Article

Educational and Psychological Measurement
2014, Vol. 74(2) 280–291

DOI: 10.1177/0013164413508774
epm.sagepub.com

A Practical Solution to Optimizing the Reliability of Teaching Observation Measures Under Budget Constraints

J. Patrick Meyer[1], Xiang Liu[1], and Andrew J. Mashburn[2]

Abstract

Researchers often use generalizability theory to estimate relative error variance and reliability in teaching observation measures. They also use it to plan future studies and design the best possible measurement procedures. However, designing the best possible measurement procedure comes at a cost, and researchers must stay within their budget when designing a study. In this study, we applied the LaGrange multiplier method to obtain facet sample size equations that minimize relative error variance (hence maximize reliability) under budget constraints. We did this for a crossed design and three nested designs that are more typical of data collection in teaching observation studies. Using an example budget and variance components similar to those found in practice, we demonstrate the use of these equations. We also show the way variance components from fully crossed designs can be combined to use our equations for a nested design.

Keywords

generalizability theory, cost, optimum reliability

[1]University of Virginia, Charlottesville, VA, USA
[2]Portland State University, Portland, OR, USA

Corresponding Author:
J. Patrick Meyer, Curry School of Education, University of Virginia, P.O. Box 400265, 405 Emmet Street South, Charlottesville, VA 22904, USA.
Email: meyerjp@virginia.edu

Teaching observation measures commonly involve multiple sources of measurement error, such as the raters judging each teacher's performance and the occasions on which performances are observed. Generalizability theory provides a set of tools for evaluating the influence of multiple sources of error on teaching observation measures and for designing optimal measurement procedures. The flexibility of this method allows researchers to collect pilot data and plan a future measurement procedure by conducting decision studies that use different data collection designs and different facet sample sizes than those used in the generalizability study. A practical limitation of the decision study is that facet sample sizes cannot be increased ad infinitum. Costs are associated with each new rater, and occasion of observation and budget restrictions limit the amount of measurement precision that is attainable. Thus, the challenge in planning a future measurement procedure is to identify facet sample sizes that maximize reliability without exceeding a budget limit.

Woodward and Joe (1973) applied constrained optimization methods to the problem and derived equations for facet sample sizes that maximized reliability under budget constraints. Although their work produced equations for optimal sample sizes in two- and three-facet crossed designs, it does not easily extend to more than three facets or nested designs (Saunders, Theunissen, & Baas, 1989). More feasible solutions involve discrete optimization (Saunders et al., 1989), the Cauchy-Schwartz inequality (Saunders, 1992), or the LaGrange multiplier method (Goldstein & Marcoulides, 1991; Marcoulides & Goldstein, 1990).

The main limitation to existing research is the focus on crossed designs, which are not the norm in teaching observation protocols. Many operational systems use one or two levels of nesting such as lessons nested within teacher (see Hill, Charalambous, & Kraft, 2012) or segments nested within lessons nested within teacher (see Mashburn, Meyer, Allen, & Pianta, 2013). Consequently, sample size equations from previous research on budget constrained reliability optimization are not applicable to teaching observation measures. The one exception is Marcoulides (1997), who provided optimal sample size equations for a nested design. The purpose of this article is to extend previous research by deriving equations for facet sample sizes that minimize error variance (hence maximize reliability) subject to budget constraints in a variety of nested decision study designs. We also demonstrate the way variance components from crossed designs can be used to obtain optimal sample sizes in nested design. Our intended audience is practicing researchers. Therefore, throughout this article we use extant data to motivate the problem and in a step-by-step fashion demonstrate the design of optimal measurement procedures in practice.

Example Costs, Budget, and Variance Components

We use the following information from our experience with the Classroom Assessment Scoring System–Secondary (CLASS-S; Pianta, Hamre, Hayes, Mintz, & LaParo, 2008) to define observation costs and budget constraints. Suppose that each teacher videotapes one lesson for a small stipend of $5. She then spends $5 for

Table 1. Variance Components for the Two- and Three-Facet Crossed Designs.

