in years 2 or 3 were excluded). An additional 160 teachers were present in the first year but did not provide data for the LPA indicator variables and thus were dropped from the analyses. This missing data pattern on the teacher ratings of the school environment (i.e., the 160 teachers who were present in year 1 but did not provide data) is reflective of the response rates on teacher-ratings of the school environment, which ranged from 71 to 74 % across the four data collection time points. Chi-square analyses indicated no significant differences between those with complete data and those missing on the LPA indicators by gender $\chi^2$ (1, $N = 1,173$) = 3.53 ($p = 0.06$), ethnicity (i.e., White vs. non-White) $\chi^2$ (1, $N = 1,165$) = 0.72 ($p = 0.40$), receipt of a graduate degree $\chi^2$ (1, $N = 1,173$) = 0.27 ($p = 0.61$), or certification status $\chi^2$ (1, $N = 1,173$) = 1.64 ($p = 0.20$). These analyses suggest that we can assume the data are missing at random (Little 1995).

## Results

### Latent Profile Analyses

A series of models with up to five latent classes was fit. The best fit for the latent profile analysis of teacher-reported perceptions and organizational factors included three profiles of

teachers (LMR $p < 0.05$ for a three-class solution, entropy = 0.88; See Table 1). Though the four- and five-class models indicated improved AIC and BIC fit indices, the non-significant LMR tests and the fact that the additional classes did not add conceptual or theoretical meaning, indicated that the three-class solution was best. The posterior probabilities were 0.93, 0.94, and 0.96, and the smallest group within this three-class solution was 13 % of the total sample. Specifically, there was a group of teachers that generally rated the organizational health as lower than the sample average (with some variability across the different scales) and their own burnout higher, comprising 12.6 % of the sample; we labeled this class the "negative" class. The second class provided ratings of the school environment and burnout which were similar to the grand-mean ratings for the sample; this class comprised 39.2 % of the sample and we labeled it "normative" given their scores were generally close to the sample mean. The third and final class generally rated the organizational health higher and their own burnout lower than average and comprised 48.2 % of the sample (hereafter referred to as the "favorable" group; see Fig. 1 for a graphical depiction of the three profiles). Interestingly, all three groups had similar ratings of efficacy.

### Growth Modeling of Teacher-Rated Classroom-Level Student Behaviors

A series of latent growth trajectories of classroom-level student behavioral TOCA-C ratings was also fit across four time points. In all but two cases (i.e., on all subscales except for Disruptive Behavior and Internalizing), the AIC, BIC, and adjusted BIC indicated minimal or no change with the addition of classes and both LMR tests were non-significant for the two-class solution, indicating a non-significant improvement in the two-class solution as compared to the one-class solution. This indicates that there was no significant variability in the growth of teachers' ratings on the TOCA subscales over the course of 3 years. In the cases of disruptive behavior and internalizing symptoms, though the two-class growth model demonstrated a significant LMR and Vuong-LMR and more notable changes in the AIC, BIC, and adjusted BIC were observed, the groups were >99 % and <1 % of the sample in both cases, indicating that the second group was not of adequate size and did not provide substantive meaning (Muthén 2004; Nylund et al. 2007). The three-class solutions of disruptive behavior and internalizing symptoms did not demonstrate significantly improved fit over the two-class solutions, as judged by the LMR tests, and demonstrated the same small subsamples as well as markedly lower entropy scores. Therefore, the one-class growth model was used for all of the TOCA subscales. In addition, the linear model had the best fit (as compared to non-linear and quadratic) for each TOCA subscale, based on comparisons of the RMSEA, CFI,

