Journal of Psychoeducational Assessment
1–11

DOI: 10.1177/0734282916629595
jpa.sagepub.com
SAGE

Emotional Bias in Classroom Observations: Within-Rater Positive Emotion Predicts Favorable Assessments of Classroom Quality

James L. Floman[1], Carolin Hagelskamp[2], Marc A. Brackett[3], and Susan E. Rivers[3]

Abstract

Classroom observations increasingly inform high-stakes decisions and research in education, including the allocation of school funding and the evaluation of school-based interventions. However, trends in rater scoring tendencies over time may undermine the reliability of classroom observations. Accordingly, the present investigations, grounded in social psychology research on emotion and judgment, propose that state emotion may constitute a source of psychological bias in raters' classroom observations. In two studies, employing independent sets of raters and approximately 5,000 videotaped fifth- and sixth-grade classroom interactions, within-rater state positive emotion was associated with favorable ratings of classroom quality using the Classroom Assessment Scoring System (CLASS). Despite various protections enacted to secure reliable and valid observations in the face of rater trends—including professional training, certification testing, and routine calibration meetings—emotional bias still emerged. Study limitations and implications for classroom observation methodology are considered.

Keywords

classroom climate, social and educational environment, education assessment, emotional intelligence, personality/individual differences, hierarchical linear/multilevel modeling, measurement

Systematic classroom observations reliably predict academic and social-emotional processes central to quality education (Pianta & Hamre, 2009). Classroom observation measures assessing emotional support, for example, predict elementary school students' reading and mathematics achievement (e.g., Rudasill, Gallagher, & White, 2010). Observation tools also serve as reliable outcome measures in school intervention research (Hamre et al., 2013), such as testing the effects of social-emotional learning programs on classroom climate and the quality of instruction (e.g.,

[1]University of British Columbia, Vancouver, Canada
[2]Public Agenda, New York, NY, USA
[3]Yale University, New Haven, CT, USA

Corresponding Author:
James L. Floman, Department of Educational and Counselling Psychology, and Special Education, University of British Columbia, 2125 Main Mall, Vancouver, British Columbia, Canada V6T 1Z4.
Email: floman@alumni.ubc.ca

Hagelskamp, Brackett, Rivers, & Salovey, 2013). Furthermore, classroom observations are used increasingly in high-stakes evaluations of teacher and school performance (Baker, Oluwole, & Green, 2013; Whitehurst, Chingos, & Lindquist, 2014).

Despite their utility, as with any mode of instrumentation, classroom observations have limitations. A range of factors contributes to variance in observational assessments. Drawing on Generalizability (G) theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972), factors that contribute to variance in observation scores, beyond true scores and random measurement error, include the following: (a) the rater (e.g., rater trends or traits), (b) the item or dimension (e.g., observation criteria or latent structure), (c) the time or setting (e.g., time of day or year), and (d) the method (e.g., live or video recording; see Hintze & Matthews, 2004).

Fluctuations or drifts in rater scoring tendencies over time, known as rater trends, are the focus of this study. A recent study using "augmented G-theory" estimated 4.5% of variation in observational assessments of classroom quality is specifically attributable to rater trends (Casabianca, Lockwood, & McCaffrey, 2015). Furthermore, drift within raters has been found to increase over time, despite calibration sessions designed to track and minimize trends in raters (Casabianca et al., 2015). Rater drift effects may thus diminish the psychometric quality of repeated classroom observations. However, if the various sources of rater trends are better understood, they could serve as points of improvement for applied observation methods.

In two studies, we examine rater state (i.e., moment-to-moment) emotion as a source of trend bias in classroom observations. Although rater severity and fatigue are likely to play roles in rater trends (Casabianca et al., 2013, 2015), rater emotion is an unstudied psychological factor that also may be a significant contributor to rater trends. Despite the fact that raters are often required to use standard criteria to assess information only witnessed during an observation, state emotions frequently alter judgments of various social phenomena (Forgas, 2014; Isbell & Lair, 2013). When the source of an emotion is identified correctly (e.g., observed, tense social exchanges make one feel anxious), emotion can enhance judgment (e.g., anxiety may lead one to correctly assess classroom climate as highly negative). Yet, when the source of an emotion is misattributed (e.g., one is anxious because of an upcoming interview, but thinks it is because of observed, tense social exchanges), this "incidental emotion" can cause judgment bias (e.g., one incorrectly assesses a classroom climate as highly negative; Schwarz & Clore, 1983, 2003).

