Prev Sci (2014) 15:146–155
DOI 10.1007/s11121-012-0357-3

# Improving the Power of an Efficacy Study of a Social and Emotional Learning Program: Application of Generalizability Theory to the Measurement of Classroom-Level Outcomes

**Andrew J. Mashburn · Jason T. Downer · Susan E. Rivers · Marc A. Brackett · Andres Martinez**

Published online: 14 February 2013


**Abstract** Social and emotional learning programs are designed to improve the quality of social interactions in schools and classrooms in order to positively affect students' social, emotional, and academic development. The statistical power of group randomized trials to detect effects of social and emotional learning programs and other preventive interventions on setting-level outcomes is influenced by the reliability of the outcome measure. In this paper, we apply generalizability theory to an observational measure of the quality of classroom interactions that is an outcome in a study of the efficacy of a social and emotional learning program called The Recognizing, Understanding, Labeling, Expressing, and Regulating emotions Approach. We estimate multiple sources of error variance in the setting-level outcome and identify observation procedures to use in the efficacy study that most efficiently reduce these sources of error. We then discuss the implications of using different observation procedures on both the statistical power and the monetary costs of conducting the efficacy study.

**Keywords** Generalizability theory · Setting-level outcomes · Interventions · Social and emotional learning

A. J. Mashburn (✉)
Psychology Department, Portland State University, P.O. Box 751, Portland, OR 97207, USA
e-mail: mashburn@pdx.edu

J. T. Downer
University of Virginia, Charlottesville, USA

S. E. Rivers · M. A. Brackett
Yale University, New Haven, USA

A. Martinez
University of Michigan, Ann Arbor, USA

Among the persistent problems facing US schools are bullying, school violence, expulsions, dropout, and student disengagement. School-based programming, such as universal social and emotional learning programs, have been developed and implemented to reduce these problems and promote positive youth outcomes by improving the quality of students' experiences in school (Greenberg et al. 2003). Social and emotional learning programs typically involve school-wide activities to create more supportive classroom and school settings, as well as classroom activities that enhance children's abilities to recognize and manage emotions, solve problems, appreciate others' perspectives, and develop interpersonal skills (Zins et al. 2004). Recent studies find that social and emotional learning programs have positive effects on the quality of interactions between teachers and students in classroom settings (Brown et al. 2010) and on student achievement (Durlak et al. 2011).

Group randomized trials involving random assignment of settings (e.g., schools, neighborhoods) to intervention and comparison conditions are commonly used in prevention research to estimate the effects of setting-level interventions on setting-level and individual-level outcomes (Bloom 2005; Murray 1998). Within this framework, setting-level measures are hypothesized to be outcomes that are directly affected by the intervention and mediators through which the intervention produces effects on individuals within settings. The statistical power of a group randomized trial to detect an intervention's direct effects on social settings and indirect effects on individuals within the settings depends, in part, upon the reliability of the setting-level measure (Raudenbush and Sadoff 2008). Generalizability theory provides a statistical framework to estimate multiple sources of error variance present in setting-level measures and to use these variance estimates to make decisions about how to

collect data in ways that reduce error, improve the reliability of the measure, and increase the statistical power of a group randomized trial.

The purpose of this paper is to demonstrate the application of generalizability theory to an observational measure of the quality of interactions in classrooms that serves as a setting-level outcome in an efficacy study of The Recognizing, Understanding, Labeling, Expressing, and Regulating emotions (RULER) Approach (Rivers and Brackett 2011). Specifically, we estimate multiple sources of variance in the Classroom Assessment Scoring System (CLASS; Pianta et al. 2008), a measure of the quality of teacher–student interactions. We then use these variance estimates to make decisions about data collection procedures in the subsequent efficacy study that reduce the sources of error and improve the measure's reliability. Finally, we illustrate the consequences of these decisions on the statistical power of the study to detect the effects of RULER on improving classroom interactions and on the monetary costs of conducting the study.

## Reliability of Setting-Level Outcome Measures

The effects of setting-level preventive interventions on individuals are achieved through a series of mediating core elements, central to which is improving the qualities of social settings (e.g., Judd and Kenny 1981; MacKinnon 1994; Snyder et al. 2006). In the case of RULER, teachers participate in professional development training to implement a yearlong curriculum involving weekly lessons that give students opportunities for practicing and honing the skills of emotional literacy in five key areas—recognizing, understanding, labeling, expressing, and regulating emotion (Rivers and Brackett 2011). These lessons are hypothesized to improve the classroom milieu and the ways that teachers and students interact with one another. The improved classroom context, in turn, is hypothesized to improve student outcomes (i.e., academic and social competence, emotional literacy). Thus, in a test of the efficacy of RULER, a measure of the quality of classroom interactions serves as the setting-level outcome and the proposed mediator through which the program impacts student outcomes.

The reliability of the measure of the quality of classroom interactions influences the statistical power of a group randomized trial to test the effects of RULER in the following way (Raudenbush and Sadoff 2008). As reliability of the measure decreases, the standard error of the mean for the measure of the quality of interactions in RULER and control classrooms increases. As a result of the decreased precision of the estimated mean for each study condition, there is an increased likelihood of making a type II error and falsely concluding that there is not a statistically significant difference in the quality of classroom interactions between the RULER and control classrooms. A consequence of the potentially false conclusion that RULER does not directly impact the quality of classroom interactions is that this measure cannot serve as a potential mediator that explains the mechanisms through which RULER ultimately influences student outcomes. Therefore, when conducting group randomized trials to evaluate the efficacy of setting-level interventions, understanding sources of error in the setting-level outcome measure and adopting data collection procedures that reduce these sources of error can improve the reliability of the measure, thereby increasing the precision of the estimates of the intervention's direct effects on settings and mediated effects on individuals.

