Journal of Educational Measurement
Spring 1999, Vol. 36, No. 1, pp. 61–71

Note on Sources of Sampling Variability in Science Performance Assessments

Richard J. Shavelson
Maria Araceli Ruiz-Primo
Edward W. Wiley
Stanford University

In 1993, we reported in Journal of Educational Measurement *that task-sampling variability was the Achilles' heel of science performance assessment. To reduce measurement error, tasks needed to be stratified before sampling, sampled in large number, or possibly both. However, Cronbach, Linn, Brennan, & Haertel (1997) pointed out that a task-sampling interpretation of a large person × task variance component might be incorrect. Task and occasion sampling are confounded because tasks are typically given on only a single occasion. The* person × *task source of measurement error is then confounded with the* pt × *occasion source. If* pto *variability accounts for a substantial part of the commonly observed* pt *interaction, stratifying tasks into homogenous subsets—a cost-effective way of addressing task sampling variability—might not increase accuracy. Stratification would not address the* pto *source of error. Another conclusion reported in* JEM *was that only direct observation (DO) and notebook (NB) methods of collecting performance assessment data were exchangeable; computer simulation, short-answer, and multiple-choice methods were not. However, if Cronbach et al. were right, our exchangeability conclusion might be incorrect. After re-examining and reanalyzing data, we found support for Conbach et al. We concluded that large task-sampling variability was due to both the person × task interaction and person × task × occasion interaction. Moreover, we found that direct observation, notebook and computer simulation methods were equally exchangeable, but their exchangeability was limited by the volatility of student performances across tasks and occasions.*

In previous research reported in *Journal of Educational Measurement*, we provided evidence showing that task-sampling variability in students' scores was the major source of measurement error in science (and other) performance assessments (Shavelson, Baxter & Gao, 1993). We also showed that variability in performance scores due to occasion sampling was substantial as well (Ruiz-Primo, Baxter, &

The work reported herein was supported, in part, under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education, and, in part, by the National Science Foundation award Number ESI-9596080. The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, the U.S. Department of Education or the National Science Foundation. We are indebted to Lee Cronbach and Ed Haertel for their helpful comments. Of course, we alone are responsible for the content of the note.

Shavelson, 1993). Finally, we reported that variability in scores due to method sampling (observation, notebook, computer simulation, short-answer, multiple-choice) was large for all but the observation/notebook pair. This last finding led us to question the exchangeability of the computer simulation and pencil-and-paper methods for hands-on performance assessment (Shavelson et al. 1993; see also Baxter & Shavelson, 1994).

Cronbach, Linn, Brennan, and Haertel (1997) directly questioned the proper interpretation of task-sampling variability. They argued that task-sampling variability was confounded with occasion-sampling variability in common practice because students typically perform assessment tasks on a single occasion. What appears to be task-sampling variability might very well be occasion sampling, or a combination of task and occasion sampling. By raising this possibility, they not only led us to question our interpretation of task-sampling variability as the major source of measurement error in performance assessment, they also led us to rethink our exchangeability conclusion.

To evaluate the Cronbach et al. argument, task and occasion sampling variability need to be disentangled. One possibility is a person × task × occasion design. This design estimates variability due to *pt* and to the residual, *pto,e*. However, in this design, *pto* cannot be separated from the overall residual error term (e). A design that separates *pto* from residual error incorporates a third facet, say rater: *ptor*. With this design, sampling variability due to *pt* and *pto* can be isolated from residual error (*ptor,e*).[1]

In this note we report a re-analysis of data collected in a person × task × occasion × rater or method G-study design (Ruiz-Primo et al., 1993; also reported as the first science data set in Shavelson et al., 1993). We then bring this re-analysis to bear on our interpretation of task-sampling variability, and on the convergence of different performance assessment methods.

The re-analysis leads us to modify our original task-sampling interpretation: task sampling variability, *confounded with occasion sampling variability*, is the major source of measurement error in students' performance scores. The consequence of this finding is of considerable import. If instability in scores were due solely to task sampling, this variability might be reduced by a stratified task-sampling procedure with only a small increase (if any) in the number of tasks needed for an assessment. However, since performance assessments are typically given on a single occasion, stratification cannot address the confounding of tasks and occasions; the only solution is to increase task samples substantially to address instability in student performance over tasks and from one occasion to another. We also modify our conclusion about the convergence of measurement methods, providing evidence that computer simulation along with direct observation and notebook methods—all methods that react to the actions taken by students in conducting a science investigation—converge. These three methods do not converge with their paper-and-pencil counterparts.