Misattribution of emotion may be particularly common among observation raters because emotional bias is more prevalent in judgments founded on ambiguous versus concrete standards (Greifeneder, Bless, & Pham, 2011), and even among widely validated classroom observation measures many criteria are open to interpretation. For example, the creators of the Classroom Assessment Scoring System (CLASS) advise, "Because of the highly inferential nature of the CLASS, scores should never be given without referring to the manual" (Pianta, La Paro, & Hamre, 2008, p. 15). Others suggest that rater effects and trends found using the CLASS may reflect the "cognitive challenges [of] assessing so many high inference dimensions at the same time" (Casabianca et al., 2015, p. 24). The cognitively taxing nature of applying observation systems such as the CLASS may increase emotional bias because emotion more heavily informs judgments that require elaborate cognitive processing (Forgas & George, 2001). Given the highly inferential rating standards and the cognitive challenges of processing many of such standards concurrently, even validated classroom observation tools may be at risk for emotional bias.

Even though many classroom observation measures require raters to complete training and pass reliability testing, and developing judgment-specific expertise can attenuate the effects of emotion on evaluations (Englich & Soder, 2009), emotion can factor into judgments rendered by expert judges too (e.g., Redelmeier & Baxter, 2009). As such, rater emotion may constitute a source of bias in observations even among well-trained raters using validated instruments. However, the link between trained rater emotions and classroom observations is unstudied. Our

research aimed to address this gap in the scientific literature to enhance the study and practice of systematic classroom observations.

Research Overview

We investigated the extent to which the state emotions of trained raters were associated with their assessments of classroom quality, using the CLASS, an established and increasingly used classroom observation measure (Hamre et al., 2013). Emotion-cognition interaction research indicates that emotion-congruent judgments are a common result of emotion misattribution, and lead to more favorable versus unfavorable assessments of people, places, and events when experiencing positive versus negative state emotions, respectively (Forgas & Eich, 2012). Accordingly, we hypothesized that rater state positive emotion would be associated with more favorable observational assessments (Hypothesis 1), and rater state negative emotion would be associated with more unfavorable assessments of classroom quality (Hypothesis 2).

Study 1

Method

Sample: Observational segments and raters. The sample for Study 1 included 2,637 individual assessments of videotaped classroom observation segments (M = 14.80min, SD = 1.39min, range = 10.00-20.00min). Videotaped classroom observations were obtained from fifth- and sixth-grade English Language Arts (ELA) classes in an urban Catholic diocese in the northeastern United States. They were part of a two-year classroom intervention study (Hagelskamp et al., 2013). Observations for Study 1 were collected across the first academic year of the intervention study from 32 schools, 92 classrooms, and 69 teachers out of 62 schools (52% return rate), 155 classrooms (59% return rate), and 105 teachers (66% return rate) participating in the first year of the study.

Classroom observations were obtained from fifth- and sixth-grade teachers who were given equipment to record their ELA classes. Each teacher was asked to provide three videotaped classroom sessions. To increase observation reliability, teachers were asked to avoid sending videos of transition periods or of the first or last 30 minutes of the school day (Hamre, Goffin, & Kraft-Sayre, 2009). Every videotape was converted into two segments of equal duration that ranged from 8 to 20 min long, as suggested by Hamre et al. (2009). Segments were considered incomplete for rating if they were less than 8 minutes long, the audio quality was virtually incomprehensible, and/or students were not visible for the majority of the video. The percentage of codable segments (i.e., passed the CLASS exclusion criteria) was 83.3%.

Sixteen unique raters (69% female) were trained on the CLASS (see the classroom observation rating procedure below) and assessed the video segments. On average, raters were randomly assigned 165 segments (range = 109-193) to assess across a 10-week period. Raters first attended the established and intensive two-day CLASS training and passed a CLASS certification test. The test required them to code a set of videotaped classroom observations with a minimum reliability rate of 80% (Pianta et al., 2008). Reliability was determined by assessing observations within one point of master codes provided by the instrument developers on 8 out of 10 dimensions. In addition, raters attended weekly calibration meetings designed to identify and remediate rater drift. Raters whose reliability dipped during a meeting attended additional rating sessions with CLASS master trainers until they achieved reliability again. At no point did unreliable raters assess any classroom observations. Each of these training and reliability protocols are outlined in the CLASS Implementation Guide, as they increase the reliability of observation ratings and raters (Hamre et al., 2009). Accordingly, raters for this study are comparable in quality with those

of other studies that employ the CLASS and follow the same recommended training and reliability procedures.

