Abstract Title Page
Not included in page count.

Title: Using Generalizability Theory to Examine Sources of Variance in Observed Behaviors within High School Classrooms

Authors and Affiliations:
Tashia Abry, Ph.D.
Assistant Research Professor
Arizona State University
T. Denny Sanford School of Social and Family Dynamics
tabry@asu.edu

Anne H. Cash, Ph.D.
Assistant Professor
Johns Hopkins University
School of Education
ahcash@jhu.edu

Catherine P. Bradshaw, Ph.D.
Professor, Associate Dean for Research and Faculty Development
University of Virginia
Curry School of Education
cpb8g@virginia.edu

Abstract Body

Problem / Background / Context:

Description of the problem addressed, prior research, and its intellectual context.

The reliability of an outcome measure directly affects statistical power in both group randomized controlled trials as well as large-scale surveys (Raudenbush & Sadoff, 2008). Generalizability theory (GT; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) offers a useful framework for estimating the reliability of a measure while accounting for multiple sources of error variance (Shavelson, Webb, & Rowley, 1989). Despite a history of use with measures of individual performance, application of GT to observational measures of individuals, classrooms, and schools has increased in recent years as the use of such measures has proliferated (e.g., Mashburn, Downer, Rivers, Brackett, & Martinez, 2013; Matsumura, Garnier, Slater, & Boston, 2008; Raudenbush & Sampson, 2002). Application of GT typically begins with a generalizability study (G-study) and is followed by a decision study (D-study; Cardinet, Tourneur, & Allal, 1976). The G-study is used to estimate variance in an observed score attributable to both the object of measurement (considered the "true score") and elements of the data collection design independent of the object of measurement (termed facets) regarded as measurement error. In turn, these variance components are used in the D-study to estimate generalizability coefficients that serve as an index of the reliability of a measure under the specific conditions in which it was administered. It is in the D-study stage that researchers can also estimate generalizability coefficients under varying hypothetical designs (e.g., altering the number of raters).

Purpose / Objective / Research Question / Focus of Research:

Description of the focus of the research.

The purpose of this study was to use GT to examine multiple sources of variance in and the reliability of school-level teacher and high school student behaviors as observed using the tool, Assessing School Settings: Interactions of Students and Teachers (ASSIST).

Improvement Initiative / Intervention / Program / Practice:

Description of the improvement initiative or related intervention, program, or practice.

The study takes place in the context of a high school randomized controlled trial (Maryland Safe and Supportive Schools [MDS3]) of Positive Behavioral Interventions and Supports (PBIS; Sugai & Horner, 2006); PBIS is a school-based decision-making framework designed to enhance students' social and academic functioning.

Setting:

Description of the research location and partners involved, if applicable.

MDS3 is a 58 high school randomized controlled trials funded by the U.S. Department of Education and the William. T. Grant Foundation. The project was designed collaboratively by the Maryland State Department of Education (MSDE), Johns Hopkins University, and Sheppard Pratt Health System based on the existing collaboration in PBIS Maryland and the lessons learned through PBIS implementation over 11 years.

Population / Participants / Subjects:

Description of the participants in the research: who, how many, key features, or characteristics.

Baseline data (collected prior to intervention) from six Maryland high schools provide the analytic sample for the present inquiry. The participating schools included a diverse student

population with a minority rate of 48% (*SD* = 25.34). The mean student enrollment of these schools was 1,253.81 (*SD* = 476.53). About half of the classroom observations were language arts classrooms (49%), 18% were conducted in math classrooms and 17% were conducted in both history and science classrooms. Approximately 61% of the observed teachers were female.

Research Design:

Description of the research design.

Two raters worked in tandem to observe 25 classrooms/occasions in each of the six high schools using the ASSIST, resulting in a partially crossed design in which observations were nested within schools and crossed by rater. A graphic depiction of the design is shown in Figure 1.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data or use of existing databases.

School interactions were assessed using the ASSIST, an observational measure of teachers' behavior management strategies and student behavior. Ten classroom tally items were selected from the ASSIST for the present analysis. Specifically, five of the selected tallies assessed teachers' classroom management techniques (proactive behavior expectations, reactive behavior management, approval, disapproval, and opportunity to respond); the other five tallies assessed student problem behavior (noncompliance, disruption, verbal aggression, physical aggression, and profanity). All 10 items were scored using a frequency count in which each discrete teacher or student behavior was tallied as one event in one category according to a hierarchical list (e.g., a verbal threat would be coded as verbal aggression but not noncompliance even though the act could be considered noncompliant). The result was a frequency count for each of the 10 categories of teacher and student behavior, ranging from 0 to 54 in the analytic sample. Descriptive statistics for the 10 tally items are reported in Table 1.