Task Sampling Variability In Performance Assessment

Cronbach et al. (1997) rightly pointed out that, because performance assessment scores are almost always collected on a single occasion, task-sampling variability is confounded with occasion sampling variability. Interpreting task-sampling variability (i.e., person × task interaction) as the principal source of measurement error ignores occasion sampling or assumes it to be negligible. Neither is justified in the absence of strong empirical support.

Cronbach et al. considered a pupil × task × judge measurement design in which each pupil performed a sample of tasks and that pupil's performance was evaluated numerically by a sample of judges. This design produces seven possible sources of variance in scores: pupil, task, judge, pupil × task, pupil × judge, task × judge and a residual (pupil × task × judge confounded with other sources of error not explicitly included in the measurement design). According to Cronbach et al. (p. 384):

> Each of the seven components includes temporal effects. Thus, relabeling the pupil component as *p,po* [where o stands for occasion and *po* is the pupil × occasion interaction], would emphasize that the pupil may perform better throughout the week of assessment than usual by virtue of some morale-inducing event or may perform worse because of illness. The full set of confounds of occasion effects in this design is *(p,po)*, *(t,to)*, *(j,jo')* [*o*' reflects the fact that with most assessments, performance is not scored in real time by an observer but rather from a notebook at a later time and one judge usually evaluates performance after another has finished[2]], ***(pt,pto)***, . . . , and ***(ptj, ptjo, ptjo', ptjoo')***. To reduce clutter, we do not use compound labels throughout but we do apply them to the fourth and seventh components [bolded by present authors] ***to warn against not uncommon misinterpretations*** [italics on words ours].

The person × task interaction (*pt,pto*, also labeled *pt,e*, by Cronbach et al.) is what we (e.g., Shavelson et al. 1993; Ruiz-Primo & Shavelson, 1997) previously have interpreted as task sampling variability—"the major source of error variability in performance assessment" (Shavelson et al., 1993, p. 215). However, this variance component consists of

> . . . [t]wo very distinct subcomponents. The *pt* portion—person-task interaction—describes the reproducible, consistent tendency of the pupil to do especially well on this task and relatively badly on that one . . . [T]he *e* (or *pto*) portion recognizes fluctuations from occasion to occasion; these arise from mood, distraction, momentary insights and confusions, and, in some tasks, guessing. Whereas the interaction could be reduced by stratifying tasks in constructing the instrument and recognizing the stratification in the analysis, the e variance could be reduced only by lengthening the test [Cronbach et al., 1997, p. 385].

The pupil × task interaction can only be separated from *e*rror in a retest design, one where pupils perform the same sample of tasks on at least two occasions. "Only a retest study—rarely practicable—could estimate the separate components" (Cronbach et al., 1997, p. 385). Cronbach et al. reported that Gao, Baxter & Shavelson disentangled task and occasion sampling variability, and "two thirds of the (pt,e) variance came from the e [pto] source" (p. 385).[3]

A generalizability study reported by Shavelson et al. (1993), based on data collected by Ruiz-Primo et al. (1993), and a re-analysis of additional data from Ruiz-Primo et al. (1993) provide science performance assessment data in retest G-study designs. Hence these G studies bear directly on the magnitude of task-sampling variability (*pt*) and the confounding of task-sampling variability with occasion-sampling variability (*pto*).

In the Ruiz-Primo et al. study, fifth and sixth grade students were tested on two *occasions*, once in the late spring and then again in the early fall of 1990, with three different performance tasks. On both occasions, each student conducted the tasks in the same order (Ruiz-Primo et al., 1993, p. 46), as presented here. The *tasks* were: (a) "Paper Towels"—given three brands of paper towels and equipment, determine which of the three towels soaks up the most water and which the least; (b) "Electric Mysteries"—given batteries, wires, and light bulbs, hook up circuits external to a series of "mystery boxes" to determine their contents (viz. a bulb, two batteries, a wire, a battery and a bulb, nothing); and (c) "Bugs"—given five sow bugs and equipment, conduct a series of investigations to determine the bugs' choice behavior in different environments (e.g., damp/dry, dark/light; for details see Shavelson, Baxter, & Pine, 1991). The tasks were drawn from activities almost universally found in fifth-sixth grade hands-on curricula.