Classroom observation rating procedure. For 10 weeks, raters completed two assessment blocks a day for 2 to 4 days per week. Typically, the first assessment block was completed in the morning (before noon) and the second assessment block was completed in the afternoon (before 4:00 p.m.). In each uninterrupted assessment block, raters watched and assessed three randomly assigned classroom observation segments (block time range = ~ 45 to 90min—with each observation segment ranging from 10 to 20min and each assessment taking from 5 to 10min). Breaks of 2.5 hours were mandated between assessment blocks to obviate rater fatigue. Random assignment of observations to raters protected against potential sources of bias during assessment—such as individual differences between raters—to obviate possible systematic biases in ratings.

Prior to each assessment block, raters completed a state positive and negative emotion survey. Reporting one's emotion directly before a judgment task reduces emotion-influenced assessments to a moderate degree by heightening conscious awareness of emotions, which often exert their impact unconsciously (McFarland, White, & Newth, 2003; Schwarz & Clore, 1983, 2003). This procedure served as a direct protection against state emotion bias, while also allowing for the detection of a rater emotion–observation score relationship. State emotion is the degree to which a person experiences an emotion at a particular moment and in a specific setting. In contrast, trait emotion is how a person feels across most moments and settings. Trait or between-rater emotion differences were not assessed because of the small rater sample ($N = 16$).

Measures

State positive and negative emotions. State positive and negative emotions were measured using the Positive Affect (PA) and Negative Affect (NA) subscales of the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988). The widely used PANAS has shown good internal reliability (αs = .89 and .85 for PA and NA, respectively; Crawford & Henry, 2004) and satisfactory test–retest reliability (rs = .74 and .70 for PA and NA across 2 months, respectively; Watson & Walker, 1996). Individuals rated the extent to which they were "currently experiencing" 10 positive (e.g., proud, interested) and 10 negative emotions (e.g., ashamed, nervous) on a 0 (*very slightly or not at all*) to 4 (*extremely*) scale. Composite scores for positive emotions and negative emotions were created by averaging across the 10 PA and 10 NA items (α = .89 and .79, respectively).

Classroom observations. The CLASS is a validated observation measure designed to assess classroom quality from preschool to high school (Hafen et al., 2012; Pianta et al., 2008). The CLASS operationalizes classroom quality as a superordinate construct focusing largely on teacher performance and classroom climate.[1] It is comprised of three domains, each of which contains multiple dimensions: *Emotional Support* (e.g., Positive and Negative Climate), *Classroom Organization* (e.g., Productivity and Behavior Management), and *Instructional Support* (e.g., Quality of Feedback and Language Modeling). Emotional support is the emotional tone of classroom social interactions, including teachers' responsiveness to students' needs and the degree of student autonomy embedded throughout the classroom. Classroom organization refers to the presence of effective classroom management practices, and the use of efficient and engaging instructional procedures and materials. Instructional support represents the extent to which teachers develop their students' higher-order rather than rote thinking skills and teachers responsively engage students via scaffolding advanced use of language.

Each of the 10 dimensions of the CLASS was rated on a 7-point scale ranging from 1-2 (*low*) to 3-5 (*mid*) to 6-7 (*high*). An overall score of classroom quality was created by averaging across the 10 CLASS dimensions. The reliability of the classroom quality measure was good (α = .84),

and consistent with CLASS validation work in pre-K to elementary school classes (α range = .79-.92; Hamre, Pianta, Mashburn, & Downer, 2007). We decided to employ the overall classroom quality score in our analysis because emotion biases present in complex social judgments may be more discernible when examined at a global rather than a local level of analysis (Casabianca et al. [2015] also reported composite-level classroom quality results). Also, composite ratings may be of greater value in ecologically valid measurement settings to guide evaluation and policy (e.g., see Pianta, 2012).