**Table 1** Fit statistics for Organizational Health Inventory latent profile analysis

| Classes | Log likelihood | AIC | BIC | Sample size adjusted BIC | Vuong-LMR | LMR | Entropy | Class proportions |
|---|---|---|---|---|---|---|---|---|
| 1 | −6636.777 | 13309.554 | 13393.787 | 13336.627 | | | | |
| 2 | −5572.472 | 11200.943 | 11331.972 | 11243.057 | $p=0.000$ | $p=0.000$ | 0.893 | 480 (60 %) 316 (40 %) |
| 3 | −5243.911 | 10563.823 | 10741.647 | 10620.977 | $p=0.045$ | $p=0.047$ | 0.882 | 384 (48 %) 312 (39 %) 100 (13 %) |
| 4 | −5111.849 | 10319.698 | 10544.319 | 10391.892 | $p=0.175$ | $p=0.779$ | 0.828 | 300 (38 %) |
| | | | | | | | | 255 (32 %) |
| | | | | | | | | 163 (20 %) |
| | | | | | | | | 78 (10 %) |
| 5 | −5044.881 | 10205.763 | 10477.179 | 10292.998 | $p=0.452$ | $p=0.455$ | 0.833 | 286 (36 %) |
| | | | | | | | | 241 (30 %) |
| | | | | | | | | 150 (19 %) |
| | | | | | | | | 97 (12 %) |
| | | | | | | | | 22 (3 %) |

Values in italics indicate best fitting class

*AIC* Akaike Information Criteria, *BIC* Bayesian Information Criterion, *LMR* Lo–Mendell–Rubin likelihood ratio test

TLI, and the SRMR. Therefore, in all analyses, the TOCA growth models were specified as linear.

Predictive TOCA-C Scale Models

The final models regressed the single-class, linear growth of TOCA-C teacher ratings on teacher demographics as well as their profile defined by perceptions regarding the school environment, teacher efficacy, and burnout. These effects were tested for the intercept and for the slopes of each outcome (i.e., the growth of TOCA-C ratings across the 3 years, as measured in the fall of the first year and the three consecutive springs; see Table 2 for a full listing of findings). Given the high entropy (>0.80) and posterior probabilities (i.e., posterior probabilities=0.93, 0.94, and 0.96, all of which were >0.90),

individual teachers were assigned a hard classification of either favorable, normative, or negative and this variable was dummy coded and treated as two categorical manifest, rather than latent, covariates in the growth models.

*Concentration Problems* Analyses revealed a significant association between the teachers' LPA class membership on ratings of concentration problems at the first time point (i.e., intercept) but not for changes in ratings over time. Specifically, the teachers in the favorable class (i.e., the group of teachers who perceived the context as the most positive) had lower ratings of concentration problems at the intercept than the average group ($\beta=-0.17$, $p<0.01$). Certified teachers also had lower ratings of concentration problems at the intercept ($\beta=-0.26$, $p=0.02$). There were no significant



**Fig. 1** Latent profiles using teacher ratings of school context, burnout, and efficacy. Data points on the x-axis represent the teacher rating scales; CL = collegial leadership, RI = resource influence, AE = academic emphasis, TA = teacher affiliation, OR = organizational climate, SS = school-wide supports, SI = selected and indicated interventions, BO = burnout, and TE = teacher efficacy
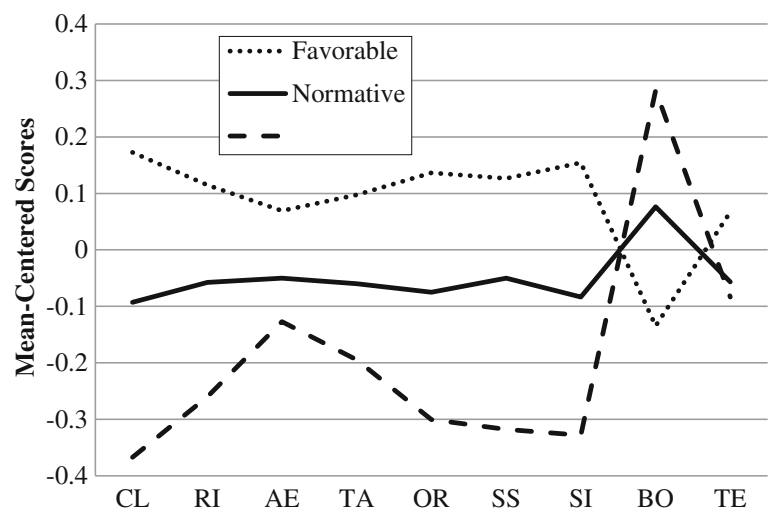
**Table 2** Relationship between teacher perceptions, beliefs, and demographics and teacher ratings of student behavior across 3 years