G-studies. ASSIST tallies were subject to multiple sources of variability—a focus of the present study—stemming from facets of the data collection approach (e.g., multiple observations and raters) and interactions between these facets. Under the design specifications, school-level ASSIST scores represent the object of measurement. School-level difference is termed universe score variance (σ_s^2), and is the target of measurement. Sources of variance independent of the universe score, then, reflect elements of measurement error. For example, as noted, multiple observations were conducted within a school. As such, variability across within-school observations ($\sigma_{o:s}^2$) constitutes error. Similarly, raters (σ_r^2) could be a source of error if, on average, some raters score behaviors with higher frequency than others. In addition to variance attributable to the main effects of these design facets, variance may also be attributable to the interactions between facets. As an example, variability attributable to a school-by-rater interaction (σ_{sr}^2) would indicate between-school variability was more pronounced for some raters compared to others. The 10 ASSIST classroom tallies are used as unique indicators of teacher and student behavior. That is, frequencies are not typically aggregated across tally categories. Therefore, a separate G-study was conducted for each of the 10 tally items. Table 2 lists the variance components part of this G-study design and their respective interpretations. This design was consistent across the 10 G-studies.

D-studies. Using variance component estimates generated in the G-studies, 10 D-studies were conducted (one for each ASSIST tally) to estimate generalizability coefficients (a reliability index) for the data collection procedure as applied. Generalizability coefficients quantify the ratio of variance attributable to the object of measurement (i.e., universe score variance, σ_s^2) to

total variance, including only the sources of variance that affect the object of measurement. In the present design, this includes any design facet that is either nested within schools or interacts with schools (in effect, excluding only σ^2_r which does not influence the relative standing of schools to one another). Calculating reliability in this way is appropriate for "relative" decisions in which the relative standing of the object of measurement (i.e., schools) is of interest, as is the case here. In addition, these D-studies were used to estimate generalizability coefficients for hypothetical designs in which the number of raters and observations per school were modified. Specifically, variance component estimates and generalizability coefficients were calculated for one and two raters under conditions ranging from five to 40 observations per school. These values were selected based on feasible alternative data collection procedures. That is, it is important to ascertain whether reliable estimates can result when relying on a single coder to conduct an observation, common in many evaluation studies. Similarly, it is helpful to know at what point increases or decreases in the number of observations conducted within a school yield substantially more or less reliable estimates. All G- and D-studies were conducted using GENOVA software (Crick & Brennan, 1983).

Findings / Outcomes:

Description of the main findings or outcomes, with specific details.

Figure 2 shows the percentage of variance attributable to each of the variance components across the eight G-studies in which variance components could be estimated. G-study results for four of the 10 tallies (two teacher and two student items) indicated a discernable amount of school-level variance in the frequency of observed teacher and student classroom behaviors. These tally categories were teacher proactive behavior expectations (6%), teacher reactive behavior management (9%), student disruption (8%), and student profanity (7%). Estimates for four other tally categories, teacher approval, teacher disapproval, teacher opportunity to respond, and student noncompliance, indicated that 1% or less of the total variance was attributable to school-level differences. Variance component estimates for the remaining two tally categories, student verbal and physical aggression could not be computed due to a lack of variability in tally frequencies (i.e., these behaviors were not observed). In five of the eight tally categories (teacher positive behavior expectations, teacher approval, teacher reactive behavior management, teacher opportunity to respond, and student disruption), the largest amount of variance was attributable to observation ($\sigma^2_{o:s}$), ranging from 71% to 88%. This indicates that, for these categories, tally frequencies varied greatly across observations within a school. This pattern was not true for teacher disapproval, student noncompliance, and student profanity in which the largest proportion of variance (88%, 98%, and 81%, respectively) was explained by the interaction between observation and rater ($\sigma^2_{r(o:s)}$), suggesting that for these categories, raters had closer agreement for some observations within a school than others. Notably, variance attributable to rater (σ^2_r) was minimal, explaining no more than 2% of the total variance in seven of the eight tallies. This suggests that, on the whole, raters were in close agreement in the way they scored ASSIST classroom tally items.