Analytic procedure. A multi-level regression model (MLM) was used to test the hypotheses that variation in rater state positive and negative emotions was related to their observational assessments of classroom quality. This approach accounted for non-independence in the data—a violation of a core assumption in linear regression. This was required because individual assessments (Level 1) were nested within assessment blocks (Level 2), which in turn were nested within individual raters (Level 3). To start, an unconditional intercept model was run. This model specified random effects at Levels 2 and 3 to estimate variation between assessment blocks as well as between raters, in addition to residual variation between individual assessments. Then, rater emotions, as reported at the beginning of each assessment block, were entered as a Level 2 predictor into the model. These positive and negative emotion scores were centered on each individual rater's positive and negative emotion means as reported across the 10-week assessment period. This procedure made it possible to estimate the within-rater association between state emotion and assessments of classroom quality, independent of between-rater (or trait emotion) differences. That is, this procedure allowed for an estimation of the degree to which raters assessed classroom quality more favorably versus unfavorably depending on fluctuations in their positive and negative emotions, respectively, relative to their own average emotion states across the 10-week period.

Results and Discussion

On average, across all assessment blocks, raters reported experiencing positive emotions "a little" ($M = 1.05$, $SD = 0.76$) and negative emotions "very slightly or not at all" ($M = 0.16$, $SD = 0.28$). These patterns of self-reported emotion, including relatively greater positive versus negative emotions, are comparable with prior research on the PANAS (Crawford & Henry, 2004).

The unconditional intercept model estimated an overall mean in classroom quality assessments of 4.29 ($SE = .06$). On average, raters assessed the overall quality of classrooms slightly above the midpoint on the 7-point scale. This is consistent with prior research that has tracked overall classroom quality scores using the CLASS over a two-year period (Casabianca et al., 2015). Intraclass correlation coefficients (ICCs) showed that the majority (89%) of variability in classroom quality was at Level 1. Level 1 included differences between individual assessments (i.e., variance that was a result of true classroom differences and random measurement error). The variability in classroom quality at Level 2 was 3%—a small portion of overall classroom quality. Level 2 variability was attributable to differences between assessment blocks (i.e., differences that were a result of temporary factors specific to the day and time the assessments occurred, including a rater's state emotion). Finally, 8% of variability in classroom quality was at Level 3. Level 3 included individual differences between raters (i.e., variance that resulted from differences in the way specific raters assessed the data).

Person-centered, state positive and negative emotions were entered in two models as Level 2 predictors. In support of Hypothesis 1, the parameter estimate for the within-rater association between positive emotions and classroom quality ratings was positive and significant, $b = .09$, $SE = .03$, $t(831.82) = 2.92$, $p = .004$. Hypothesis 2 was not supported. Negative emotions were not associated with classroom quality ratings, $b = .10$, $SE = .07$, $t(885.46) = 1.47$, $p = .14$.

Agreement regarding the optimal technique for calculating effect sizes in MLM remains unclear (Peugh, 2010; Roberts & Monaco, 2006); a well-matched equivalent of R^2 or Hedge's g does not exist for MLM. One approach that researchers have employed is to convert t values produced by MLM into Hedge's g or Cohen's d coefficients (see Fromme, Corbin, & Kruse, 2008; Oishi, Lun, & Sherman, 2007). It is important to note that these effect size conversion equations require df_{within} and MLM only provides an approximate df_{within}—potentially biasing the effect size estimate produced by this approach. Athough imperfect, we employed this technique (using conversion Equations 6 and 7 specified in Rosnow, Rosenthal, & Rubin, 2000; see the Appendix) to provide information on the relative significance of the emotion bias. The effect size for the positive emotion–classroom observation rating association at Level 2 (i.e., temporal variance between assessment blocks), which itself accounted for a small (3%) proportion of total classroom quality variance, was medium: Cohen's $d = 0.36$ (Cohen, 1992).

Temporary increases in rater positive emotions were associated with higher classroom quality ratings. These results suggest that state positive emotion played (at least a small) role in trained raters' classroom quality assessments, despite receiving reliability training and reporting their emotions prior to each rating, behaviors which can reduce emotion-influenced judgments (Englich & Soder, 2009; McFarland et al., 2003). This effect is consistent with the literature that shows positive emotions favorably bias high-inference social judgments (Forgas, 2014; Isbell & Lair, 2013). No evidence for a negative emotion bias emerged. The predicted emotion-congruency effect was not observed (see the General Discussion for explanations).

