SPECIAL ISSUE
Historical and Contemporary Perspectives on Educational Evaluation

education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 26 Number 48 April 16, 2018 ISSN 1068-2341

Methodological Perspectives: Standardized (Summative) or Contextualized (Formative) Evaluation?

Richard J. Shavelson
Stanford University (Emeritus)
United States

Citation: Shavelson, R. J. (2018). Methodological perspectives: Standardized (summative) or contextualized (formative) evaluation? *Education Policy Analysis Archives, 26*(48). http://dx.doi.org/10.14507/epaa.26.3813 This article is part of the Special Issue, *Historical and Contemporary Perspectives on Educational Evaluation: Dialogues with the International Academy of Education,* guest edited by Lorin W. Anderson, Maria de Ibarrola, and D. C. Phillips.

Abstract: A critical issue in educational evaluation is whether evaluations should focus on standardized (summative, often quantitative) or contextualized (formative or often qualitative) evidence. The author of this article advises readers to beware of false dichotomies. The big issue is not whether evaluations should be "standardized" or "contextualized" but rather whether the evidence collected rigorously addresses the policy and/or practice questions driving the evaluation. The questions asked, in turn, lead to evaluation designs which may be standardized (summative), contextualized (formative) or both. Three general questions drive research and evaluation: (1) Descriptive—What's Happening? (2) Causal—Is there a systematic effect? and (3) Process or mechanism—Why or how is it happening? Depending on the nature of the question, summative and/or formative data might be collected. Equally important are politics, measurement methods

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 1/15/2018
Revisions received: 2/18/2018
Accepted: 2/18/2018

and modeling in conducting evaluations. Ignore these matters at your peril. Concrete examples show how assumptions and misperceptions can upend or change the outcomes of evaluation; they are drawn from political, measurement and statistical modeling contexts.
Keywords: Summative Evaluation; Formative Evaluation; Evaluation Methods; Politics of Evaluation

Perspectivas metodológicas, ¿evaluación estandarizada(sumativa) o contextualizada (formativa)?

Resumen: Un asunto critico de la evaluación educativa es si esta última debiera basarse en evidencia estandarizada (sumativa, a menudo cuantitativa) o contextualizada (formative, a menudo cualitativa). El autor de este artículo aconseja a los lectores cuidarse de las falsas dicotomías. El meollo del asunto no es si la evaluación debiera ser estandarizada o contextualizada, sino más bien si la evidencia recuperada responde rigurosamente a las preguntas de política y/o de prácticas que guían la evaluación. Las preguntas que se formulan, a su vez, dirigen los diseños de evaluación que pueden ser estandarizados (sumativos), contextualizados (formativos) o ambos. Tres preguntas generales conducen la investigación y la evaluación: 1) Descriptiva - ¿qué está pasando? 2) Causal, ¿hay un efecto sistemático? 3) De procesos o mecanismos: Por qué o cómo pasa lo que está pasando. Dependiendo de la naturaleza de las preguntas, se puede recopilar información sumativa o formativa. Igualmente importantes son las políticas, los métodos e instrumentos de medición y los diseños al conducir las evaluaciones. Ignorarlas es asumir fuertes riesgos. Ejemplos concretos obtenidos de contextos políticos, estadísticos y de medición muestran como estos supuestos o falsas percepciones pueden fortalecer o cambiar los resultados de la evaluación.
Palabras-clave: evaluación sumativa; evaluación formativa; métodos de evaluación; políticas de evaluación

Perspectivas metodológicas, padronizadas (somativas) ou contextualizadas (formativas)?

Resumo: Uma questão crítica da avaliação educacional é se esta deve ser baseada em evidências padronizadas (sumativa, muitas vezes quantitativa) ou contextualizada (formativa, muitas vezes qualitativa). O autor deste artigo aconselha os leitores a cuidar de falsas dicotomias. O cerne da questão não é se a avaliação deve ser padronizada ou contextualizada, mas sim se a evidência recuperada responde rigorosamente às questões políticas e/ou práticas que orientam a avaliação. As perguntas formuladas, por sua vez, direcionam os desenhos de avaliação que podem ser padronizados (sumativos), contextualizados (formativos) ou ambos. Três questões gerais orientam pesquisa e avaliação: 1) Descritivo - o que está acontecendo? 2) Causal, existe um efeito sistemático? 3) Processos ou mecanismos: Por que ou como acontece o que está acontecendo? Dependendo da natureza das perguntas, informações somativas ou formativas podem ser coletadas. Igualmente importantes são as políticas, métodos e instrumentos de medição e projetos durante a realização de avaliações. Ignorá-los é assumir fortes riscos. Exemplos concretos obtidos a partir de contextos políticos, estatísticos e de medição mostram como essas suposições ou falsas percepções podem fortalecer ou alterar os resultados da avaliação.
Palavras-chave: avaliação somativa; avaliação formativa; métodos de avaliação; políticas de avaliação