| | Perceptions (LPA) | | Teacher demographics | | | | |
|---|---|---|---|---|---|---|---|
| TOCA Subscale | Negative[a] | Favorable[a] | Gender | Race | Years | Degree | Certified |
| Concentration problems | | | | | | | |
| Intercept | 0.12 (0.07) | −0.17 (0.00) | 0.11 (0.06) | −0.05 (0.36) | 0.00 (0.53) | −0.03 (0.46) | −0.26 (0.02) |
| Slope | −0.03 (0.43) | −0.01 (0.35) | 0.02 (0.52) | −0.02 (0.50) | 0.00 (0.31) | 0.00 (0.91) | 0.02 (0.79) |
| Disruptive behavior | | | | | | | |
| Intercept | 0.13 (0.02) | −0.07 (0.03) | 0.07 (0.24) | −0.08 (0.05) | 0.00 (0.10) | −0.06 (0.03) | −0.19 (0.04) |
| Slope | −0.03 (0.17) | −0.03 (0.07) | 0.01 (0.68) | −0.03 (0.03) | 0.00 (0.73) | 0.00 (0.80) | 0.02 (0.77) |
| Emotion dysregulation | | | | | | | |
| Intercept | 0.14 (0.06) | −0.08 (0.10) | 0.15 (0.05) | 0.09 (0.07) | 0.00 (0.18) | −0.05 (0.23) | −0.33 (0.01) |
| Slope | −0.06 (0.03) | −0.03 (0.05) | 0.04 (0.27) | −0.05 (0.04) | 0.00 (0.19) | 0.00 (0.92) | 0.07 (0.18) |
| Family involvement | | | | | | | |
| Intercept | −0.06 (0.53) | 0.20 (0.00) | −0.21 (0.12) | −0.01 (0.87) | 0.00 (0.32) | 0.15 (0.01) | 0.12 (0.42) |
| Slope | −0.01 (0.80) | 0.05 (0.04) | 0.07 (0.18) | 0.08 (0.02) | 0.00 (0.80) | −0.01 (0.73) | 0.02 (0.71) |
| Internalizing | | | | | | | |
| Intercept | 0.10 (0.14) | −0.10 (0.03) | 0.16 (0.02) | 0.09 (0.09) | 0.00 (0.90) | −0.01 (0.86) | −0.28 (0.04) |
| Slope | −0.02 (0.51) | −0.02 (0.09) | 0.01 (0.87) | −0.02 (0.41) | 0.00 (0.62) | −0.01 (0.53) | 0.09 (0.09) |
| Prosocial behavior | | | | | | | |
| Intercept | −0.11 (0.12) | 0.21 (0.00) | −0.16 (0.03) | 0.07 (0.26) | 0.00 (0.07) | 0.05 (0.29) | 0.28 (0.03) |
| Slope | 0.03 (0.37) | 0.03 (0.17) | 0.01 (0.58) | 0.03 (0.24) | 0.00 (0.65) | 0.00 (0.84) | −0.04 (0.54) |

Estimates are tabled, with p-values in parentheses. Significant findings are in italics. The intercepts have been grand-mean centered so that the intercept reflects the average TOCA rating in the sample. Low indicates that the teacher was hard classified as having below average perceptions of the environment and high indicates above average perceptions. Gender is coded, *1* male; Race is coded, *1* White; Years is a continuous variable indicating years of experience; Degree is coded, *1* has graduate degree; and Certified is coded, *1* is a certified teacher. $N = 702$

[a] The reference group is the normative or average class

relationships between the perceptions of the school environment or demographics on the slopes of concentration problems (i.e., growth over time). Concentration ratings appeared to be stable over time.

*Disruptive Behavior* Significant effects for perceptions of the environment and demographics also emerged for the intercept of disruptive behavior. The latent class of teachers characterized by negative perceptions had higher ratings of disruptive behaviors ($\beta = 0.13$, $p = 0.02$) than the normative group, whereas the favorable group had lower ratings of disruptive behaviors than the normative group ($\beta = -0.07$, $p = 0.03$) at the intercept. White teachers ($\beta = -0.08$, $p = 0.05$), those with a graduate degree ($\beta = -0.06$, $p = 0.03$), and those with a state certification ($\beta = -0.19$, $p = 0.04$) had lower ratings of disruptive behavior at the intercept. White teachers also showed lower levels of growth in their ratings of disruptive behavior over time ($\beta = -0.03$, $p = 0.03$) than teachers who were not White. Other demographics and perceptions were not significantly related to ratings of disruptive behavior over time.