Each of the tasks was administered by five different *methods*: (a) *direct observation* of hands-on performance, (b) *notebook* based on the observed investigation in which students recorded their procedures and findings, (c) *computer simulation* of the Electric Mysteries and Bugs tasks (Paper Towels depended on saturation and could not be easily simulated), (d) *multiple-choice* in which the item stems described steps in an investigation and possible next steps or conclusions served as alternatives, and (e) *short-answer* where the steps or findings in an investigation were described and the student described the next steps or interpreted the finding. Finally, *raters* (graduate students) were trained to use an analytic scoring method to evaluate performance on the Bugs and Paper Tower tasks; the Electric Mysteries task could be scored objectively (reported accuracy of the contents of the box and the diagram of hook up to box). From these data, then, two different generalizability studies were possible: (a) person × task × occasion × rater for the Bugs and Paper Towels tasks, and (b) person × task × occasion × method for all three tasks.

We reproduce the results of Shavelson et al. (1993, Table 3, p. 223) person (n = 26) × rater (n = 2) × task (n = 2; Bugs and Paper Towels) × occasion (n = 2) G study in our Table 1. As can be seen, the variance component for *pt* is .063 and for *pto* is 1.16. Task sampling variability (*pt,pto*) does indeed account for the lion's share of error variance in performance assessment, and the major source of this variability appears to be the confound with occasion.

To use all three tasks and focus on performance rather than pencil-and-paper (see below), the Ruiz-Primo et al. (1993) data were analyzed in a person (n = 27) × occasion (n = 2) × task (n = 3; Paper Towels, Electric Mysteries, Bugs) × method (n = 2: direct observation, notebook) G-study design. Our analysis of these data is presented in Table 2. Not surprisingly, the findings of this G study are consistent with those reported in Table 1: The variance component for *pt* (0.65) is smaller than the variance component for *pto* (0.79).

The findings of a large person × task × occasion interaction jibes with Ruiz-Primo et al.'s (1993) report that students tended to change their approach to each investigation from one occasion to the next. Note, however, that the estimated variance component for the person × occasion effect is zero (Tables 1 and 2). Even though students approached the tasks differently each time they were tested, the

Table 1
Variance Component Estimates for the Person x Rater x Task x Occasion
G Study Using the Science Data
(from Shavelson, Baxter & Gao, 1993, Table 1, p. 223)

Source of Variability	n	Estimated Variance Component	Percent Total Variability
Person (p)	26	0.07	4
Rater (r)	2	0.00[a]	0
Task (t)	2	0.00[a]	0
Occasion (o)	2	0.01	1
pr		0.01	1
pt		0.63	32
po		0.00[a]	0
rt		0.00	0
ro		0.00	0
to		0.00[a]	0
prt		0.00[a]	0
pro		0.01	0
pto		1.16	59
rto		0.00[a]	0
prto,e		0.08	4
$\hat{\rho}^2$		.04	
φ		.04	

[a] A small, negative variance component was set to zero.

aggregate level of their performance, averaged over the tasks, did not vary from one occasion to another.

A commonly proposed solution to the "task-sampling problem" is to recommend the stratification of tasks. By stratifying, homogeneous subsets of tasks would be formed (assuming the appropriate stratification variable could be found). The person × task interaction within any one subset would be smaller than the interaction estimated without stratification. Moreover, the argument goes, the pooled within-strata person × task interaction (over subsets) would be smaller than the un-stratified $p \times t$ interaction. By stratification, then, the $p \times t$ variance component would be reduced with the consequence that fewer tasks would be needed in an assessment.

Table 2
Variance Component Estimates for the Person x Occasion x Task x Method G Study
(See Ruiz-Primo, Baxter & Shavelson, 1993, for details)

Source of Variability	n	Estimated Variance Component	Percent Total Variability
Person (p)	27	0.33	13.11
Task (t)	3	0.00[a]	0
Occasion (o)	2	0.10	4.14
Method (m)	2	0.00[a]	0
pt		0.65	25.83
po		0.00[a]	0
pm		0.01	0.50
to		0.03	1.08
tm		0.12	4.67
om		0[a]	0
pto		0.79	31.35
ptm		0.02	0.63
pom		0[a]	0
tom		0[a]	0
ptom,e		0.47	18.70

[a] A small, negative variance component was set to zero.