Methodological Perspectives: Standardized (Summative) or Contextualized (Formative) Evaluation?

A critical issue confronting countries such as Mexico is whether education evaluations should focus on standardized (summative, often quantitative) or contextualized (formative or often qualitative) evidence. The advice I offer in this chapter can be summarized as follows: beware of false dichotomies. I argue that the big issue is *not* whether evaluations should be "standardized" or "contextualized" but rather whether the evidence collected rigorously addresses the policy and/or practice questions driving the evaluation. The questions asked, in turn, lead to evaluation designs that may be standardized (summative), contextualized (formative) or both. Three general questions drive research and evaluation: (1) Descriptive—"What's happening?" (2) Causal—Is there a systematic effect? and (3) Process or mechanism—Why or how is it happening? Depending on the nature of the question, summative and/or formative data might be collected. Equally important are politics, measurement methods and modeling in conducting evaluations. Ignore these matters at your peril. Concrete examples show how assumptions and misperceptions can upend or change the outcomes of evaluation; they are drawn from political, measurement and statistical modeling contexts.

Initial Reactions

Methodological issues—the choice between standardized (summative, often quantitative) or contextualized (formative, often qualitative) evaluation—were considered to be key in the evaluation of basic education, at least for the Dialogues with the International Academy of Education. My immediate reaction was: beware of false dichotomies. This seemed obvious. Evaluations should be driven by the policy or practice question(s) that generated the need for evaluation in the first place and not the particular set of methods used in carrying out the evaluation. But of course I was naïve.

I was vaguely aware of a co-occurring event—a teacher strike—affecting all of Mexico. Even though I was preparing for the symposium at The National Institute for Educational Evaluation (INEE) I didn't connect the symposium and teacher strikes at the time (September 1-4, 2016). But the teacher strike surfaced in vivid detail while at the Symposium:

> The strike was launched in May to ramp up the union's rejection of the government's education reform, introduced by Peña Nieto [Mexican President] in 2013, on the basis that the policies threaten public education with creeping privatization and fail to respond to education needs of rural and Indigenous students.
>
> After a marathon session inside the "Ernesto Che Guevara" auditorium, teachers affiliated with the militant National Coordinator of Education Workers in the southern Mexican state of Chiapas voted Thursday to accept a government proposal and end their strike and return to classes.
> [https://dorsetchiapassolidarity.wordpress.com/2016/09/18/teachers-in-chiapas-mexico-vote-to-end-strike/ Retrieved 12/27/16].

One of the major teacher concerns was that summative evaluation had not and was unlikely to reveal the difficult teaching conditions among the poorest schools in the country. The government was asking a different question than the teachers were asking: How might education be run more efficiently and effectively from a distant vantage point (versus what is the impact of distal policies on the poorest students' access to quality education).

In what follows I argue that evaluations should be driven by the nature of the question that gave rise to the need for evaluation and not the particular method to be used in the evaluation: Method should follow from the question. The government's question focused on economics and effects; the teachers' question focused on the human condition of their students and what it takes to educate them. Hence the debate: "Standardized or Contextualized Evaluation?" I then turn to "what matters"—politics, measurement and modeling, and the perceptions and assumptions underlying each. Concrete examples are used to make the case context and assumptions underlying evaluation matter a great deal.

Methods and Questions

In Gilbert and Sullivan's *The Mikado*, the emperor, the Mikado, sings of his virtues: "A more humane Mikado never/Did in Japan exist." One such virtue, he sings, is: "My object all sublime/ I shall achieve in time—/To let the punishment fit the crime/The punishment fit the crime". The point I mean to make here parallels the Mikado: To let the evaluation method fit the question; the method fit the question. In *Scientific Research in Education* (Shavelson & Towne, 2002) we succinctly described three different questions that drive both research and evaluation: (a) Descriptive—"what's happening"; (b) causal—"is there a systematic effect (viz. what is the cause of what's happening)?" and (c) mechanism—"how or why is it happening?"