Component	Notation	Two-facet design	Three-facet design
Teacher, *t*	$\sigma^2(\tau)$	0.186	0.227
Segment, *s*	$\sigma^2(s)$	—	0.013
Lesson, *l*	$\sigma^2(l)$	0.009	0.005
Rater, *r*	$\sigma^2(r)$	0.261	0.192
r×*s*	$\sigma^2(rs)$	—	0.007
r×*l*	$\sigma^2(rl)$	0.001	0.009
t×*r*	$\sigma^2(tr)$	0.072	0.053
s×*l*	$\sigma^2(sl)$	—	0.006
t×*s*	$\sigma^2(ts)$	—	0.026
t×*l*	$\sigma^2(tl)$	0.225	0.085
r×*s*×*l*	$\sigma^2(rsl)$	—	0.005
t×*r*×*s*	$\sigma^2(trs)$	—	0.053
t×*r*×*l*	$\sigma^2(trl)$	0.233	0.017
t×*s*×*l*	$\sigma^2(tsl)$	—	0.298
t×*r*×*s*×*l*	$\sigma^2(trsl)$	—	0.321

postage to mail a \$5 digital videotape to a team of researchers who then have raters view and code the videotape. The cost of one rater coding a videotape is \$30 (two hours of work at \$15 per hour). Let c represent the cost of one rater observing one lesson that is taught by a single teacher; then the total cost for this single observation is $c = \$45$.

To define a budget, let $B^* = Bk$ represent the total budget for observing k teachers, and let B represent the budget per teacher. In our examples below, we will define our cost constraints using a total budget of $B^* = \$120,000$ for a study that involves a total of $k = 75$ teachers. The budget per teacher is then $B = \$120,000/75 = \$1,600$. We aim to design a study that gives us the most reliability that we can afford with this budget.

With costs and budgets defined, we can define a simple cost constraint for a two-facet design using $cn_l n_r \leq B$, where n_l represents the number of lessons and n_r represents the number of raters. This constraint can be interpreted to mean that the cost for all raters observing a teacher in all lessons must be less than or equal to our per teacher budget. It easily extends to $cn_l n_r n_s \leq B$ in a three-facet design, where n_s represents the number of segments. The interpretation is the same, but the cost per teacher increases because of the need to observe multiple segments. In each type of design, other cost constraints are possible, such as those that include one-time costs for a lesson and one-time costs for a rater (see Marcoulides, 1993; Saunders, 1992).

Table 1 lists variance components for a two-facet and three-facet crossed design. We derived these variance components from a study of various observational designs (see Mashburn et al., 2013). This article focuses on the variance components for the instructional support scale of CLASS-S.[1]

Optimal Sample Sizes for the $t \times l \times r$ Design

The generalizability coefficient is an index of reliability that is suitable for norm-referenced testing and relative decisions (Brennan, 2001). It is defined as the ratio of universe score variance, $\sigma^2(\tau)$, to observed score variance. It is given by $E\rho^2 = \sigma^2(\tau)/[\sigma^2(\tau) + \sigma^2(\delta)]$. In a $t \times l \times r$ crossed design,[2] every teacher, t, uses the same lessons, l, and raters, r, observe every teacher in every lesson. Implementing this design in practice requires standardization of the lessons so that every teacher is teaching the same thing. Treating teachers as the object of measurement, relative error is

$$\sigma^2(\delta) = \frac{\sigma^2(tl)}{n_l} + \frac{\sigma^2(tr)}{n_r} + \frac{\sigma^2(tlr)}{n_l n_r}.$$

We can maximize the generalizability coefficient by minimizing the amount of relative error variance, $\sigma^2(\delta)$. If there are no budget constraints, this minimum is obtained when facet sample sizes (e.g., n_l and n_r) are infinitely large. However, budget constraints are a reality of practical research, and no one can afford an infinite number of facets. Costs increase as facet sample sizes increase. The trick is finding the largest facet sample sizes that one can afford.