Study 2

The goal of Study 2 was to replicate the findings in Study 1 by gathering further evidence that state emotion is associated with observational assessments of classroom quality. An unexpectedly high percentage of findings in psychological science fail to receive direct replication, potentially undermining trust in the field. A call for direct replications was recently issued to redress this major problem (Pashler & Wagenmakers, 2012). Toward this end, a second set of raters' observational assessments was used to test the proposed hypotheses. The data set was structured in the same fashion as in Study 1, but it consisted of assessments from a different cohort of raters, who assessed a new set of classroom observations.

Method

Sample: Observational segments and raters. The sample for Study 2 consisted of 2,300 individual assessments of videotaped classroom observation segments ($M = 14.42$min, $SD = 2.55$min, range = 10.00-17.50min). The observations were collected from a new, second year of data collection from the same schools for the same intervention study used in Study 1. The data collection and video segmentation procedures along with the segment exclusion criteria were the same as in Study 1 as well. A total of 62 schools, 196 fifth- and sixth-grade classrooms, and 95 teachers participated in the second year of the study with 58 schools (94% return rate), 142 classrooms (65% return rate), and 68 teachers (72% return rate) providing classroom observation data. Nearly all (97.15%) observation segments obtained during this data collection period were codable (i.e., passed the CLASS exclusion criteria).

Fifteen raters (73% female) assessed the classroom observations. On average, raters assessed 153 segments (range = 91-197). The same training, reliability, and weekly calibration procedures described in Study 1 were used. Each of these steps were taken in accordance with standards established by the CLASS creators, as they contribute to greater rater quality and rating reliability (Hamre et al., 2009). As such, the quality of raters in this study is comparable to that of similar studies that also adhered to the CLASS training and coding protocols.

Classroom observation rating procedure. The same observation data collection procedure used in Study 1 was employed.

Measures

Positive and negative emotions. Scores for overall positive emotion and negative emotion were created by averaging across the 10 PA and 10 NA items of the PANAS (α = .93 and .83, respectively). The reliability levels are within range of those found in PANAS validation research (Crawford & Henry, 2004).

Classroom observations. A score for overall classroom quality was calculated by averaging across the 10 dimensions of the CLASS (α = .83). The reliability is consistent with that found in Study 1, and validation research on the CLASS (α range = .79-.92; Hamre et al., 2007).

Results and Discussion

Similar to Study 1, raters reported "a little" positive ($M = 1.20$, $SD = 0.81$) and "very slightly or not at all" negative emotion ($M = 0.20$, $SD = 0.31$) across assessment blocks. The unconditional intercept model estimated an overall mean in classroom quality assessments of 4.22 ($SE = .04$). This is just above the midpoint on the 7-point rating scale, as was found in Study 1 and in recent longitudinal data on the CLASS (Casabianca et al., 2015). ICCs showed that the majority (93%) of variability in classroom quality was at Level 1. Level 1 is made of differences between individual assessments (i.e., variability that was a result of true classroom differences and random measurement error). Again, a small amount of variability in classroom quality (3 %) was at Level 2. This variance is attributable to differences between assessment blocks (i.e., differences that were a result of temporary factors specific to the day and time the assessments occurred, including rater state emotions). Finally, 4% of the variability was at Level 3, which is comprised of individual differences between raters (i.e., variability that resulted from differences in the way raters assessed the data).

Person-centered, state positive and negative emotions were entered as Level 2 predictors of classroom quality. In support of Hypothesis 1, the parameter estimate for the within-rater association between positive emotions and classroom quality ratings was positive and significant, $b = .14$, $SE = .04$, $t(751.86) = 2.88$, $p = .000$. Again, Hypothesis 2 did not receive support. The parameter estimate for the within-person association between negative emotions and classroom quality ratings was not significant, $b = -.001$, $SE = .07$, $t(771.69) = -.08$, $p = .93$.