*Emotion Dysregulation* There were no significant associations between the teacher perception profile and the intercept of emotion dysregulation; however, there were significant effects of teacher perceptions for the slope of emotion dysregulation such that both the negative ($\beta = -0.06$, $p = 0.03$) and favorable ($\beta = -0.03$, $p = 0.05$) groups showed lower rates of growth in their ratings of emotion dysregulation over time as compared to the normative group. Male teachers had higher ratings of emotion dysregulation ($\beta = 0.15$, $p = 0.05$) and certified teachers had lower ratings ($\beta = -0.33$, $p = 0.01$) at the intercept. White teachers also showed less growth in their ratings than teachers who were not White ($\beta = -0.05$, $p = 0.04$). Other teacher demographics and teachers' perceptions and beliefs were not related to the intercept of emotion dysregulation.

*Internalizing Symptoms* The teachers in the favorable latent class had lower ratings of student internalizing symptoms ($\beta = -0.10$, $p = 0.03$) at the intercept than the average group. Also, male teachers had higher ratings ($\beta = 0.16$, $p = 0.02$) at the intercept than female teachers, and certified teachers had lower ratings than those who were not certified ($\beta = -0.28$, $p = 0.04$). Teacher demographics and the perceptions of the school environment were not related to the slope of ratings of internalizing symptoms.

*Prosocial Behavior* The favorable latent class of teachers showed significantly higher ratings of prosocial behavior (β =0.21, *p* <0.01) at the intercept than the normative group. Also, male teachers had lower ratings at the intercept than female teachers (β=−0.16, *p* =0.03) and certified teachers had higher ratings than those who were not certified (β=0.28, *p* = 0.03). Teacher demographics and their perceptions and beliefs were not related to the slope of ratings of prosocial behavior.

*Family Involvement* The favorable latent class of teachers had significantly higher ratings of family involvement for the intercept (β=0.20, *p* <0.01) and greater growth (β=0.05, *p* =0.04) than the normative group. White teachers also showed greater growth in their ratings of family involvement as compared to teachers who were not White (β=0.08, *p* =0.02). Other teacher demographics were not related to the intercept or slope of family involvement.

## Discussion

The purpose of this paper was to explore whether distinct profiles of teachers' beliefs and perceptions of context could be identified and whether these profiles were functionally related to trajectories of teacher-rated student behavior across 3 years. It should be noted that these analyses are at the teacher, not student, level as students change classes each year. Thus, data regarding three different cohorts of students were analyzed. The LPA results revealed three distinct profiles of teachers' beliefs and perceptions; (a) teachers who rated the school environment and their own burnout and efficacy to handle behavioral problems normatively or in line with the grand mean (39.2 %; normative class), (b) those who saw the school more positively and had lower burnout (48.2 %; favorable class), and (c) those who perceived the environment and their own burnout more negatively (12.6 %; negative class).

In addition, there was some evidence of the functional association between the latent teacher profiles and the intercept ratings of student behavior as well as growth in ratings over time. More specifically, the profiles identified were related to the intercept of all TOCA-C subscales except for emotion dysregulation. Teachers in the favorable class viewed students more positively at the intercept (i.e., they provided lower ratings of concentration problems, disruptive behavior, and internalizing as well as higher ratings of prosocial behaviors and family involvement). Teacher ratings of emotion dysregulation increased more slowly for teachers with favorable and negative perceptions than the normative class. Teachers with favorable perceptions showed greater increases of their ratings of family involvement over time than teachers with a normative profile.

Teacher demographics were also associated with the intercept and slope for several TOCA subscales. Most consistently,

White teachers and certified teachers provided more favorable ratings of students both at the intercept and across time. Teacher gender and having a graduate degree were less consistently related to the outcomes, and the years of experience variable was not related to any of the TOCA-C intercepts or slopes. Although several tests were conducted to assess these outcomes, the high number of significant findings (i.e., 23) suggests that these findings are not likely due to chance; however, they should be should be interpreted with caution.