To minimize relative error subject to the cost constraint, $cn_l n_r \leq B$, we find the stationary points of the LaGrange function, $F(n_l, n_r, \lambda) = \sigma^2(\delta) - \lambda(cn_l n_r B)$. Marcoulides and Goldstein (1990) showed these points to be

$$n_l = \sqrt{\frac{\sigma^2(tl)}{\sigma^2(tr)}\frac{B}{c}} \quad \text{and} \quad n_r = \sqrt{\frac{\sigma^2(tr)}{\sigma^2(tl)}\frac{B}{c}}, \tag{1}$$

which agree with the work of Woodward and Joe (1973). Using our example costs and budget and the variance components listed in Table 1, the optimal sample sizes are $n_l = 10.5$ and $n_r = 3.4$. We round these numbers to 11 and 3, respectively, given that we want to use an entire lesson and a whole rater. Substituting these facet sample sizes into the expression for relative error leads to a value of $\sigma^2(\delta) = 0.052$ and an optimal reliability of $E\rho^2 = 0.78$. Thus, our study must involve 11 lessons and 3 raters to achieve the largest possible reliability estimate that we can afford. The only way to achieve a higher reliability in this scenario is to either increase the total budget or cut costs in some other way.

Optimal Sample Sizes for the $r \times (l : t)$ Design

We extended the work of Marcoulides and Goldstein (1990) and Marcoulides (1997) by applying the LaGrange multiplier method to three nested designs. In the first design, lessons are unique to each teacher, but every rater observes every lesson nested within teacher, $r \times (l : t)$. This design mimics actual teaching practice in that

Table 2. The Effect of Rounding on Reliability and Total Costs for the $r\times(s:l:t)$ Design.

n_r	n_l	$\sigma^2(\delta)$	$E\rho^2$	Cost per teacher
1.9	9.37	0.068	0.77	$1,602.27
1	9	0.106	0.68	$810
1	10	0.101	0.69	$900
2	8	0.073	0.76	$1,440
2	9	0.068	0.77	$1,620
2	10	0.064	0.78	$1,800

teachers can teach whatever lesson they choose; there is no standardization of lessons. Relative error variance is given by

$$\sigma^2(\delta)=\frac{\sigma^2(tr)}{n_r}+\frac{\sigma^2(l:t)}{n_l}+\frac{\sigma^2(r\times[l:t])}{n_rn_l}.$$

Using the LaGrange function $F(n_l,n_r,\lambda)=\sigma^2(\delta)-\lambda(cn_ln_r-B)$, the optimal sample sizes for the number of raters and lessons are (see the appendix)

$$n_l=\sqrt{\frac{\sigma^2(l:t)}{\sigma^2(tr)}\frac{B}{c}}\quad\text{and}\quad n_r=\sqrt{\frac{\sigma^2(tr)}{\sigma^2(l:t)}\frac{B}{c}}. \tag{2}$$

These equations agree with Equations 13 and 14 in Marcoulides (1997) when there is only one type of cost associated with a rater and lesson.

Although we use variance components from a $r\times(l:t)$ design in these equations, we do not have to collect data in this manner. A benefit of generalizability theory is that decision studies can involve the same data collection design as the generalizability study or something different. In particular, we can collect data with a crossed design and consider various nested designs by combining variance components in the crossed design to obtain nested design components. Brennan (2001) explains the rules for combining variance components. To use Equation 2, with variance components from a crossed design, the variance component for lessons nested within teachers is $\sigma^2(l:t)=\sigma^2(l)+\sigma^2(tl)$, where the terms on the right-hand side are obtained from the crossed design. Substituting these terms in Equation 2 results in

$$n_l=\sqrt{\frac{\sigma^2(l)+\sigma^2(tl)}{\sigma^2(tr)}\frac{B}{c}}\quad\text{and}\quad n_r=\sqrt{\frac{\sigma^2(tr)}{\sigma^2(l)+\sigma^2(tl)}\frac{B}{c}}. \tag{3}$$

Using variance components from Table 2, the optimal sample sizes are $n_l=10.7$ and $n_r=3.3$. These values are slightly different from those from the fully crossed design. Comparing the equations in 1 and 3 makes evident that their difference is mainly due to the magnitude of $\sigma^2(l)$. If this value is zero, the optimal sample sizes from the two

designs will be the same. Otherwise, the nested design will require more lessons and fewer raters to reach the optimal reliability level in each design.