There are a number of methods (e.g., interviews, observation) that can be used to assess setting-level outcomes, and in selecting measures for a group randomized trial, consideration must be given to the tradeoffs between validity, reliability, and affordability (Snyder et al. 2006). Observation is an appealing method because it is naturalistic and authentic. However, there are a number of sources of error that may be present in an observed score attributable to characteristics of the rater, stable and temporary characteristics of the setting, and characteristics of the interchanges between individuals and the setting (Cairns and Green 1979). Traditional approaches for evaluating the reliability of setting-level measures come from classical test theory that considers one of these potential sources of error variance at a time. For example, indices of inter-rater reliability, such as kappas and intraclass correlations, assess the magnitude of rater error and test–retest reliability statistics evaluate the stability of scores over time. However, observational measures of social settings comprise multiple sources of error variance, as well as potential interactions between some sources of error variance (Cairns and Green 1979; Cronbach et al. 1972).

Generalizability theory, introduced by Cronbach et al. (1972), is a statistical framework for simultaneously estimating the multiple sources of variance present in an observed score. Generalizability theory was originally applied to evaluate and improve the reliability of person-level measures, such as achievement tests and performance-based assessments (Cronbach et al. 1972). It has since been used to understand sources of variance and improve the reliability of observational assessments of individuals' behaviors (e.g., Hintze 2005; Marcoulides 1989; Suen and Ary 1989) and assessments of the qualities of classroom settings (e.g., Erlich and Borich 1979; Meyer et al. 2011). Application of generalizability theory occurs in two stages (e.g., Shavelson and Webb 1991; Suen 1990). First, a generalizability study is conducted to obtain data and estimate sources of observed score variance, including variance attributed to the object of measurement and variance attributed to multiple sources independent of the object of measurement that are considered sources of error in the measure. These variance

components are applied in the second stage, the decision study, to estimate generalizability coefficients—an index of reliability—for the measure using the procedures adopted in generalizability study and under a hypothetical set of procedures. For example, results of the decision study can be used to compare generalizability coefficients for a measure using different data collection procedures (e.g., number of occasions, number of raters), which can inform subsequent decisions about collecting data in ways that most efficiently reduce error.

In the efficacy study of RULER, the quality of classroom interactions are assessed using the CLASS (Pianta et al. 2008), an observational measure of the quality of emotional support, classroom organization, and instructional support. Data collection procedures for the CLASS in the efficacy study of RULER involve rating the quality of classroom interactions on multiple occasions within a day, for multiple days within each classroom, and by multiple raters of each occasion. The resulting observed score that is intended to represent the quality of classroom interactions has a complex variance structure comprising multiple sources of variability related to classrooms, days within classrooms, occasions within days, raters, as well as some interactions between these variances.

Specifically, classroom quality is the object of measurement with the CLASS, and the variability between classrooms represents universe score variance $\left(\sigma^2_{c}\right)$, which reflects the construct that the measure is intended to assess. There are additional sources of variance that are independent of the universe score that represent error in the measurement of classroom quality. For example, we assume each day that a classroom was observed is a unique occurrence to that classroom; therefore, days are nested within classrooms and variability across days within classrooms $\left(\sigma^2_{d:c}\right)$ is a source of error in the measurement of the classroom. Similarly, observations collected multiple times within a day are unique occurrences to that given day and classroom. Thus, occasions are nested within days and classrooms, and the variability across each occasion $\left(\sigma^2_{o:d:c}\right)$ represents an additional source of error variance in the measurement of the quality of the classroom. There may also be error variability attributed to raters $\left(\sigma^2_{r}\right)$, whereby some raters are more lenient than others and assign scores that are higher, on average, than those scores of more harsh raters.

In addition to variance components associated with the main effects of classroom, day within classroom, occasion within day within classroom, and rater, there are interactions among the crossed, non-nested, sources of variance. For example, between-rater variability may be more pronounced in some classrooms ($\sigma^2_{cr}$, classroom-by-rater interaction), on some days within classrooms ($\sigma^2_{r(d:c)}$, rater-by-day interaction), and on some occasions within days within classrooms ($\sigma^2_{r(o:d:c)}$, rater-by-occasion interaction). Table 1 describes the theoretical sources of variance that are present in observational ratings using the CLASS for this generalizability study design.

## Conducting the Generalizability Study

To test the effects of RULER on the quality of teacher–student interactions and students' academic and social competence and emotional literacy skills, a multisite group randomized trial was conducted within one school district in Brooklyn and Queens, New York. Sixty-four schools volunteered to participate and were randomly assigned to the RULER or control group (32 schools per condition). Prior to the initiation of the group randomized trial (i.e., at baseline, prior to randomization), videotaped observations of fifth and sixth grade classrooms within these schools were collected and rated using the CLASS. These ratings served as the baseline assessment of the quality of classroom interactions for the efficacy study, and they were used for the generalizability study to generate estimates of sources of variance in measures of emotional support, classroom organization, and instructional support. The next subsections describe the procedures, measure, analysis, and results from the generalizability study.