The impact of stratification can be examined, if imperfectly, with the $p \times t \times o \times m$ G study data. Two tasks, Electric Mysteries and Bugs, represent distinctly different areas of science. Although the study was not designed with randomly selected strata from science, let's assume that two strata were selected, one being electricity and the other animal choice behavior—that is, Electric Mysteries and Bugs are our two strata. Following the data in hand, within the Electric Mysteries stratum assume six mystery boxes were sampled (varying the contents of the boxes); within the Bugs stratum assume three tasks were randomly chosen to examine environmental choice behavior (dark/light, damp/dry, 2 × 2 factorial combination). This p × *s*tratum × o × m × *t*ask-within-*s*tratum G study allows us to examine the contention that the within-strata $p \times t$ interaction would be appreciably smaller, relatively, than the $p \times t$ interaction in an un-stratified design (Table 2).

The results of the $p \times (t{:}s) \times o \times m$ G study are presented in Table 3. The major source of measurement error is, once again, the person × task-within-stratum × occasion ($p \times [t{:}s] \times o$) variance component, accounting for 40 percent of total

Table 3
Variance Component Estimates for the Person x Stratum x Occasion x Method x Task:Stratum G Study

Source of Variability	n	Estimated[*] Variance Component	Percent Total Variability
Person (p)	27	0.0216	12.53
Strata (s)	2	0.0000[a]	0
Occasion (o)	2	0.0023	1.32
Method (m)	2	0.0007	0
Task:Stratum (t:s)	3,6	0.0100	5.79
ps		0.0147	8.54
po		0.0001	0.06
so		0.0042	2.42
pm		0.0018	1.06
sm		0.0000	0.02
om		0.0001	0.06
pt:s		0.0085	4.92
t:so		0.0000[a]	0
pso		0.0138	7.97
psm		0.0000[a]	0
pom		0.0000[a]	0
som		0.0000[a]	0
pt:so		0.0683	39.60
pt:sm		0.0045	2.61
t:som		0.0000[a]	0
psom		0.0033	1.89
pt:som		0.0192	11.14

*Minque estimates for unbalanced design.
[a] Small negative variance component set to zero.

variation in scores, compared with 31 percent in the un-stratified design (Table 2). Combining the person × task-within-stratum (*p* × [*t*:*s*]) and the person × task-within-stratum × occasion (*p* × [*t*:*s*] × *o*) variance components to estimate the typical confound in performance assessment when only one testing occasion is used, we find that *p* × (*t*:*s*), *p* × (*t*:*s*) × *o* accounts for 44.52 percent of total variation (Table 3) while *pt*,pto in the un-stratified design accounts for 57.18 percent (Table 2).

Although task-sampling variability is reduced, relatively, in the stratified design, the difference is fairly small.

To what degree will this reduction in task-sampling variability reduce the total number of tasks needed to reach .80 reliability in science performance assessment? The answer is, "not much, if at all." The total number of assessment tasks needed for a reliability of .80 depends not only on the number of tasks within strata, but also on the number of strata: Total $N_t = n_s \times n_t$. The good news about stratification is that it reduced task-sampling variability slightly. However, this reduction comes with a price—introducing strata-sampling variability into scores. In our stratified G study, the variability due to strata was non-negligible: the person × strata interaction accounted for 8.54 percent of total variability, *pso* accounted for 7.97 percent, and *psom* for 1.87 percent. Assuming one occasion, one method, and two tasks per strata, we calculated that the total number of tasks (N_t) needed for a reliability of .80 would be 27 in the un-stratified G study, and 38 in the stratified design![4]

Before over-interpreting this finding, recall that the original study was not intended to be a stratified design. A long line of research suggests that somewhere between 6 and 20 tasks may be needed to get reliable performance-assessment scores for individual students, depending on the nature of the domain and tasks (Ruiz-Primo & Shavelson, 1997).