Description: What's Happening?

The question of what's happening—perhaps the most pressing question for Mexico's teachers—calls for detailed description of a particular situation or event. As Yogi Berra, the catcher and "bard" for the New York Yankees baseball team, once said, "If you want to know what's going on, you have to go out and look at what is going on." *Formative evaluation* does; it is often used to address descriptive questions. Descriptive questions invite both quantitative and qualitative methods so as to, for example, characterize a population, characterize the scope and severity of a problem from various viewpoints, develop a theory or conjecture, or track change over time. Descriptive questions can also include associations among variables, such as school characteristics (size, location, economic base) that are related to (say) the provision of music and art instruction.

Numerous methods can be used to address descriptive questions ranging from detailed ethnography, to case study, to observation, to interviews, to a probability survey, to descriptive statistics, to statistical comparisons of groups, to statistical estimates of relationships (e.g., socio-economic status and school achievement). In the USA, the National Assessment of Educational Progress ("the Nation's Report Card") is perhaps the best example of descriptive standardized (summative, quantitative) education evaluation.

Often in evaluation the questions are such that "mixed methods" need to be used to address them. For example, in the late 1970s Holland and Eisenhart (1990) asked why so few women who entered college in nontraditional majors (e.g., science, mathematics and engineering) ended up in those majors and careers. At the time, several possible explanations were under consideration: (1) lack of adequate preparation for that major, (2) discrimination against women, and (3) aversion to competition with men. They first conducted an ethnographic study of women in nontraditional majors at two residential colleges—one historically black and the other historically white. Volunteer students from each campus, 23 in all, were matched on background (e.g., high-school grade-point-average, major, college activities, and college peers). Half were planning a traditional major; half a non-traditional major. Over a year's time, using participant observation and open-ended interviews, they developed models to describe how the women participated in campus life. The models showed

three different kinds of commitment to school work: (1) views about the value of school work, (2) reasons for doing school work, and (3) perceived cost (monetary and time) of doing school work. From each of these models, Holland and Eisenhart predicted what each woman would do immediately after college—continue schooling, get a job in her field or outside her field, get married, etc. At the end of four years and again at the end of three more years, a follow up was conducted with each woman. In every case, the commitment to school work model predicted the women's futures better than precollege preparation, discrimination, or competition.

Causality: Is There a Systematic Effect?

Evaluation designs that attempt to identify systematic effects have as their root intent to establish a cause-and-effect relationship. Summative evaluation is, essentially, about establishing program (causal) effects or impact (e.g., Fu et al., 2016). Causal evaluation is built on both theory and descriptive studies (see above). The search for causal effects cannot be conducted in a vacuum: *ideally a strong theoretical base as well as extensive descriptive information* is in place to provide the foundation for understanding causal relationships. Consequently, for summative evaluation, a program should have gone through a development period (three or more years) and be in consistent running order before testing for causal effects.

In addressing questions of cause, both summative and formative methods can be applied. In general, the summative "gold standard" is the randomized controlled experiment (control and treatment groups with units randomly assigned to condition). When such experiments are not feasible, either logistically or ethically, alternative methods can be used, such as quasi-experiments (pretests, treatment and control; no randomization), longitudinal causal models (multiple waves of data on the same units), instrumental variables, propensity-score matching, and regression discontinuity modeling (e.g., Shadish, Cook & Campbell, 2002). Qualitative methods can also be used to infer causality such as ethnography and multiple case studies.

Perhaps the best-known experiment in the USA was the Tennessee Class-Size Reduction study carried out in the mid-1980s (see Shavelson & Towne, 2002, for a summary). The State Legislature asked if reducing class size would have a positive impact on students' achievement and funded a large-scale experiment to find out. A total of 11,600 elementary school students and their teachers in 79 schools across the state were randomly assigned to one of three conditions: (1) regular class size (22-26 students), (2) regular class with a full-time teacher's aide (22-26 students/two adults), or (3) reduced class size (13-17 students). The experiment began with a cohort of students in kindergarten and ended four years later when they completed third grade and all entered fourth grade in regular size classes. The experiment showed that students in reduced size classes outperformed their peers in either regular-size or regular size with aide classrooms. It also showed the effects to be greatest for minority and inner-city children. And finally, it showed that those students in the reduced-size classes persisted at a greater rate than peers in taking college entrance examinations and in their performance on those examinations (Krueger & Whitmore, 2001). In the end, however, the Tennessee legislature decided *not* to reduce class sizes in the state because it would be too costly!