Teacher ratings of student behavior were also relatively stable across time, despite the fact that these data were collected across three school years (i.e., in the fall and spring of the first year and in two subsequent springs) and were about three different cohorts of students. Growth in teacher ratings over time also did not significantly vary across teachers. This suggests that teachers may take a normative approach in completing student behavior ratings whereby their overall perceptions of students are stable, even though the students change. When variability was explored further in preliminary hierarchical linear models (i.e., where the variance estimates for each rating served as the outcomes), it was re-confirmed that variance was low across the six ratings, but that interestingly, the variability was highest for the two positive scales (i.e., family involvement and prosocial behavior). Future research should examine this phenomenon in other samples and further examine the within classroom variations of student ratings.

Limitations and Future Directions

This study is unique in its analytic approach and its goal of exploring the elements of teacher ratings which may be related to perceptions of context; however, it is not without limitations. The scope and extent of our data are limited solely to teacher ratings of all elements (i.e., teacher-provided demographics, perceptions, and student ratings) and therefore we cannot validate the teacher rating data. To do so, an additional source of data, such as observations, would be needed and is an area for future research. Since teachers were the sole providers of the surveys collected, when data were not provided, there also was no way to include them in the analysis, which led to listwise deletion. More specifically, when teachers did not provide the context, efficacy, and burnout data used in the LPA, they were not eligible for inclusion in the models. Attrition analyses based on demographics indicated that there were no significant differences in teacher demographics for those teachers included and excluded from the study, however, and therefore missing data were treated as missing at random (Little 1995).

The fact that only teachers in elementary schools were included in this research allowed it to be conducted with some practical ease, but was also a limitation of this study. Elementary school teachers spend nearly the entire day with their students, emphasizing the importance of their

perceptions of their students and eliminating further difficulty in the modeling. However, these findings may not be generalizable to middle and high schools; further, data from these school levels would likely present additional research opportunities. For example, having multiple observers/raters of the same student would allow for validation of the data but would also complicate the analyses because each classroom is a different compilation of students.

Because data are from multiple years and include ratings of three cohorts of students, the findings are not student-specific; rather, they are teacher-specific. This is also a strength as it is a unique approach to studying the issue of rater effects, but it does mean that conclusions cannot be drawn about students. Further, there are likely unmeasured nuances about the assignment of students to classrooms which are not accounted for in these analyses; not only is the process of classroom assignments not random, but it may be guided by knowledge of teacher skills as well as student behavior and achievement. Beyond the non-random assignment of students to classrooms, there are also possible classroom and school-level interventions being implemented which could have varied over time.

Finally, this study does not provide information regarding the direction of the relationship between teachers' perceptions of the context and ratings of students. For example, it is possible that teachers become more disillusioned with the context because they are assigned difficult students. Alternatively, their perceptions may skew how they perceive students. Further, their negative perceptions of the context and students can impact how they interact with students. Though this is of interest in the current study, the data do not allow these nuances to be teased out. More comprehensive data regarding the assignment of students to classrooms, school leadership, and additional sources of data regarding student behavior and school context are necessary to further elucidate the complexity of factors that influence teacher perceptions and their ratings of students. With this additional data, future research could identify potential factors that influence the development of teacher perceptions over time and assess the relationship between this process and growth in teacher ratings of students; it is quite possible that these are parallel processes that influence each other.

Implications for Prevention Science

These findings suggest that teachers' ratings of students' behavior are associated with their own experiences and perceptions of the school, such that teachers who have negative perceptions of the school and experience high burnout generally rate their students less favorably than their colleagues. Consistent with the life-course social fields theory (Kellam 1990), teacher perceptions of their students likely influence the way in which teachers interact with their students and in turn the way that their students interact with each other. Teachers' ratings on the

TOCA-C were closely linked with the teachers' latent class membership, but these associations were most consistent for the intercept ratings, rather than actual changes over time (i.e., slope). Taken together, these findings suggest that there may be "set" level differences between teachers in their overall perception of the level of student behavior problems, which are attributable at least in part to their personal experiences and perceptions of context. This is an important area for prevention research, as many prevention trials have relied heavily on teacher-ratings of student behavioral and mental health concerns (e.g., Ialongo et al. 1999; 2001).