If the data from this one, rare, retest G-study are representative of the relative magnitudes of the task and occasion sampling variability in performance assessment scores, and we believe they are (e.g., McBee & Barnes, 1998), stratifying tasks will not solve the task/occasion sampling problem. Stratification will not completely reduce the magnitude of the $p \times t$ interaction because it is confounded with the $p \times t \times o$ interaction. Even with stratification *pt,pto* will still be large. Indeed, in our recent research (Solano-Flores, Jovanovic, Shavelson, & Baxter, in press) we found that stratification of tasks based on an item shell did not reduce task/occasion sampling variability. The most reasonable recourse for addressing the magnitude of task-sampling error is to substantially increase the number of tasks included in an assessment (i.e., dividing $p \times t$ and $p \times t \times o$—i.e., *pt,pto*—by n_t).[5] Stratifying tasks to reduce the person × task variability will only minimally, at best, address the large *pto* component.

Exchangeability: Convergence of Measurement Methods

Due to cost considerations (Stetcher & Klein, 1997) and definitional ambiguity (Baxter & Shavelson, 1994), a variety of methods have been called science "performance assessments." Such methods include hands-on investigations using actual laboratory apparatus (either directly observed by a judge or indirectly evaluated from a student's notebook), computer simulation of the very same investigation, and pencil-and-paper tests that approximate the hands-on version (short-answer and even multiple-choice). We ruled out paper-and-pencil tests as performance assessment on conceptual grounds. They do not react to the actions taken by the student conducting an investigation, and they are only partially consistent with current reform that places hands-on investigations at the center of the curriculum. A sole diet of paper-and-pencil testing methods would send the wrong signal to students and teachers. Nevertheless, we have included them in our research to see if they

Table 4
Correlations Among Measurement Methods*
(based on Shavelson, Baxter & Gao, 1993, Table 6, p. 229)

	Electric Mysteries			Bugs		
	DO	NB	CS	DO	NB	CS
NB	.84			.71		
CS	.55	.52		.44	.49	
SA	.53	.48	.42	.38	.30	.35

*DO=Direct Observation, NB=Notebook, CS=Computer Simulation, and SA=Short-Answer.

could be distinguished from performance assessments on empirical grounds as well.

Shavelson et al. (1993) re-examined the Ruiz-Primo et al. (1993) data for convergence of four assessment methods: direct observation (DO) of hands-on performance, student notebook (NB) reporting on a hands-on investigation, computer simulation of the investigation, and short-answer about the same (but not hypothetical) investigation. (The multiple-choice scores were too unreliable to include in their analysis.) Using data from 186 fifth and sixth grade students' performance on two tasks (electric mysteries and choice behavior or sow bugs) with all four methods, they found higher correlations between scores based on DO and NB ($r = .84$ for electric mysteries and .71 for bugs) than on any other pairing of methods (Table 4). Indeed, except for the DO-NB correlations, the other correlations clustered around .52 for electric mysteries and .38 for bugs. These and other data led Shavelson et al. to conclude that " . . . at best for the data reported here, student performance is dependent on the methods sampled. Methods do not converge" (p. 230).

Actually, the convergence picture is even bleaker than they realized. The strongest convergence evidence was for DO and NB. However, unlike the other methods, these two methods were based on the very same investigation: a student conducted a hands-on investigation while two observers evaluated performance in real time and the student recorded her investigation in a notebook. Both the computer simulation and short answer methods were spaced about one month apart from each other and DO and NB. The DO-NB correlations, then, are presumably raised by *po* and *pto* components.

Retest data collected by Ruiz-Primo et al. (1993) on a random sample of 29 students from the larger (186) sample permits us to calculate the DO-NB correlation with performances separated by several months. For electric mysteries, we found $r_{\text{DO1,NB2}} = .48$ and $r_{\text{DO2,NB1}} = .56$; for bugs, $r_{\text{DO1,NB2}} = .48$ and $r_{\text{DO1,NB1}} = .57$. Averaging the two estimates, we get $r_{\text{DO,NB}} = .520$ and .525 for electric mysteries and bugs, respectively.[6] These correlations jibe with those between computer simulation and both DO and NB, and somewhat less so for short answer.

Disattenuating these correlations in a multivariate G study, Gao, Shavelson, Brennan, and Baxter, (1996) showed a pattern of high correlations among the direct observation, notebook, and computer simulation methods but not short-answer or multiple-choice, even when DO and NB were based on the same investigation.

In summary, direct observation, notebook and computer simulation—all methods that react to the actions taken by students—converge to about the same degree. Nevertheless, student performance across methods is unstable.