As stated, D-studies were used to estimate generalizability coefficients (i.e., reliability estimates) for observed and hypothetical data collection designs. Generalizability coefficients were estimated for designs in which one or two raters conduct between five and 40 observations within a school. Figure 3 presents generalizability coefficients for six of the 10 tally categories in which generalizability coefficients could be estimated and were non-zero. Generalizability coefficients for the observed design (2 raters conducting 25 observations within each school)

were .62 for teacher positive behavior expectations, .07 for teacher disapproval, .72 for teacher reactive behavior management, .27 for student noncompliance, .70 for student disruption, and .74 for student profanity. Notably, differences in generalizability coefficients are apparent when comparing one versus two rater scenarios for a given number of observations within a school, indicating a relatively modest gain in reliability in a two-rater design compared to a one-rater design. Increasing the number of observations within a school also resulted in increases in generalizability coefficients, although these differences also appeared relatively modest. Differences in generalizability coefficients between one-rater 25/observation and one-rater/40 observation designs were comparable across tally categories (typically less than .10). In sum, these analyses suggest relatively limited improvements in reliability associated with a decrease in raters or increases in the number of observations within a school. Returns in reliability are directly related to the design facets that explain the most amount of variance. That is, for tally items in which higher proportions of variance are attributable to raters, generalizability coefficients will decrease more in one-rater designs compared to those tally items in which less variance is attributable to rater. Likewise, for tally items in which higher proportions of variance are attributable to observation, there will be a larger return in reliability for increasing the number of observations within a school compared to tallies in which less variance is attributable to observation.

Conclusions:

Description of conclusions, recommendations, and limitations, based on findings.

PBIS aims to improve the quality of the school environment as a way of fostering the positive social and academic development of students. Observational measures such as the ASSIST are an important component of program evaluation designs. GT provides a framework for estimating multiple sources of error in an outcome measure and the resulting reliability. As seen in the present study, variations in the number of raters used to conduct observations and the number of observations occurring within a school had direct implications for the reliability of ASSIST tally items. Specific recommendations can be made based on these results and will be discussed in more detail in the presentation.

The primary limitation of this study was a small sample size for the object of measurement (high schools). Therefore all results presented here must be interpreted with caution. Given the small sample size, it was especially notable that for four of the tallies, discernable amounts of school-level variance were detected (6 to 9%) and the observed design specifications yielded adequate reliability estimates (.62 to .74). Thus, studies using the ASSIST may be especially well-equipped to detect school-level differences in these behaviors, even in very small sample sizes. A secondary limitation was that the present study included a limited number of design facets. Future analyses would be strengthened by including additional potential sources of variance such as day of the week, subject matter observed, etc.

In closing, our results indicate that raters can be trained to observe selected behaviors with high levels of agreement and that studies using the ASSIST may be well-positioned to capture school-level differences for the behaviors of teacher proactive behavior expectations, teacher reactive behavior management, student disruption, and student profanity. In addition, our study demonstrates GT as a useful tool when evaluating and planning for the use of observational measures in large-scale program evaluations.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Cardinet, J., Tourneur, W., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, *13*, 119-135. doi: 10.1111/j.1745-3984.1976.tb00003.x

Crick, J. E. & Brennan, R. L. (1983). Manual for GENOVA: A generalized analysis of variance system [Computer software]. Iowa City: IA: The American College Testing Program.

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: John Wiley & Sons.

Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2013). 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. *Prevention Science*. Advance online publication. doi: 10.1007/s11121-012-0357-3

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale." *Educational Assessment*, *13*, 267-300. doi: 10.1080/10627190802602541

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

Raudenbush, S. W. & Sampson, R. J. (2002). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology*, *29*, 1-41. doi: 10.1111/0081-1750.00059

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922-932. doi: 10.1037/0003-066X.44.6.922

Sugai, G., & Horner, R. H. (2006). A promising approach for expanding and sustaining School-Wide Positive Behavior Support. *School Psychology Review*, *35*, 245-259.

Appendix B. Tables and Figures

Not included in page count.

Table 1

Descriptive Statistics for ASSIST Classroom Tally Items

	M	*SD*	Min	Max
Teacher				
Proactive behavior expectations	3.91	3.35	0	20
Approval	2.24	2.86	0	15
Disapproval	0.11	0.36	0	3
Reactive behavior management	2.59	3.66	0	24
Opportunity to respond	16.14	11.55	0	54
Student				
Noncompliance	0.04	0.32	0	4
Disruptive	7.85	5.79	0	30
Verbal aggression	0	0	0	0
Physical aggression	< .01	0.06	0	1
Profanity	0.16	0.71	0	10

Table 2
Variance components estimated in the G-studies and their interpretations

Variance component	Symbol	Interpretation
School	σ_s^2	The object of measurement; universe score variance that reflects average differences in tally frequencies across schools.
Observation	$\sigma_{o:s}^2$	Variance attributable to average differences in tally frequencies across observations within a school.
Rater	σ_r^2	Variance attributable to average differences in tally frequencies across raters.
School x rater	σ_{sr}^2	Variance attributable to average differences in school-level tally frequencies across raters.
Observation x rater	$\sigma_{r(o:s)}^2$	Variance attributable to average differences in tally frequencies across observations and raters within a school.