In addition, it seems that when examined across time and different cohorts of students, teachers tend to rate students normatively. In other words, it does not appear that teachers vary their ratings of different groups of students from year to year, even though the students do in fact vary year to year. Rather, they may simply recalibrate their ratings based on the sample they rate each year. This is potentially disconcerting from a prevention research perspective, as it suggests that there may be some type of halo effect (Thorndike 1920), whereby teachers provide overall impressions of their students rather than differentiating between specific behaviors across years. This may be indicated by ratings across many domains that are rated similarly (e.g., students being rated as poor in behavior and academics, see Espinosa and Laffey 2003). When providing ratings for outcome analyses, halo effects will diminish variability and therefore make it more difficult to detect treatment effects. In practice, the default to a halo impression of students by teachers can taint how they perceive and therefore respond to students (Pianta 1999). In addition, if teachers view student behavior as stable, they may also view it as unchangeable, which can serve as a barrier to the uptake and implementation of prevention programs. On the other hand, changing teachers' perceptions of their school context, burnout, and efficacy may result in changes to their perceptions and actions in response to student behavior. This has implications for student success and the prevention of negative outcomes as well as highlights considerations prevention researchers should make when relying solely on teacher ratings of student behavior as outcomes.

## References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213–232. doi:10.1037/0033-2909.101.2.213.

Berg-Nielsen, T. S., Vika, A., & Dahl, A. A. (2003). When adolescents disagree with their mothers: CBCL-YSR discrepancies related to maternal depression and adolescent self-esteem. *Child: Care, Health and Development, 29*, 207–213.

Bradshaw, C. P., Pas, E. T., Goldweber, A., Rosenberg, M., & Leaf, P. J. (2012). Integrating school-wide positive behavioral interventions and supports with tier 2 coaching to student support teams: The PBISplus model. *Advances in School Mental Health Promotion, 5*, 177–193.

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*, 19–35.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont: Wadsworth.

Debnam, K., Pas, E. T., & Bradshaw, C. P. (2012). Secondary and tertiary support systems in place among schools implementing universal School-wide Positive Behavioral Interventions and Supports (SWPBIS): A descriptive analysis. *Journal of Positive Behavior Interventions, 14*, 142–152. doi:10.1177/1098300712436844.

Dobbs, J., & Arnold, D. H. (2009). Relationship between preschool teachers' reports of children's behavior and their behavior towards those children. *School Psychology Quarterly, 24*, 95–105.

Dwyer, S. B., Nicholson, J. M., & Battistutta, D. (2006). Parent and teacher identification of children at risk of developing internalizing or externalizing mental health problems: A comparison of screening methods. *Prevention Science, 7*, 343–357.

Espinosa, L. M., & Laffey, J. M. (2003). Urban primary teacher perceptions of children with challenging behaviors. *Journal of Children and Poverty, 9*, 135–156.

Glisson, C., & Green, P. (2006). The effects of organizational culture and climate on the access to mental health care in Child welfare and Juvenile Justice systems. *Administration and Policy in Mental Health and Mental Health Services Research, 33*, 433–448.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*, 949–967.

Hoy, W., & Feldman, J. (1987). Organizational health: The concept and its measure. *Journal of Research and Development in Education, 20*, 30–38.

Hoy, W. K., & Woolfolk, A. E. (1993). Teachers' sense of efficacy and the organizational health of schools. *The Elementary School Journal, 93*, 356–372.

Ialongo, N. S., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *American Journal of Community Psychology, 27*(5), 599–641. doi:10.1023/a:1022137920532.

Ialongo, N. S., Poduska, J. M., Werthamer, L., & Kellam, S. G. (2001). The distal impact of two first-grade preventive interventions on conduct problems and disorder in early adolescence. *Journal of Emotional and Behavioral Disorders, 9*, 146–160.

Jennings, P. A., & Greenberg, M. T. (2009). The Prosocial Classroom: Teacher social and emotional competence in relation to child and classroom outcomes. *Review of Educational Research, 79*, 491–525. doi:10.3102/0034654308325693.

Kam, C.-M., Greenberg, M. T., & Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science, 4*, 55–63.

Kellam, S. (1990). Development epidemiological framework for family research on depression and aggression. In G. R. Patterson (Ed.), *Depression and aggression in family interaction* (pp. 11–46). Hillsdale: Erlbaum.