**Table 1** Theoretical sources of variance in observational ratings using the CLASS

| Source of variability | Variance | Definition |
|---|---|---|
| Classroom | $\sigma^2_{c}$ | The universe score variance that reflects average differences across classrooms in ratings of the quality of interactions. |
| Rater | $\sigma^2_{r}$ | Variance that arises because ratings of the quality of interactions vary on average across raters. |
| Day | $\sigma^2_{d:c}$ | Variance that arises because ratings of the quality of interactions vary on average from day-to-day within a classroom. |
| Occasion | $\sigma^2_{o:c:d}$ | Variance that arises because ratings of the quality of interactions vary on average from one occasion to the next within a day. |
| Classroom × rater | $\sigma_{cr}$ | Variance that arises because the between-classroom difference in ratings of the quality of interactions varies from rater-to-rater. |
| Rater × day | $\sigma^2_{r(d:c)}$ | Variance that arises because the between-rater difference in ratings of the quality of interactions varies from day-to-day. |
| Rater × occasion | $\sigma^2_{r(o:c:d)}$ | Variance that arises because the between-rater difference in ratings of the quality of interactions varies from one occasion to the next within a day. |

## Procedures

Principals attending a regularly scheduled meeting with the district superintendent volunteered their fifth and sixth grade English language arts classrooms to participate in the group randomized trial (94 % of the principals volunteered). School sizes ranged from 178 to 656 students ($M$=325.9, SD=97.1), with a student-to-teacher ratio ranging from 11.0 to 25.1 to 1 ($M$=24.5, SD=3.74). Across schools, between 5 and 100 % of students were minorities ($M$=67 %, SD=32 %) and between 0 and 95 % received free/reduced lunch ($M$=23 %, SD=32 %). On average, the fifth and sixth grade teachers had taught for 14.4 years and had worked at their current school for 10.7 years. Forty percent of teachers had worked toward or completed a master's degree, 36 % had earned BAs, and 24 % did not specify. Schools had between two and four fifth and sixth grade English language arts classrooms ($M$=2.67, SD=0.84).

Across all participating schools, there were 170 fifth or sixth grade English language arts classrooms of which 155 were eligible to participate (some teachers taught three or more English language arts classes and, to reduce their burden, they were asked to provide data on two of their classrooms, one randomly selected from each grade level). Of those 155 eligible classrooms, 96 provided videotapes (62 % response rate) for the baseline assessments. There were no statistically significant demographic differences between teachers who returned videotapes and those who did not, but teachers who returned videotapes were more likely to be from schools with a lower percentage of English language learners ($p$<0.05). Each videotape included a recording of a 30-min lesson involving the same teacher and students on a single day. Research assistants converted each videotaped lesson into two separate 15-min occasions, which represented either the first or the second half of the class period. Nearly every occasion was 15 min in length ($M$=14.8 min), although rare, shorter occasions were considered viable and included if they were a minimum of 8 min in length. Ninety classrooms had videotaped lessons for 3 separate days (yielding 6 occasions), 3 classrooms had videotaped lessons for 2 separate days (4 occasions), and 3 classrooms had a videotaped lesson for only 1 day (2 occasions).

Twenty-six raters were trained to use the CLASS and all passed an initial reliability test as well as weekly reliability checks. Multiple raters were randomly assigned to observe and rate each occasion, and 92 % of the occasions were rated by three or more raters. Occasions were randomly distributed across the coding period, and lists of the occasion assignments were reviewed to ensure that a rater did not observe and rate occasions from the same classroom during a single coding session. Raters watched no more than two occasions from different classrooms per coding session, and they took a minimum of a 1-h break between coding sessions to reduce fatigue. This procedure generated a data set comprising multiple-coded occasions, days, and classrooms to allow for estimates of the sources of variance identified in Table 1. Upon completing an observation of each occasion, raters judged whether the camera was stationary, the occasion was continuous, the teacher was audible and visible, and the students were audible and visible, and they also provided an overall codability score for the occasion. From among 535 total occasions observed, 8 % were judged to be uncodable by at least 1 rater, and all ratings of that occasion were excluded. This resulted in 495 unique occasions, with an average of 5.10 occasions per classroom and 1.83 occasions per day within each classroom.

## Measure

The CLASS (Pianta et al. 2008) is an observational measure of the quality of teacher–student interactions in classroom settings. CLASS raters watch live or videotaped classroom interactions for a specified amount of time (typically 15 to 30 min), and then rate the quality of interactions along ten dimensions using a seven-point scale. For each dimension, specific behavioral indicators are provided as descriptive guidelines for "low" scores (ratings of 1 or 2), "mid" scores (ratings of 3, 4, or 5), and "high" scores (ratings of 6 or 7). Scores on the CLASS dimensions are averaged to create three domain scores in the following way: emotional support comprises four dimensions—positive climate, negative climate (reversed), teacher sensitivity, and regard for student perspectives; classroom organization comprises three dimensions—behavior management, productivity, and instructional learning formats; and instructional supports comprise three dimensions—concept development, quality of feedback, and language modeling. Inter-rater agreement (ratings within one point on the seven-point rating scale) for these dimensions in other studies range from 0.72 to 0.89 (Brown et al. 2010). The CLASS is predictive of students' development of academic (e.g., Mashburn et al. 2008) and social skills (e.g., Pianta et al. 2008), and it is sensitive to detecting intervention effects (e.g., Brown et al. 2010).

## Analysis

From the data collected during the generalizability study, estimates were computed for seven variance components that are expressed in Eq. 1:

$$Y_{r(o:c:d)} = \mu + \alpha_c + \beta_r + \gamma_{d:c} + \pi_{o:d:c} + (\alpha\beta)_{cr} + (\alpha\gamma)_{r(d:c)} + (\beta\pi)_{r(o:d:c)} \quad (1)$$

where $\mu$ is the observed overall mean and $\alpha_c$ through $(\beta\pi)_{r(o:d:c)}$ are mutually independent random variables with

means of zero and variances denoted as $\sigma^2_c$, $\sigma^2_r$, $\sigma^2_{d:c}$, $\sigma^2_{o:d:c}$, $\sigma^2_{cr}$, $\sigma^2_{r(d:c)}$, and $\sigma^2_{r(o:d:c)}$, respectively. Programs in SPSS and SAS were used to estimate these variance components and are available by writing to the first author.