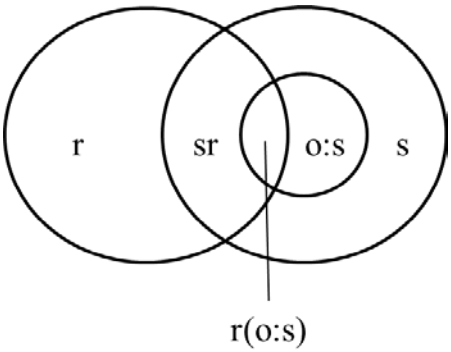

Figure 1. Graphic depiction of the partially crossed G-study design where s = school; o:s = observations (nested in schools); r = rater; sr = school by rater interaction; r(o:s) = observation by rater interaction.

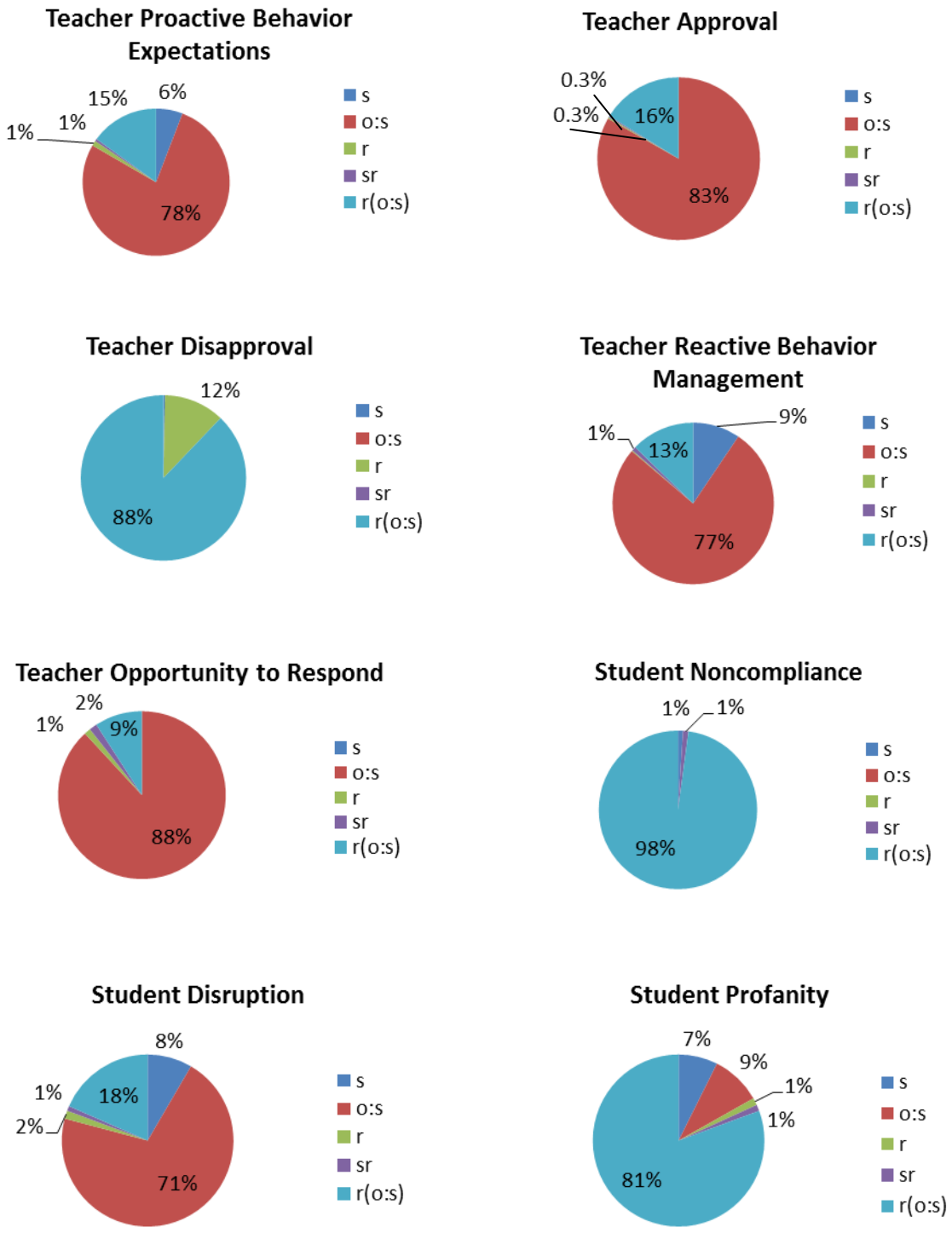

Figure 2. Proportion of variance explained by variance components in each of four tally categories with non-zero school-level variance.