To calculate MLM effect sizes, Study 2 used the same approach as Study 1, to tentatively quantify the relative significance of the emotional bias (e.g., see Fromme et al., 2008). The *t*-value produced by the significant MLM analysis was converted into Cohen's *d* (using steps specified by Rosnow et al., 2000; see the Appendix). The effect size for the association between state positive emotions and observation ratings of classroom quality at Level 2 (i.e., temporal differences between assessment blocks), which itself accounted for a small amount of variance in total classroom quality (3%), was small: Cohen's $d = 0.21$ (Cohen, 1992).

Compared with Study 1, the effects differed somewhat in size, but the pattern of the effects was the same: state positive, but not negative emotions, were positively associated with rater assessments of classroom quality. As such, Study 2 replicated the findings from Study 1 with a new set of raters and a different set of classroom observations—suggesting that raters' state positive emotions may introduce a small degree of bias into systematic classroom observations.

General Discussion

Observations of classroom processes are contributing increasingly to high-stakes decisions in education (Baker et al., 2013; Whitehurst et al., 2014) and school-based intervention research

(Hamre et al., 2013). However, the social psychology research on emotion and cognition suggests that the inferential and cognitively demanding nature of classroom observations may raise their susceptibility to emotional judgment biases. Accordingly, Studies 1 and 2 found that rater state positive emotion predicted favorable observational assessments of classroom quality. State negative emotion was not related to these ratings. The results provide the first evidence of a relationship between trained rater emotions and their use of an observational instrument to assess ecologically valid classroom interactions.

These findings are consistent with established research in social psychology which shows that state emotions factor into a diverse array of human judgments, especially of complex, ambiguous social phenomena (Forgas, 2014; Isbell & Lair, 2013). The results suggest that classroom observations may be subject to inflationary biases from positive emotions akin to other methods of assessment, such as interviews (Redelmeier & Baxter, 2009), and other domains of assessment, such as job applications (Baron, 1987, 1993). Although more work is needed, this research contributes to growing evidence that emotion may be a factor, often unaccounted for in educational measurement, which may constitute a source of bias in need of correction (e.g., Brackett, Floman, Ashton-James, Cherkasskiy, & Salovey, 2013).

A positive but not negative emotion bias was found in classroom quality ratings. It is plausible that a negative emotion bias was not detected because of limitations in how negative emotions were measured. The PANAS has been criticized for lacking sensitivity to moderate and low arousing emotions, particularly for negative emotions. In support of this notion, and similar to a large-scale validation study of the PANAS (Crawford & Henry, 2004), reports of negative emotions were particularly limited in range (i.e., a floor effect was found) in this study. It is also plausible that negative emotions occurred too rarely to meaningfully factor into rater assessments, as negative emotions generally arise with less frequency in daily life than positive emotions in non-clinical adult samples (Diener, Kanazawa, Suh, & Oishi, 2014).

Study Limitations, Future Research Directions, and Implications

Our studies have limitations. Because emotions were measured and not manipulated, it remains unclear whether rater positive emotions caused bias in their observation ratings. It is important to note that emotions were measured directly prior to (not after) each assessment block, suggesting the expected direction of causality. However, research that induces state positive, negative, and neutral emotions in raters before they render their assessments is needed.

Another limitation of this research is that state positive emotions explained a relatively small portion of the trends in classroom quality, which itself accounted for a small amount of variance in overall classroom quality. That said, these findings were replicated and emerged under highly controlled circumstances. Raters reported their emotions directly prior to each assessment block, and further, all raters underwent training, attended weekly calibration meetings, and passed a reliability test to become certified raters. Both heightened emotion awareness (McFarland et al., 2003) and expertise can attenuate the effects of emotion on judgment (Englich & Soder, 2009). Bearing this in mind, it is noteworthy that even minor emotional bias was found under these atypically well-controlled rating conditions.

The influence of emotions in less controlled conditions (i.e., non-randomized live classroom observations conducted by untrained school administrators or professional evaluators) may be more significant. In support of this notion, recent research suggests that live (non-randomized) versus videotaped (randomized) classroom observation ratings produce greater rater trends (Casabianca et al., 2013), and school administrators render more favorable observation ratings than independent raters (Kane & Cantrell, 2013). Whether the presence of these factors translates into greater emotional bias in rater scoring requires further study.