## Results

Table 2 presents the estimates for seven sources of variance in observational ratings of emotional support, classroom organization, and instructional support and the number of measurement occasions available to compute each of these sources of variance. Results indicate that the percentage of the total variance that is attributable to the between-classroom portion $\left(\sigma^2_c\right)$ ranges from 23 to 31 %; this is the variance that reflects the construct that the CLASS is intended to assess. Across each of the three CLASS domains, a relatively small portion of the total variance (3–11 %) was attributable to differences in ratings from 1 day to the next within a classroom $\left(\sigma^2_{d:c}\right)$. The classroom-by-rater variance $\left(\sigma^2_{cr}\right)$ was also small (1–6 %), indicating that between-classroom variability was relatively consistent across raters. Similarly, rater-by-day variance $\left(\sigma^2_{r(d:c)}\right)$ was also very small, indicating that the magnitude of the between-rater variability did not vary from 1 day to the next.

Average ratings varied considerably across the different raters, and this between-rater variance $\left(\sigma^2_r\right)$ was substantial for instructional support (18 %) and emotional support (14 %) and relatively small for classroom organization (4 %). Another large source of error was the variability attributed to differences in scores from one occasion to the next within a day $\left(\sigma^2_{o:d:c}\right)$, reflecting fluctuations in the quality of classroom interactions that occur from one 15-min occasion to the next. The single greatest source of variance for all three CLASS domains (between 27 and 33 % of the total variance in observed scores) was rater-by-occasion variance $\left(\sigma^2_{r(o:d:c)}\right)$. This source of error variance comes from the interaction between rater variance and occasion variance, and it refers to inconsistencies between raters from one occasion to the next within a given day. In other words, raters had stronger agreement for one occasion within a day than for the other occasion within the same day.

One possible explanation for this source of error is that the different types of activities that occurred in the classroom during each occasion gave rise to more or less consistency in raters' perceptions of the quality of interactions. For example, on one occasion within a day, there may be laughter, smiling, and unambiguous displays of sensitivity by the teacher toward students, and these clear displays of emotional support are likely to generate high levels of rater agreement for this occasion. In contrast, the second occasion within the same day may involve activities that have few interactions between students and teachers, such as independent work by students at their desks, resulting in few clear indicators that align with the items' descriptions of an emotionally supportive classroom. As a result, raters have less objective information to make a rating and must use more subtle events to make inferences about the level of emotional support during this occasion, which is likely to produce disagreement between raters for this occasion. Thus, the interaction between the rater and the occasion—likely caused by differing events across the occasions that affect the consistency in raters' scores—is the greatest source of error in the measurement of classroom quality.

## Conducting the Decision Study

Using the variance estimates from Table 2, we conducted a decision study to identify procedures for observing classrooms in the efficacy study of RULER that maximize the reliability of the CLASS. More specifically, the variance estimates from the generalizability study were used to compute a generalizability coefficient, λ, an index of reliability

**Table 2** Estimates of seven sources of variance in observational ratings of emotional support, classroom organization, and instructional support

| | | | Emotional support | | Classroom organization | | Instructional support | |
|---|---|---|---|---|---|---|---|---|
| Source of variability | | *n* occasions | Variance | Percent total | Variance | Percent total | Variance | Percent total |
| Classroom | $\sigma^2_c$ | 96 | 0.23 | 28 | 0.24 | 31 | 0.31 | 23 |
| Rater | $\sigma^2_r$ | 26 | 0.11 | 14 | 0.04 | 4 | 0.25 | 18 |
| Day | $\sigma^2_{d:c}$ | 3 | 0.03 | 3 | 0.08 | 11 | 0.07 | 5 |
| Occasion | $\sigma^2_{o:d:c}$ | 2 | 0.18 | 22 | 0.13 | 17 | 0.26 | 19 |
| Classroom × rater | $\sigma^2_{cr}$ | 293 | 0.05 | 6 | 0.04 | 5 | 0.02 | 1 |
| Rater × day | $\sigma^2_{r(d:c)}$ | 268 | 0.00 | 0 | 0.00 | 0 | 0.01 | 1 |
| Rater × occasion | $\sigma^2_{r(o:d:c)}$ | 99 | 0.22 | 27 | 0.25 | 32 | 0.46 | 33 |
| Total | | | 0.81 | | 0.77 | | 1.38 | |

Estimates were obtained using the VARCOMP procedure with Restricted Maximum Likelihood in SAS and SPSS

that is the ratio of universe score variance $\left(\sigma_c^2\right)$ to the total variance, excluding those sources that do not affect the object of measurement. This statistic is appropriate when the decisions that are made from the decision study concern the relative standing of individuals, in this case, classrooms. This is in contrast to absolute decisions that can be made in a decision study that concern the absolute levels of performance, such as whether the classroom achieved a specific level of quality above or below a defined threshold (Brennan and Kane 1977). In addition to the between-classroom variability, the sources of variance that influence the relative standing of classrooms to one another in the efficacy study of RULER and are relevant sources of error in the computation of λ are $\sigma^2_{d:c}$, $\sigma^2_{o:d:c}$, $\sigma^2_{cr}$, $\sigma^2_{r(d:c)}$, and $\sigma^2_{r(o:d:c)}$. Each of these sources of variance is either nested within classrooms ($\sigma^2_{d:c}$ and $\sigma^2_{o:d:c}$) or involves an interaction between raters and classroom $\left(\sigma^2_{cr}\right)$ or raters and a facet nested within classroom ($\sigma^2_{r(d:c)}$ and $\sigma^2_{r(o:d:c)}$).

$$\lambda = \frac{\sigma_c^2}{\sigma_c^2 + \frac{\sigma^2_{d:c}}{D_c} + \frac{\sigma^2_{o:d:c}}{O_d D_c} + \frac{\sigma^2_{cr}}{R_o} + \frac{\sigma^2_{r(d:c)}}{R_o D_c} + \frac{\sigma^2_{r(o:d:c)}}{R_o O_d D_c}} \quad (2)$$