Using $\sigma^2(r\times[l:t]) = \sigma^2(rl) + \sigma^2(trl)$ in the equation for relative error variance and rounding the optimal sample sizes to 11 and 3, the relative error variance is 0.052, and the generalizability coefficient is 0.78. These values are also similar to the values from the fully crossed design, but this result is mainly due to $\sigma^2(rl)$ being so close to zero. It will not always be the case that the rounded optimal sample sizes, relative error variance, and generalizability coefficient from the crossed design are the same as those in the $r\times(l:t)$ design. It just worked out that way in this example.

Optimal Sample Sizes for the $t\times r\times(s:l)$ Design

Obtaining optimal sample sizes in a three-facet design is possible, but the solution involves extensive equations and imaginary numbers. It is unlikely that a researcher would use such equations in practice. An alternative is to numerically optimize the LaGrange function through computer intensive methods or adjust the expression for relative error variance in some way to simplify the problem. Marcoulides and Goldstein (1990) chose the latter and obtained an upper bound to the relative error variance by eliminating facet sample size terms from the equation. Their choice of terms to eliminate was arbitrary and can lead to optimized relative error variance that notably differs from the numerical solution.

We approached the problem from a different perspective. Many teaching observation measures operationally define the number of segments, n_s, that must be collected within a lesson. For example, CLASS-S operational procedures suggest two segments per lesson (Pianta et al., 2008), and the Mathematical Quality of Instruction typically involves four segments (Hill et al., 2008). Therefore, we fixed the number of segments to a constant value in the optimization of the LaGrange function. We only optimized the function with respect to the number of raters, number of lessons, and the LaGrange multiplier.

In the $t\times r\times(s:l)$ design, every rater observes every teacher's segment that is nested within the lesson, $s:l$. To achieve this design in practice, the lessons must be standardized in some fashion so that every teacher completes the same lessons. Every rater must view every lesson for every teacher, which is best achieved through videotaped lessons. Relative error variance for this design is given by

$$\sigma^2(\delta) = \frac{\sigma^2(tr)}{n_r} + \frac{\sigma^2(tl)}{n_l} + \frac{\sigma^2(trl)}{n_r n_l} + \frac{\sigma^2(t\times[s:l])}{n_s n_l} + \frac{\sigma^2(tr\times[s:l])}{n_r n_s n_l}.$$

Using the LaGrange function $F(n_l, n_r, \lambda) = \sigma^2(\delta) - \lambda(cn_s n_l n_r - B)$, the optimal sample sizes for the number of raters and lessons are then

$$n_r = \frac{\sqrt{\sigma^2(tr)\frac{B}{c}}}{\sqrt{n_s\sigma^2(tl) + \sigma^2(t\times[s:l])}} \quad \text{and} \quad n_l = \frac{\left\{n_s\sigma^2(tl) + \sigma^2(t\times[s:l])\right\}\sqrt{\sigma^2(tr)\frac{B}{c}}}{n_s\sigma^2(tr)\sqrt{n_s\sigma^2(tl) + \sigma^2(t\times[s:l])}}. \quad (4)$$

We can substitute variance components from a $t\times r\times(s:l)$ design into Equation 4 along with budget information and costs to obtain the optimal number of raters and lessons. We can also use variance components from a completely crossed design such as those in the last column of Table 1 by combining crossed effects that are confounded in the nested design. Nested variance components expressed as confounded crossed effects are $\sigma^2(t\times[s:l])=\sigma^2(ts)+\sigma^2(tsl)$ and $\sigma^2(t\times r\times[s:l])=\sigma^2(trs)+\sigma^2(trsl)$. There are other confounded crossed effects in the nested design, but these two are the only ones needed for the optimal sample size equations and relative error variance computations.