It is also important in future work to examine the role of trait emotions in rating behavior, as state emotion and trait emotion are correlated (see Watson et al., 1988), and the stability of trait emotion makes it likely to produce systematic biases in judgment when randomization of raters to observations is unfeasible. Furthermore, an investigation of the role of other possible between-rater differences, such as emotional intelligence or personality, in observational assessments is needed given the notable amount of variance they contribute to overall classroom quality ratings (i.e., 4% and 8% in Studies 1 and 2; see also Casabianca et al., 2015).

Conclusion

Our findings suggest that rater state emotion may play at least a small role in observations of classroom processes, even when they are conducted by trained and certified raters using a validated observation tool under near ideal assessment conditions. As classroom observations become an increasingly common method for high-stakes assessments of educational performance and school interventions, elucidating the prevalence and sources of observation biases under varying rating conditions is important. Tracking and minimizing rater biases will increase the quality of classroom observations. This will, in turn, improve the value of research in the field and the information decision-makers employ to enhance learning and education.

Appendix

Conversion Equations 6 and 7 specified in Rosnow, Rosenthal, and Rubin (2000), which were used to calculate our MLM-based effect sizes:

Equation 6: To obtain Hedge's g from the MLM-generated t

$$g = \frac{2t}{\sqrt{N'}}.$$

Equation 7: To obtain Cohen's d from Hedge's g

$$d = g\sqrt{\frac{N}{df_{\text{within}}}}.$$

Acknowledgments

The authors express their appreciation to Nicole Elbertson, Maria Reyes, Michelle Bertoli, and Mark White.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The University of British Columbia Doctoral Affiliated Fellowship supported this research. The William T. Grant Foundation (Grant 8364 and 180276) and the NoVo Foundation supported this research.

Note

1. Fifth-grade and sixth-grade classrooms were assessed using the Classroom Assessment Scoring System (CLASS) K-3 because the CLASS Upper Elementary was not yet available during data collection.

References

Baker, B. D., Oluwole, J., & Green, P. C., III. (2013). The legal consequences of mandating high-stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archives*, *21*, 1-71. doi:10.14507/epaa.v21n5.2013

Baron, R. A. (1987). Interviewer's moods and reactions to job applicants: The influence of affective states on applied social judgments. *Journal of Applied Social Psychology*, *17*, 911-926. doi:10.1111/j.1559-1816.1987.tb00298.x

Baron, R. A. (1993). Interviewers' moods and evaluations of job applicants: The role of applicant qualifications. *Journal of Applied Social Psychology*, *23*, 253-271. doi:10.1111/j.1559-1816.1993.tb01086.x

Brackett, M. A., Floman, J. L., Ashton-James, C., Cherkasskiy, L., & Salovey, P. (2013). The influence of teacher emotion on grading practices: A preliminary look at the evaluation of student writing. *Teachers and Teaching: Theory and Practice*, *19*, 634-646. doi:10.1080/13540602.2013.827453

Brennan, R. L. (2001). *Generalizability theory: Statistics for social science and public policy*. New York, NY: Springer-Verlag.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*, 311-337. doi:10.1177/0013164414539163

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *73*, 757-783. doi:10.1177/0013164413486987

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159. doi:10.1037/0033-2909.112.1.155

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, *43*, 245-265. doi:10.1348/0144665031752934

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: John Wiley.

Diener, E., Kanazawa, S., Suh, E. M., & Oishi, S. (2014). Why people are in a generally good mood? *Personality and Social Psychology Review*, *4*, 1-22. doi:10.1177/1088868314544467

Englich, B., & Soder, K. (2009). Moody experts—How mood and expertise influence judgmental anchoring. *Judgment and Decision Making*, *4*, 41-50.

Forgas, J. P. (2014). On the regulatory functions of mood: Affective influences on memory, judgments and behavior. In J. P. Forgas & E. Harmon-Jones (Eds.), *Motivation and its regulation: The control within* (pp. 169-192). Sussex, UK: Psychology Press.

Forgas, J. P., & Eich, E. (2012). Affective influences on cognition: Mood congruence, mood dependence, and mood effects on processing strategies. In A. F. Healy & R. W. Proctor (Eds.), *Handbook of psychology: Experimental psychology* (Vol. 4, pp. 61-82). New York, NY: Wiley.