As Eq. 2 illustrates, in the denominator, each of the sources of error variance that affects the relative standing of classrooms is divided by the number of units used to compute the variance estimate. For example, the classroom-by-rater variance $\left(\sigma^2_{cr}\right)$, rater-by-day variance $\left(\sigma^2_{r(d:c)}\right)$, and rater-by-occasion variance $\left(\sigma^2_{r(o:d:c)}\right)$ are each divided by the average number of raters who observe each unique occasion ($R_o$). Thus, increasing the number of raters per occasion reduces these sources of error variance, which in turn, decreases the total variability in the denominator and increases the generalizability coefficient for the measure. Similarly, all sources of error variance, except for classroom-by-rater $\left(\sigma^2_{cr}\right)$, are reduced by increasing the number of days each classroom is observed ($D_c$). Further, the between-occasion variance $\left(\sigma^2_{o:d:c}\right)$ and the rater-by-occasion variance $\left(\sigma^2_{r(o:d:c)}\right)$ are also reduced by increasing the number of occasions that are observed during each day that a classroom is observed ($O_d$). In sum, based on Eq. 2, the reliability of the observed score can be improved by manipulating the observation procedures in three ways: (1) increasing the number of raters who observe each unique occasion ($R_o$), (2) increasing the number of days each classroom is observed ($D_c$), and (3) increasing the number of occasions that ratings are made on each day in each classroom ($O_d$).

For the purposes of informing decisions about observation procedures to use in the RULER efficacy study, we set the number of occasions that are observed within each day to two 15-min occasions, which was done in consideration of the amount of time that teachers are teaching English language arts lessons during a given day. Under this constraint, we computed the generalizability coefficient for different numbers of days each classroom was observed (1, 2, and 4) and different numbers of unique ratings made for each occasion (1, 2, and 4). Table 3 presents the generalizability coefficient for emotional support, classroom organization, and instructional support for nine possible scenarios in which the observations can be conducted. It is important to note that the interpretation of these generalizability coefficients is different than the traditional standards for evaluating reliability. Coefficients from reliability statistics common in classical test theory consider a single source of error variance at a time that is independent of the true score. The coefficient λ in this case considers five sources of variance to be error. Thus, generalizability coefficients are generally lower than traditional reliability coefficients, and the coefficients in Table 3 may be compared to each other, but they should not be compared to the standards that are typical for other measures of reliability.

Results from Table 3 help inform decisions about how to conduct classroom observations in ways that maximize the generalizability coefficient for each of the three CLASS domains. For example, under the nine scenarios, scenarios 3, 5, and 7 require roughly the same amount of resources for conducting observations. Specifically, each scenario requires that eight observations are conducted within each classroom (scenario 3 involves 4 days with two occasions per day that are rated by one rater; scenario 5 involves 2 days with two occasions per day that are each rated by two raters; scenario 7 involves 1 day with two occasions per day that are each rated by four raters). Despite equal resources needed to observe under these three scenarios, the decision study indicates that scenarios 3 and 5 produce substantially higher reliability estimates than scenario 7. Thus, with a fixed budget that can support eight observations for each classroom, assigning four raters to observe two occasions on 1 day for each classroom is the least optimal way to utilize resources for conducting classroom observations.

Scenarios 1 and 9 represent the least and most resource-intensive procedures, respectively. Scenario 1 involves a single day of observing each classroom that results in two occasions, each of which is observed by a single rater for a total of two ratings per classroom. Scenario 9 involves 4 days of observing each classroom that results in eight observation occasions, each of which is rated independently by four different raters. Observations conducted under scenario 9 would yield a total of 32 ratings for each classroom; however, the tradeoff to the increased resources needed for observing classrooms under the conditions in scenario 9 compared to scenario 1 is the improvement in the reliability of the measure (Table 3). Specifically, for scenario 9, the reliability of the measure of emotional support, classroom organization, and instructional support is 0.84, 0.88, and 0.86, respectively. For scenario 1, the reliability of the

**Table 3** Generalizability coefficient for emotional support, classroom organization, and instructional support for nine different observation scenarios

| | Observation procedures | | | | | |
|---|---|---|---|---|---|---|
| | Number of days/ classroom ($D_c$) | Number of occasions/ day ($O_d$) | Number of raters/ occasion ($R_o$) | Emotional support λ | Classroom organization λ | Instructional support λ |
| Scenario 1 | 1 | 2 | 1 | 0.48 | 0.51 | 0.44 |
| Scenario 2 | 2 | 2 | 1 | 0.60 | 0.65 | 0.60 |
| Scenario 3 | 4 | 2 | 1 | 0.70 | 0.74 | 0.73 |
| Scenario 4 | 1 | 2 | 2 | 0.57 | 0.62 | 0.54 |
| Scenario 5 | 2 | 2 | 2 | 0.70 | 0.75 | 0.70 |
| Scenario 6 | 4 | 2 | 2 | 0.79 | 0.83 | 0.81 |
| Scenario 7 | 1 | 2 | 4 | 0.64 | 0.69 | 0.61 |
| Scenario 8 | 2 | 2 | 4 | 0.76 | 0.81 | 0.76 |
| Scenario 9 | 4 | 2 | 4 | 0.84 | 0.88 | 0.86 |

measure of emotional support, classroom organization, and instructional support is 0.48, 0.51, and 0.44, respectively. In the next section, we demonstrate how the reliability of the CLASS using observation procedures described in scenarios 1 and 9 influences the statistical power of the efficacy study to detect effects of RULER on classroom emotional support.