Mechanism: How or Why Did It Happen?

Perhaps the ultimate (if largely unattainable) goal of program-impact evaluation is to explain the observed effect with one or more causal mechanism(s)—mechanism(s) that give rise to the effect. To see the importance of mechanisms, consider the case of cigarette smoking and cancer. Legislative and legal battles were fought over the question of whether smoking caused lung cancer. Plenty of studies had established a correlation between smoking and cancer but it wasn't until the

biological mechanism was found that the legislative and legal case was closed and, as they say, the rest is history (https://www.ncbi.nlm.nih.gov/books/NBK53010/, retrieved 12/27/2016).

The mechanism question has stymied evaluators in explaining the causal impact of class-size reduction in the Tennessee study. One possible explanation is that teachers "teach better" giving individual students more attention than in regular-size classes. But observational evidence shows that teachers do not change the way they teach in reduced-size classes. Another possible explanation is that students behave better; troublemakers are found out sooner. This may possibly contribute to the effect initially. A third explanation is that students may be more engaged in learning because, once again, they cannot hide. And so it goes; the search for mechanisms continues.

Both standardized and contextual evaluations can be deployed to address the mechanism question. For example, observational studies have been conducted to test the idea that students behave better and are more attentive to teaching in reduced-size classes than in regular size classes. Small experiments have been tried in training teachers to attend to individual students in their classes.

Closing Thoughts on Questions and Methods

There is no one "right" method; the adequacy of method depends on the question it is intended to address. Often more than one method is needed to fully understand the impact of a program or policy. As Lee J. Cronbach once told me as we discussed *Scientific Research in Education*, randomized experiments are nothing more than single- or multi-site case studies. Bring in a new site, or study the same site years later, and different conclusions might be reached. Caution is needed. Until we have a better handle on why a particular program or policy works *in what contexts*, generalization and transfer are problematic. I caution humility.

What Matters: Politics, Measurement and Modeling

Myriad things matter in carrying out education evaluations. The three things that I have found most impactful are politics, measurement, and modeling. Politics matter. Whatever the object of evaluation, the evaluation is embedded in multiple contexts. When policy on a large scale is the object of evaluation such as education policy in Mexico, politics matter a great deal. Ignore politics at your peril. Moreover, measurement matters. Whatever the target (construct) of interest, different ways of measuring may produce different results—reliability, validity and utility must be aligned with the measurement's intended purposed. Modeling matters. Different ways that standardized (quantitative) and contextualized (qualitative) information is modeled to address evaluation questions (especially their underlying assumptions) can produce very different results.

Politics Matter

The teacher strike in Mexico is a vivid example of how politics matter in education evaluation and policy. In part the teachers were concerned about the ways in which their performance was to be measured and evaluated; they questioned the validity of the measurements in their local contexts.

Before arriving in Mexico City, however, I had a different experience in mind. In the 1980s California embarked on a remarkable education reform. The reform was intentionally systemic. It aligned student learning outcomes (with emphasis on inquiry and constructivism) with curricular reform and with assessment-of-learning reform (http://www.cacollaborative.org/sites/default/files/CA_Collaborative_CLAS.pdf, retrieved 1/2/17). The California Learning Assessment System (CLAS) set out to move assessment from multiple-choice testing to performance assessment with high fidelity simulation of doing science,

mathematics and writing. Instead of asking students in science classes to select the most appropriate option for controlling variables on a multiple-choice test, students were asked to carry out hands-on investigations where they had to decide on what variable to vary and what variable to control, how to control the variable, and then interpret the results. CLAS focused on innovative assessment for summative evaluation and set a 10-year horizon for full implementation. In the meantime, a matrix-sampled multiple-choice test that had been in place for years was to serve initially in the assessment and then phased out. CLAS also collected additional formative assessment information with, for example, writing tasks embedded in classrooms. Teachers scored student performance and teachers' scores were moderated to assure a common scale. After10 years' time, CLAS was to be fully implemented with (a) performance assessments, (b) embedded classroom tasks, and (c) teacher-provided additional evidence for evaluative purposes. The old multiple-choice system would be relegated to an audit function to determine large gaps between scores on these tests and the main evidence coming from the innovative system.