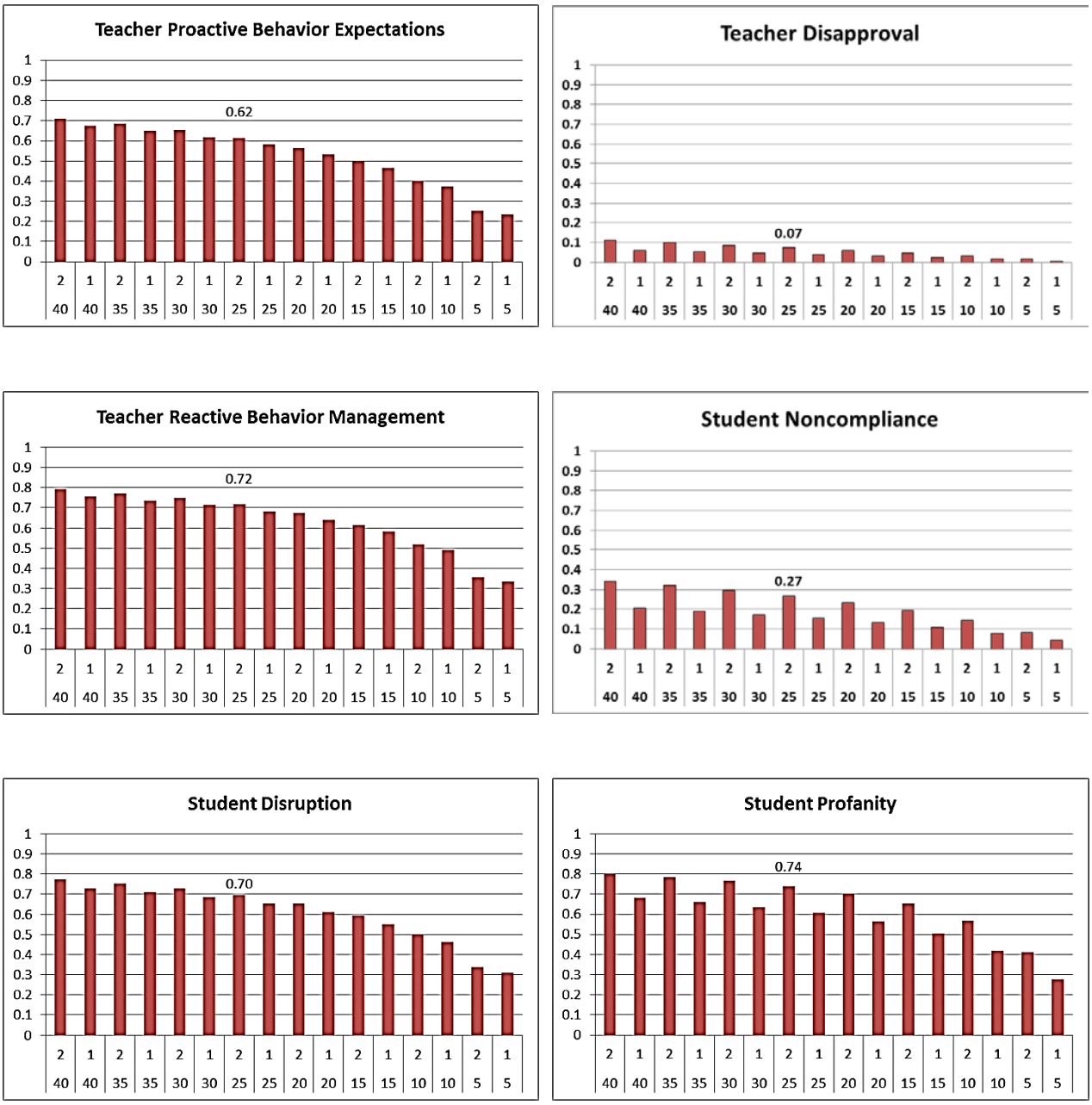

Figure 3. Generalizability coefficients for observed and hypothetical data collection designs. Coefficients for the observed data collection design are indicated by the bar with the numeric data label. The top row of the x-axis corresponds to the number of raters; the bottom row corresponds to the number of observations conducted within a school.