Kellam, S.G. (1994). Testing theory through developmental epidemiologically-based prevention research (pps. 37–57). In A. Cazares & A. Beatty (Eds.), *Scientific Methods for Prevention Research*. National Institute on Drug Abuse Research Monograph No. 139. DHHS Pub. No. 94-363.

Kellam, S. G., & Rebok, G. W. (1992). Building developmental and etiological theory through epidemiologically based preventive intervention trials. In J. McCord & R. E. Tremblay (Eds.), *Preventing antisocial behavior: interventions from birth through adolescence* (pp. 162–195). New York: Guilford Press.

Kellam, S. G., Branch, J. D., Agrawal, K. C., & Ensminger, M. E. (1975). *Mental health and going to school: the Woodlawn program of assessment, early intervention and intervention and evaluation*. Chicago: University of Chicago Press.

Kellam, S. G., Ling, X., Merisca, R., Brown, C. H., & Ialongo, N. S. (1998). The effect of the level of aggression in the first grade classroom on the course and malleability of aggressive behavior into middle school. *Development and Psychopathology, 10*, 165–185.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2009). Teacher Observation of Classroom Adaptation-Checklist (TOCA-C): Development and factor structure. *Measurement and Evaluation in Counseling and Development, 42*, 15–30.

Little, R. J. (1995). Modeling the dropout mechanism in repeated measures studies. *Journal of the American Statistical Association, 90*, 1112–1121.

Lo, Y., Mendell, N., & Rubin, D. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767–778.

Mangelsdorf, S. C., Schoppe, S. J., & Buur, H. (2000). The meaning of parental reports: A contextual approach to the study of temperament and behavior problems in childhood. In V. J. Molfese & D. L. Modfese (Eds.), *Temperament and personality development across the lifespan* (pp. 121–140). London: Erlbaum.

Mascendaro, P. M., Herman, K. C., & Webster-Stratton, C. (2012). Parent discrepancies in ratings of young children's co-occurring internalizing symptoms. *School Psychology Quarterly, 27*, 134–143. doi:10.1037/a0029320.

Maslach, J., & Jackson, S. E. (1981). The measurement of experienced burnout. *Journal of Occupational Behaviour, 2*, 99–113.

Maslach, C., Jackson, S. E., & Leiter, M. P. (1996). *Maslach Burnout Inventory manual* (3rd ed.). Palo Alto: Consulting Psychologists Press, Inc.

McDermut, J. F., Haaga, D. A. E., & Bilek, L. A. (1997). Cognitive bias and irrational beliefs in major depression and dysphoria. *Cognitive Therapy and Research, 21*, 459–476.

Merrell, K. W. (2001). Assessment of children's social skills: Recent developments, best practices, and new directions. *Exceptionality, 9*, 3–18.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345–368). Thousand Oaks: Sage.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles: Author.

Nagin, D. S. (2005). *Group-based modeling of development over the life course*. Cambridge: Harvard University Press.

Nylund, K. L., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14*, 535–569.

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*, 88–96.

Pianta, R. (1999). *Enhancing relationships between children and teachers*. Washington: American Psychological Association.

Ramaswamy, V., Desarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science, 12*, 103–124.

Renk, K., & Phares, V. (2004). Cross-informant ratings of social competence in children and adolescents. *Clinical Psychology Review, 24*, 239–254.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Skaalvik, E. M., & Skaalvik, S. (2007). Dimensions of teacher self-efficacy and relations with strain factors, perceived collective teacher efficacy, and teacher burnout. *Journal of Educational Psychology, 99*, 611–625.

Skiba, R. J. (1989). The importance of construct validity: Alternative models for the assessment of behavioral disorders. *Behavioral Disorders, 14*, 175–185.

Stanger, C., & Lewis, M. (1993). Agreement among parents, teachers, and children on internalizing and externalizing behavior problems. *Journal of Clinical Child Psychology, 22*, 107–115. doi:10.1207/s15374424jccp2201_11.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25–29.

Webster-Stratton, C., & Hammond, M. (1988). Maternal depression and its relationship to life stress, perceptions of child behavior problems, parenting behaviors, and child conduct problems. *Journal of Abnormal Child Psychology, 16*, 299–315.

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

Wright, D., & Torrey, G. K. (2001). A comparison of two peer-referenced assessment techniques with parent and teacher ratings of social skills and problem behaviors. *Behavioral Disorders, 26*, 173–182.