Forgas, J. P., & George, J. M. (2001). Affective influences on judgments and behavior in organizations: An information processing perspective. *Organizational Behavior and Human Decision Processes*, *86*, 3-34. doi:10.1006/obhd.2001.2971

Fromme, K., Corbin, W. R., & Kruse, M. I. (2008). Behavioral risks during the transition from high school to college. *Developmental Psychology*, *44*, 1497-1504. doi:10.1037/a0012614

Greifeneder, R., Bless, H., & Pham, M. T. (2011). When do people rely on affective and cognitive feelings in judgment? A review. *Personality and Social Psychology Review*, *15*, 107-141. doi:10.1177/1088868310367640

Hafen, C. A., Allen, J. P., Mikami, A. Y., Gregory, A., Hamre, B., & Pianta, R. C. (2012). The pivotal role of adolescent autonomy in secondary school classrooms. *Journal of Youth and Adolescence*, *41*, 245-255. doi:10.1007/s10964-011-9739-2

Hagelskamp, C., Brackett, M. A., Rivers, S. E., & Salovey, P. (2013). Improving classroom quality with the RULER Approach to Social and Emotional Learning: Proximal and distal outcomes. *American Journal of Community Psychology*, *51*, 530-543. doi:10.1007/s10464-013-9570-x

Hamre, B. K., Goffin, S. G., & Kraft-Sayre, M. K. (2009). *Classroom Assessment Scoring System implementation guide: Measuring and improving classroom interactions in early childhood settings*. Charlottesville, VA: Teachstone.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., & Hamagami, A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal*, *113*, 461-487. doi:10.1086/669616

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms*. New York, NY: Foundation for Childhood Development. Retrieved from http://fcd-us.org/sites/default/files/BuildingAScienceOfClassroomsPiantaHamre.pdf

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, *33*, 258-270.

Isbell, L. M., & Lair, E. C. (2013). Moods, emotions, and evaluations as information. In D. Carlston (Ed.), *The Oxford handbook of social cognition* (pp. 435-462). New York, NY: Oxford University Press.

Kane, T. J., & Cantrell, S. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

McFarland, C., White, K., & Newth, S. (2003). Mood acknowledgment and correction for the mood-congruency bias in social judgment. *Journal of Experimental Social Psychology*, *39*, 483-491. doi:10.1016/S0022-1031(03)00025-8

Oishi, S., Lun, J., & Sherman, G. D. (2007). Residential mobility, self-concept, and positive affect in social interactions. *Journal of Personality and Social Psychology*, *93*, 131-141. doi:10.1037/0022-3514.93.1.131

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528-530. doi:10.1177/1745691612465253

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*, 85-112. doi:10.1016/j.jsp.2009.09.002aco

Pianta, R. C. (2012). *Implementing observation protocols: Lessons for K-12 education from the field of early childhood*. Washington, DC: Center for American Progress. Retrieved from http://cdn.americanprogress.org/wp-content/uploads/issues/2012/05/pdf/observation_protocols.pdf

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*, 109-119. doi:10.3102/0013189X09332374

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System: Manual, K-3*. Baltimore, MD: Paul H. Brookes.

Redelmeier, D. A., & Baxter, S. D. (2009). Rainy weather and medical school admission interviews. *Canadian Medical Association Journal*, *181*, 933. doi:10.1503/cmaj.091546

Roberts, J. K., & Monaco, J. P. (2006, April). *Effect size measures for the two-level linear multilevel model*. Paper presented at the annual conference of the American Educational Research Association, San Francisco, CA.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. New York, NY: Cambridge University Press.

Rudasill, K. M., Gallagher, K. C., & White, J. M. (2010). Temperamental attention and activity, classroom emotional support, and academic achievement in third grade. *Journal of School Psychology*, *48*, 113-134. doi:10.1016/j.jsp.2009.11.002

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*, 513-523. doi:10.1037//0022-3514.45.3.513

Schwarz, N., & Clore, G. L. (2003). Mood as information: 20 years later. *Psychological Inquiry*, *14*, 296-303. doi:10.1080/1047840X.2003.9682896

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063-1070. doi:10.1037/0022-3514.54.6.1063

Watson, D., & Walker, L. M. (1996). The long-term stability and predictive validity of trait measures of affect. *Journal of Personality and Social Psychology*, *70*, 567-577. doi:10.1037/0022-3514.70.3.567

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: The Brookings Institute. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2014/05/13%20teacher%20evaluation/