Using our example budget and costs and the variance components in Table 1, we get optimal sample sizes of $n_r=1.95$ and $n_l=9.1$. The sample size for the number of segments was not part of the optimization. We fixed that value at $n_s=2$. Rounding the number of raters and number of lessons to 2 and 9, respectively, we get an optimal relative error variance of 0.065 and an optimal generalizability coefficient of 0.78.

Optimal Sample Sizes for the r×(s : l : t) Design

This design differs from the previous one in that lessons are no longer standardized. Teachers are free to teach whatever lesson they choose. Consequently, lessons are unique to each teacher and, naturally, segments are nested within the lesson. Every rater observes every segment within lesson within teacher in this design. The relative error variance is

$$\sigma^2(\delta)=\frac{\sigma^2(tr)}{n_r}+\frac{\sigma^2(l:t)}{n_l}+\frac{\sigma^2(r\times[l:t])}{n_rn_l}+\frac{\sigma^2(s:l:t)}{n_sn_l}+\frac{\sigma^2(r\times[s:l:t])}{n_rn_sn_l}.$$

As done in the previous section, the number of segments is fixed to the value n_s. Optimization of the LaGrange function, $F(n_l,n_r,\lambda)=\sigma^2(\delta)-\lambda(cn_sn_ln_r-B)$, is only done with respect to the number of raters and lessons. Optimal sample sizes for the number of raters and number of lessons are then

$$n_r=\frac{\sqrt{\sigma^2(tr)\frac{B}{c}}}{n_s\sigma^2(l:t)+\sigma^2(s:l:t)}\quad\text{and}\quad n_l=\frac{[n_s\sigma^2(l:t)+\sigma^2(s:l:t)]\sqrt{\sigma^2(tr)\frac{B}{c}}}{n_s\sigma^2(tr)\sqrt{n_s\sigma^2(l:t)+\sigma^2(s:l:t)}}. \quad (5)$$

We can collect data as a $r\times(s:l:t)$ design and substitute estimated variance components into Equation 5 to obtain the optimal number of raters and lessons. We can also use variance components from a crossed design, such as those in the last column of Table 1. Nested design variance components composed of confounded effects in the crossed design are

$$\sigma^2(l:t)=\sigma^2(l)+\sigma^2(tl),$$

$$\sigma^2(r\times[l:t])=\sigma^2(lr)+\sigma^2(tlr),$$

$$\sigma^2(s:l:t) = \sigma^2(s) + \sigma^2(sl) + \sigma^2(ts) + \sigma^2(tsl),$$

and

$$\sigma^2(r\times[s:l:t]) = \sigma^2(sr) + \sigma^2(slr) + \sigma^2(tsr) + \sigma^2(tslr).$$

As noted before, there are other confounded crossed effects in the nested design, but these are the only ones needed for the optimal sample size equations and relative error variance.

Using our example budget and costs and the variance components in Table 1, we get optimal sample sizes of $n_r = 1.9$ and $n_l = 9.37$. The sample size for the number of segments was not part of the optimization. We fixed it to $n_s = 2$. Rounding the number of raters and number of lessons to 2 and 9, respectively, we get a relative error variance of 0.068 and a generalizability coefficient of 0.77.

Impact of Rounding on the Total Budget

The LaGrange multiplier method often results in optimal sample sizes that are not integers. The simple solution to this problem is to round the results to the nearest whole number. We used this approach in the previous examples. However, the problem with rounding is that it could lead to costs that exceed or fall short of the total budget. In the example for the $r\times(s:l:t)$ design, rounding resulted in a cost per teacher of \$1,620, which is \$20 more than we would like to spend. While this amount seems trivial, it puts us \$1,500 over budget for the entire sample of 75 teachers. A simple solution to the problem of rounding is to consider all possible permutations of rounding up and rounding down sample sizes (see Goldstein & Marcoulides, 1991). Table 2 shows these permutations for the $r\times(s:l:t)$ design. The first row of Table 2 lists the optimal sample sizes, but the results are not integers. The remaining rows show the permutations for rounding up and rounding down sample sizes to the nearest integer. The one exception is the solution with two raters and eight lessons. We added this row because the solution with the highest reliability value without going over budget left us \$700 under budget. Given that we could spend more money, we considered the solution with two raters and eight lessons, and it turned out to provide a reliability of 0.76 and kept us \$160 under our per teacher budget. Thus, the best solution (i.e., highest level of reliability without going over budget) is to use two raters and eight lessons.