Conclusions

Cronbach et al. (1997) were quite right to point out that, in typical science performance situations, task-sampling and occasion-sampling variance is confounded. We have presented empirical evidence on the magnitude of this confound. The magnitude is large; the task-occasion confound (person × task × occasion) is larger than the task-sampling variability (person × task) in the data we analyzed. This finding has important practical consequences in large-scale assessment: increasing task samples will reduce measurement error, and stratification of tasks (because of the occasion confound) won't.

The importance of occasion sampling led us to re-evaluate our findings on the exchangeability of performance-assessment measurement methods. Originally we had concluded that only direct observation and notebook methods were exchangeable. The evidence we now have, evidence that addresses the task-occasion confound, leads us to conclude that: (a) direct observation, notebook, and computer simulation are equally exchangeable, (b) paper-and-pencil methods are not exchangeable for performance assessments, and (c) student performance over occasions is volatile. The volatility of student performance, perhaps arising from partial knowledge of the science they are applying in performance assessment, limits the exchangeability of any of the methods. One possible implication of this finding is that multiple assessments are needed to gauge performance, perhaps using a mix of methods to triangulate on a student's performance (Gao, Shavelson, & Baxter, 1994).

Notes

[1] Another benefit of this three-facet (task, occasion, rater), crossed design is that the estimated variance components can be used to evaluate other, plausible testing situations. For example, one might argue that each time a student encounters a task, a unique occasion occurs. The variance estimates from the three-facet design can be combined to examine a design with occasion nested within task. We are thankful to Ed Haertel for pointing out this alternative use of the three-facet design. This complexity is not introduced in the paper; we can make our point without it.

[2] Lee Cronbach pointed out that even if the judges evaluate performance at the same time, the requirement that they do so independently creates two occasions.

[3] Actually Cronbach et al. were referring to Table 3 in Shavelson et al. (1993), not Table 1 in Gao, Shavelson, and Baxter (1994).

[4] For details of alternative decision-study designs, contact the first author.

[5] Of course, more than one occasion could be sampled as well. But this is very expensive and unlikely to occur in practice.

[6] In this small sample, the DO-NB correlations were .88 and .90 at time 1 and 2, respectively, for electric mysteries, and .72 and .73 for bugs, respectively. These correlations are very close to those based on all 186 students (Table 2).

References

Baxter, G. P., & Shavelson, R. J. (1994) Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research, 21*, 279–298.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373–399.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994) Generalizability of large-scale performance assessments in science: Promises and Problems. *Applied Measurement in Education, 7*, 323–342.

Gao, X., Shavelson, R. J., Brennan, R. L., & Baxter, G. P. (1996) "A Multivariate Generalizability Theory Approach to Convergent Validity of Performance-Based Assessment." Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY, April 9, 1996.

McBee, M. M., & Barnes, L.L.B. (1998). The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education, 11*(2), 179–194.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*(1), 41–53.

Ruiz-Primo, M. A., & Shavelson, R. J., (1997). Rhetoric and reality in science performance assessments: An update. *Journal of Research on Science Teaching, 33*(10), 1045–1063.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215–232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991) Performance assessment in science. *Applied Measurement in Education* (Special Issue) (R. Stiggins and B. Plake, Guest Editors), *4*(4), 347–362.

Solano-Flores, G., Jovanovic, J., Shavelson, R. J., & Bachman, M. (in press). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education.*

Stetcher, B. M. & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19*(1), 1–14.

Authors

RICHARD J. SHAVELSON is the I. James Quillen Dean, Stanford University School of Education and Professor of Education and (by courtesy) Psychology, Stanford University, Stanford, CA 94305-3096; richs@leland.stanford.edu. *Degree:* PhD, Stanford University. *Specializations:* cognitive science, psychometrics, analysis of performance and cognition, teacher use assessment and policy issues.

MARIA ARACELI RUIZ-PRIMO is Research Associate, School of Education, Stanford University, Stanford, CA 94305-3096; aruiz@leland.stanford.edu. *Degree:* PhD, Stanford University. *Specializations:* alternative assessments in science, including performance assessments, concept maps, and student journals.

EDWARD W. WILEY is PhD Candidate in Psychological Studies in Education, Stanford University, Stanford, CA 94305-3096; ewiley@leland.stanford.edu. *Degrees:* BA, Mathematics and MA, Educational Psychology, University of Nebraska-Lincoln. *Specializations:* educational measurement, statistics.