## Implications of Reliability for the Power of the Efficacy Study of RULER

The statistical power of a group randomized trial to detect effects on setting-level outcomes is influenced by the reliability of the outcome measure (Raudenbush and Sadoff 2008), and power analyses that do not adjust for the unreliability of the outcome measure implicitly assume that the outcome has been assessed without error. However, as we have shown, this is an untenable assumption in the context of the efficacy study of RULER. In this section, we account for the reliability of the outcome measure in determining the power of the efficacy study to detect moderate-sized differences in quality of interactions between classrooms assigned to the RULER and control groups. Specifically, using the Optimal Design Software (Liu et al. 2009), we calculated the power of the efficacy study to detect a statistically significant difference in emotional support between classrooms in RULER and control schools using the outcome reliability of the measure of emotional support in scenario 1 (λ=0.48) and scenario 9 (λ=0.84).

The design of the study is a multisite group randomized trial with a group-level outcome, treatment is randomly assigned at the school level, and the setting-level outcome is assessed at the classroom level. The power calculation was conducted under the following conditions: 64 schools ($K$), 3 classrooms per school ($J$), an estimate of 30 % of the true outcome variance explained at the school level, α=0.05, an estimate of the intraclass correlation for the between-school variance in the outcome ($\rho$)=0.10, and an expected effect size difference in emotional support between RULER and control classrooms ($\delta$)=0.50. This moderate-sized effect is a reasonable expectation for the magnitude of setting-based interventions that are intended to improve setting-level outcomes (e.g., Brown et al. 2010).

Figure 1 plots the association between the reliability of the classroom-level measure of emotional support and the statistical power of the efficacy study of RULER. There is a nonlinear association between reliability and statistical power, such that, when reliability is low (e.g., between 0.0 and 0.4), increases in reliability result in large improvements in power. However, when reliability is relatively high (between 0.7 and 1.0), increases in reliability result in relatively smaller improvements to the power of the study. Figure 1 also illustrates that, in the case of the efficacy study of RULER, when emotional support is measured with a reliability of 0.48 as in the case of scenario 1, the statistical power of the study to detect a difference between RULER and control conditions is 0.65. When the outcome is measured with a reliability of 0.84 as in the case of scenario 9, the statistical power of the study is 0.85. Figure 2 further illustrates the implications of low reliability of the setting-level outcome measure on the statistical power of the study by plotting the association between the statistical power and the number of schools in the study when the reliability of the outcome measure is 0.48 as in scenario 1. To achieve a statistical power of 0.85 for the efficacy study of RULER using the observation procedures described by scenario 1 (λ=0.48), the number of schools needed in the study would increase from 64 to 102 (Fig. 2).

The tradeoff between the reliability of the outcome measure and the statistical power of the group randomized trial is also evident when contrasting the three scenarios (3, 5, and 7) that require the same number of observations per classroom but adopt different observation procedures (Table 3). Specifically, the reliability for the measure of emotional support for scenarios 3, 5, and 7 is 0.70, 0.70, and 0.64, respectively. To achieve a

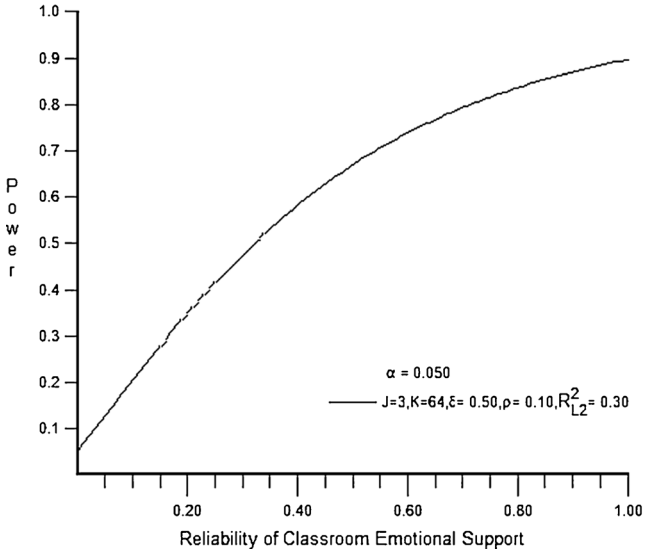


**Fig. 1** Association between the reliability of classroom emotional support and the statistical power of the RULER efficacy study

statistical power of 0.85 for each of these three scenarios, 75 schools are needed in scenarios 3 and 5 and 85 schools are needed in scenario 7.

## Implications for the Costs of Conducting Efficacy Studies

As we have shown, increasing the reliability of the CLASS improves the statistical power of the efficacy study of RULER. In this section, we consider the monetary costs associated with conducting a generalizability study and with adopting the observation procedures identified in the decision study that maximize the reliability of the CLASS. In the RULER study, observing and rating videotapes from classrooms at the baseline phase was part of the design of the efficacy study, and a total of 535 observation occasions were available. Each observation occasion was rated, on average, 3.39 times, resulting in a total of 1,814 ratings. Each rating took approximately 30 min to complete; thus, the number of person-hours required to rate these occasions was 907 h. Raters were paid $16/h, and the estimated additional cost for conducting the generalizability study was $14,512 or approximately $151 for each of the 96 classrooms.