Evaluating all permutations of rounding sample sizes is a simple way to avoid going over budget. However, Saunders, Theunissen, and Baas (1989) described an integer programming method for obtaining optimal sample sizes that does not require rounding. The benefit of their method is that it leads to integer-based sample sizes, and no rounding is necessary. Perhaps the main disadvantage is that integer programming is computationally intensive, and algorithms must be adapted to every decision study design. All things considered, our advice for using the equations in this article is to check the results to make sure that your budget has not been exceeded. If it has,

then evaluate other permutations for rounding up and rounding down facet sample sizes.

Discussion

Researchers are often faced with the task of planning a study that involves teaching observation measures. The goal is to set a budget and design a study that maximizes reliability. Building on the work of Marcoulides and Goldstein (1990) and Marcoulides (1997), we used the LaGrange multiplier method to obtain equations for optimal facet sample sizes in three different nested designs. We chose three designs that are typical of teaching observation measures.

Our equations for the $r\times(l:t)$ are similar to those from the $r\times l\times t$ design. Indeed, these two designs will produce the same optimal sample sizes when $\sigma^2(l)=0$. Otherwise, the nested design will require more lessons and fewer raters to obtain the optimal reliability level.

Our equations from the three-facet designs are not directly comparable to the work of others (e.g., Marcoulides & Goldstein, 1990) because of the way we simplified the expression for relative error variance. We fixed the number of segments to a constant in the optimization of the LaGrange function, whereas Marcoulides and Goldstein defined an upper bound to relative error variance. Therefore, we did not evaluate the conditions under which equations from the three-facet nested and crossed designs would produce similar results.

We attempted to compare results from our two nested designs with results from a fully crossed design that also simplified the problem by fixing the number of segments to a constant value. However, even with the number of segments fixed to a constant, the solution for stationary points in the fully crossed designs involved complex numbers. We did not include those equations here because we felt they would be too difficult for a researcher to use in practice. It would be easier to use numerical methods to obtain the optimal solutions.

In summary, we derived facet sample size equations that minimize relative error variance and maximize reliability for three nested designs. We then showed the way variance components from a crossed design can be combined to obtain optimal sample sizes in a nested design. We applied these equations to a budget scenario based on our experience and variance components derived from the data collected with the CLASS-S instrument.

Limitations

A limitation of the current article is we only considered a simple cost function with fixed costs per rater and occasion. Consequently, the sample size equations presented herein are not applicable to more elaborate cost functions, and they cannot be applied to situations that involve variable costs. To demonstrate the way results change with different cost functions, we return to the $r\times(l:t)$ and consider the use of variable

costs subject to the same constraint that our per teacher budget is $1,600. Suppose that we have a group of novice raters who earn $15 per hour. In keeping with our previous description, the cost for this type of rater is $45. Now suppose we also have a group of expert raters who earn $30 per hour. The cost for this second type of rater is then $75—2 hours of work plus the $15 cost of asking the teacher to create and mail the video tape. Using equations given by Marcoulides (1997), we know the optimal sample sizes are 14 lessons and 5 raters. However, labor is divided such that three novice raters judge nine lessons per teacher, and two expert raters judge five lessons per teacher. These results lead to a relative error variance of 0.0345 and a generalizability coefficient of 0.84. Thus, using variable costs we are able to afford more lessons and raters and higher level of reliability than we could afford using fixed costs.

Appendix

To demonstrate the procedures for maximizing reliability under budget constraints, we return to the $r\times(s:l:t)$ design. There are three facet sample sizes in this design, but the problem is simplified by fixing the number of segments, n_s, to a constant value. In our example, we fixed the number of segments to 2. Using relative error variance and budget constraints defined earlier the optimum facet sample sizes are obtained through the following steps.

1. Define the LaGrange Function, $F(n_l, n_r, \lambda) = \sigma^2(\delta) - \lambda(c2n_l n_r - B)$.
2. Find the partial derivatives of F and set the results equal to zero.

$$\frac{\partial F}{\partial n_r} = -\frac{\sigma^2(t\times r)}{n_r^2} - \frac{\sigma^2(r\times[l:t])}{n_r^2 n_l} - \frac{\sigma^2(r\times[s:l:t])}{2n_r^2 n_l} - 2\lambda n_l = 0,$$

$$\frac{\partial F}{\partial n_l} = -\frac{\sigma^2(r\times l:t)}{n_r n_l^2} - \frac{\sigma^2(l:t)}{n_l^2} - \frac{\sigma^2(s:l:t)}{2n_l^2} - \frac{\sigma^2(r\times s:l:t)}{2n_r n_l^2} - 2\lambda n_r = 0,$$

$$\frac{\partial F}{\partial \lambda} = -2n_r n_l + \frac{B}{c} = 0.$$

3. Solve the system of equations for n_r and n_d to find the equations for optimal sample sizes. These results were listed earlier in the article. Note that it becomes increasingly difficult to solve this system of equations as the number of facets increases. Goldstein and Marcoulides (1991) described a bisection procedure that quickly finds the roots of this system. The advantage of their approach is that the system can be quickly solved with the help of a computer, and it does not require a researcher to solve the system by hand.

Authors' Note

This article was presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

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: This research was supported by the William T. Grant Foundation (Grant No. 11456).

Notes

1. When we analyzed these data with a crossed design, we ended up with three negative variance component estimates. We changed these negative estimates to positive ones, while also adjusting their magnitude under the constraint that the sum of variance components must be the same as in the actual analysis. Our purpose for using variance components in this study is to simply demonstrate the use of optimal sample size equations. It is not to describe features of the CLASS-S instrument. Therefore, we felt justified in using a ''historical fiction'' version of CLASS-S variance components.
2. Brennan (2001) uses capital letters to represent facets in a decision study design. We use lowercase letters to refer to decision study facets to keep the notation simple.

References

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Goldstein, Z., & Marcoulides, G. A. (1991). Maximizing the coefficient of generalizability in decision studies. *Educational and Psychological Measurement*, *51*, 79-88.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430-511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teaching observation systems and a case for the generalizability study. *Educational Researcher*, *41*, 56-64.

Marcoulides, G. A. (1993). Maximizing power in generalizability studies under budget constraints. *Journal of Educational Statistics*, *18*, 197-206.

Marcoulides, G. A. (1997). Optimizing measurement designs with budget constraints: The variable cost case. *Educational and Psychological Measurement*, *57*, 808-812.

Marcoulides, G. A., & Goldstein, Z. (1990). The optimization of generalizability studies with resource constraints. *Educational and Psychological Measurement*, *50*, 761-768.

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2013). *The effect of observation length and presentation order on the reliability and validity of classroom observations* (Unpublished manuscript).

Pianta, R. C., Hamre, B., Hayes, N., Mintz, S., & LaParo, K. M. (2008). *Classroom Assessment Scoring System–Secondary (CLASS-S)*. Charlottesville: University of Virginia.

Saunders, P. F. (1992). Alternative solutions for optimization problems in generalizability theory. *Psychometrika*, *57*, 351-356.

Saunders, P. F., Theunissen, T. J., & Baas, S. M. (1989). Minimizing the number of observations: A generalization of the Spearman-Brown formula. *Psychometrika*, *54*, 587-589.

Woodward, J. A., & Joe, G. W. (1973). Maximizing the coefficient of generalizability in multi-facet decision studies. *Psychometrika*, *38*, 173-181.