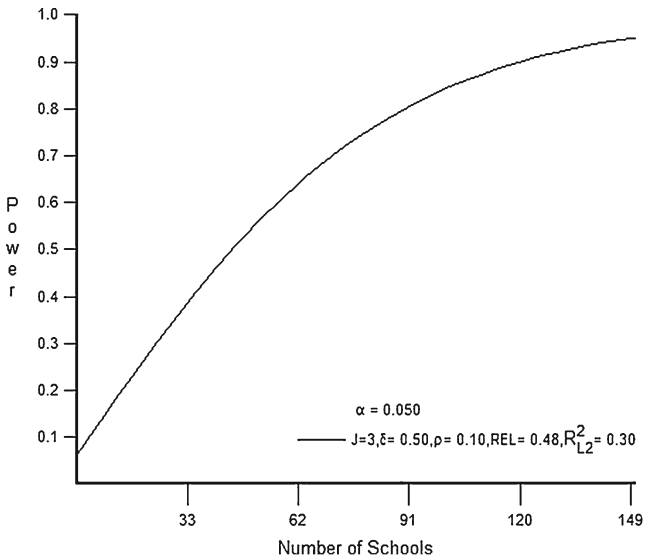


**Fig. 2** Number of schools needed for the RULER efficacy study to achieve statistical power of 0.85 when the reliability of classroom emotional support is 0.48

As described above, the statistical power of the efficacy study of RULER on classroom emotional support is equivalent under the following two conditions: (1) 64 schools with an outcome reliability of 0.84 achieved using observation scenario 9 and (2) 102 schools with an outcome reliability of 0.48 achieved using observation scenario 1. However, there are substantial differences in the costs of conducting the study between these two conditions. Under the first condition involving 64 schools using observation scenario 9, observation cost is estimated to be $49,152 (64 schools, 3 classrooms per school, 4 separate days, 2 occasions per day, and 4 raters per 30-min occasion, who are paid $16/h). Under the second condition involving 102 schools using observation scenario 1, observation cost is estimated to be $4,896 (102 schools, 3 classrooms per school, 1 day, 2 occasions per day, 1 rater per 30-min occasion, who is paid $16/h). Thus, the cost of measuring the setting-level outcome in scenario 1 is $44,256 less than the cost in scenario 9.

The tradeoff to the reduced observation costs in scenario 1 is the additional costs required to increase the number of schools in the study from 64 (32 RULER and 32 control) to 102 (51 RULER and 51 control). For the RULER study, the estimated cost of implementing the workshops and coaching intervention alone is $3,118 per school; thus, increasing the number of schools in the RULER condition by 19 (from 32 to 51) increases the estimated cost of the efficacy study by $59,242, which is approximately 33 % more than the additional costs for conducting the extensive observations of classrooms in the 64 schools using scenario 9. This is a very conservative estimate of the cost savings, given that these figures do not take into account the costs for collection of other data (e.g., student and teacher surveys, archival data), travel, and staffing that would be required to conduct the study with all 38 schools added to the RULER and control groups. The cost savings related to increased reliability of the outcome measure are also evident when comparing scenarios 3, 5, and 7. Although the costs for conducting observations are equal for these three scenarios, scenario 7 requires ten more schools than scenarios 3 and 5 to achieve statistical power equal to 0.85. The addition of five of these schools to the RULER group would increase the cost of implementing RULER in the efficacy study by $15,590.

## Discussion

A common focus of prevention science research is to develop, implement, and test the effects of interventions designed to improve the quality of social settings in order to promote positive outcomes for individuals. As we have demonstrated, generalizability theory provides a statistical framework to identify the sources of error in the setting-level outcome measure, estimate these sources, and make decisions about procedures for collecting data that reduce these sources. Within the context of testing the efficacy of RULER on improving the quality of classroom interactions, decisions about the number of days to observe each classroom and the number of raters who observe and rate each occasion had a substantial effect on the reliability of the outcome measure, the resulting statistical power of the study to detect the intervention's effects, and the costs of conducting the study.

Based on the lessons learned in this demonstration, prevention scientists should consider the following when applying generalizability theory for the purposes of improving the reliability of a setting-level outcome in a group randomized trial. First, careful consideration should be given to specifying the sources of error variance that are salient for a particular measure used in a subsequent efficacy study. For example, in this case, we theoretically defined that the rater-by-occasion variance is a source of error in measuring the quality of classroom interactions. This assumes that emotional support, classroom organization, and instructional support are stable characteristics within an English language arts lesson, and the substantial variability attributable to differences in raters' scores from one occasion to the next within a given lesson represents error in the measurement of these outcomes. An alternative specification is that rater-by-occasion variance is an attribute of the object of measurement itself (i.e., the instability of scores by raters is part of the construct that is being assessed), and this variance should not be conceived as error.

Second, the generalizability study should be conducted using observation procedures and settings that directly align with those planned for the efficacy study. Specifically, as was the case in this study, observations in both the generalizability study and the efficacy study involve the same measure (CLASS), a subsample of the same social settings (fifth and sixth grade classrooms), the same types of activities observed within the settings (English language arts lessons), and the same procedures for training the raters. It is a limitation that 38 % of teachers who participated in the efficacy study did not provide videotapes for the generalizability study. This difference between the sample in the generalizability study and the sample in the efficacy study may affect how well the variance estimates from the generalizability study apply to the efficacy study, as well as the accuracy of the subsequent decisions made about data collection procedures. Future applications of generalizability theory for improving the power of a group randomized trial study should as closely as possible align the observation procedures and sample in the generalizability study with those in the efficacy study to help ensure that the resulting decisions are directly relevant for the efficacy study.

Third, when conducting a generalizability study, it is important to select occasions to observe and rate that will provide estimates of as many sources of variance as possible, given coding resources. The best practice for conducting this generalizability study is to use a fully crossed design in which two or more raters observe and rate multiple occasions within a day for multiple days within each classroom for multiple classrooms. However, there may be limitations to conducting a generalizability study using a fully crossed design, and if this is the case, steps must be taken to identify which estimated sources of variances are confounded by unestimated sources and then make assumptions about the magnitudes of the unestimated sources.

Fourth, the optimum procedures for producing reliable estimates identified in a decision study must be interpreted in light of practical constraints related to observing classrooms during the efficacy study. For example, there are likely to be limits to the number of days that teachers and school administrators will allow observations to be conducted. Therefore, it may place undue burden on teachers and may not be feasible to adopt observation scenarios that involve many days of observing classrooms. It may also not be feasible to conduct observations prior to the efficacy study for the purposes of conducting the generalizability study. There are also constraints related to the number of occasions that can be observed within a given day depending upon limits imposed by teachers and school administrators and whether the object of measurement is a discrete event (an English language arts lesson) or a general assessment of the classroom during a typical day. Further, if live observations are utilized, there are limits to the number of independent ratings that can be made of a specific occasion within a classroom because having more than two raters in a classroom will disrupt the typical classroom processes that are the targets of assessment.

In addition to these lessons learned from this demonstration, we also have identified the following research questions that should be further investigated to improve the procedures of conducting classroom observations, G studies, and D studies. What are the potential tradeoffs to the reliability and validity of ratings collected from videotapes compared to live observations? What is the optimum length of an observation occasion and is 15 min sufficient? To what extent do observational measures of teachers comprise reactivity effects caused by the presence of cameras or raters in the classroom? How well do the variance estimates from the generalizability study taken from a sample of classrooms, days, occasions, and raters per occasion approximate the population variance values for these variance estimates? This last question is particularly relevant for

the rater-by-occasion variance, which in this study was the largest source of variance and was estimated from a relatively small number of occurrences (99) when the same rater provided scores for both occasions within the same classroom on the same day.

In conclusion, as demonstrated in this paper, generalizability theory provides a framework for understanding theoretical sources of error in observational measures of settings, estimating these sources of error, and using estimates to make decisions about the procedures for observing settings in ways that improve the reliability of the outcome measure. To apply this method, careful consideration must be taken in identifying the salient sources of error variance in the setting-level outcome measure, designing a generalizability study that produces accurate estimates of each variance component, and interpreting the results of the decision study in light of practical constraints that may impose limits on the number of days, occasions, and ratings per occasion that can occur. Following these procedures when evaluating effects of interventions to improve social settings has the potential to improve the reliability of the setting-level outcome measure, increase the statistical power of an efficacy study, reduce the costs of conducting an efficacy study, and improve the accuracy of the inferences that can be made about the effects of social and emotional learning and other interventions on improving social settings.

**Acknowledgments** This research was supported by grants from the William T. Grant Foundation (#8364 and #11456). The authors wish to thank J. Patrick Meyer, Howard Bloom, and Steven Raudenbush for their comments about this manuscript.

## References

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277–289. doi:10.1111/j.1745-3984.1977.tb00045.x.

Brown, J. L., Jones, S. M., LaRusso, M., & Aber, J. L. (2010). Improving classroom quality: Teacher influences and experimental impacts of the 4Rs program. *Journal of Educational Psychology, 102*, 153–167. doi:10.1037/a0018160.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments* (pp. 115–172). New York: Sage.

Cairns, R. B., & Green, J. A. (1979). How to assess personality and social patterns: Observations or ratings? In R. B. Cairns (Ed.), *The analysis of social interactions: Methods, issues, and illustrations* (pp. 209–226). Hillsdale: Lawrence Erlbaum.

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: Wiley.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*, 405–432. doi:10.1111/j.1467-8624.2010.01564.x.

Erlich, O., & Borich, C. (1979). Occurrence and generalizability of scores on a classroom interaction instrument. *Journal of Educational Measurement, 16*, 11–18. doi:10.1111/j.1745-3984.1979.tb00081.x.

Greenberg, M. T., Weissberg, R. P., O'Brien, M. U., Zins, J. E., Fredericks, L., Resnik, H., & Elias, M. J. (2003). Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning. *American Psychologist, 58*, 466–474. doi:10.1037/0003-066X.58.6-7.466.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34*, 507–519.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5*, 602–619. doi:10.1177/0193841X8100500502.

Liu, X., Spybrook, J., Congdon, R., Martinez, A., & Raudenbush, S. W. (2009). *Optimal design for longitudinal and multilevel research* (*version 2*) [*computer software*]. Available at http://www.wtgrantfoundation.org.

MacKinnon, D. P. (1994). Analysis of mediating variables in prevention intervention studies. In A. Cazares & L. A. Beatty (Eds.), *Scientific methods for prevention intervention research. NIDA research monograph 139, DHHS pub. 94-3631* (pp. 127–153). Washington, DC: US Department of Health and Human Services.

Marcoulides, G. (1989). The application of generalizability analysis to observational studies. *Quality and Quantity, 23*, 115–127. doi:10.1007/BF00151898.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., et al. (2008). Measures of pre-k quality and children's development of academic, language and social skills. *Child Development, 79*, 732–749. doi:10.1111/j.1467-8624.2008.01154.x.

Meyer, J. P., Henry, A. E., & Mashburn, A. J. (2011). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment Journal, 16*, 227–243. doi:10.1080/10627197.2011.638884.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal, 45*, 365–397. doi:10.3102/0002831207308230.

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS)*. Baltimore: Paul H. Brookes.

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness, 1*, 138–154. doi:10.1080/19345740801982104.

Rivers, S. E., & Brackett, M. A. (2011). Achieving standards in the English language arts (and more) using The RULER Approach to social and emotional learning. *Reading & Writing Quarterly, 27*, 75–100. doi:10.1080/10573569.2011.532715.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park: Sage.

Snyder, J., Reid, J., Stoolmiller, M., Howe, G., Brown, H., Dagne, G., & Cross, W. (2006). The role of behavior observation in measurement systems for randomized prevention trials. *Prevention Science, 7*, 43–56. doi:10.1007/s11121-005-0020-3.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale: Lawrence Erlbaum.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale: Lawrence Erlbaum.

Zins, J. E., Bloodworth, M. R., Weissberg, R. P., & Walberg, H. J. (2004). The scientific base linking social and emotional learning to school success. In J. E. Zins, R. P. Weissberg, M. C. Wang, & H. J. Walberg (Eds.), *Building academic success on social and emotional learning: What does the research say?* (pp. 3–22). New York: Teachers College Press.