The newly elected governor of California had ridden into office in part on the coattails of a promise to the State's citizens that he would get rid of the California Assessment System that produced scores for schools but, due to matrix sampling, did not produce scores for individual students. Voters in California were fed up with students spending time taking tests that did not produce information about how they, as individuals, were doing. The new Governor promised scores for each student and mandated that all students at a grade level take the same multiple-choice test. While this produced individual level scores it narrowed what could be tested and in turn narrowed the curriculum. The Governor said that CLAS had gotten it wrong: it had placed priority on innovative assessment while delaying implementation of common multiple-choice testing. He could not deliver on his promise of scores for each student in the state. So he fired the CLAS director and stopped support for the innovative testing program. The state has, ever since, used individual multiple-choice tests to assess student performance (although with recent reform that may or may not change). In this case politics is interwoven with what assessments were historically used and thought to be understood by the public. So familiarity with well-established methods was also an enemy of change, and this was capitalized on by political forces. *Politics matter; ignore politics at your peril* (cf. McDonnell & Weatherford, 2016).

Measurement and Modeling Matter

What you measure and how you measure it matters. The CLAS multiple-choice science tests and performance assessments measured somewhat different things (constructs) and they differed substantially in how they measured them. These measurement properties matter when scores from these tests are used in models bearing on education evaluation questions. To make my case I use a research and development program in Colombia aimed at estimating colleges' contribution to learning, their *value added*. I then summarize a report colleagues and I did on the use of value added in teacher evaluation.

The Colombian government mandates the use of value-added measures in the evaluation of its colleges and universities. Of particular concern are mostly private institutions serving low-income or low-achieving students. The goal was to have objective data on which to base decisions for accreditation and support. Colombia is in a unique position among countries around the world; it tests all high school graduates with one examination, the SABER 11, and a parallel examination for all college leaving students, the SABER PRO (e.g., Shavelson et al., 2016).

Value added is a fairly simple notion but one that becomes hugely complicated when implemented in practice. Value added is simply the difference between a student's: (a) predicted

college-leaving score (e.g., SABER PRO) based on some pretest score (e.g., SABER 11) and (b) her actual observed score (e.g., on the SABER PRO):

Value added = observed SABER PRO score—predicted SABER PRO score.

Now the complication: it matters which subtest of the SABER 11 and SABER PRO you use in getting the predicted and observed score. Different pretests lead to different interpretations and results of value added. Moreover, if you use more than one subtest of the SABER 11 you change the definition of value added and findings. Finally, if you include other predictors—such as socioeconomic status—the definition of value added changes.

The assumptions underlying the use of value added are daunting. Value-added measures attempt to provide causal estimates of the effect of colleges on student learning. Consequently they make the usual causal modeling assumptions (Holland, 1986; Reardon & Raudenbush, 2009):

- Manipulability: Students could theoretically be exposed to any treatment (i.e., go to any college).
- No interference between units: A student's outcome depends only upon his or her assignment to a given treatment (e.g., no peer effects).
- The metric assumption: Test score outcomes are on an interval scale.
- Homogeneity: The causal effect does not vary as a function of a student characteristic.
- Strongly ignorable treatment: Assignment to treatment is essentially random after conditioning on control variables.
- Functional form: The functional form (typically linear) used to control for student characteristics is the correct one.

These assumptions lead to additional questions such as: (a) What is the treatment and compared to what? If College A is the treatment what is the control or comparison? What is the duration of the treatment (e.g., 3, 4, 5, 6, 6+ years)? What treatment is of interest—teaching-learning adjusting for institutional context effects? Peer effects? (b) What is the unit of comparison? The institution or college or major? If students change institution, college or major what is the comparison? (c) What should be measured—generic skills (e.g., critical thinking) or domain-specific skills (mathematics). How should it be measured (e.g., multiple-choice, short answer, performance assessment). What pretests ("covariates") should be used in the modeling (a parallel test to the outcome? Multiple pretests? Institutional context (e.g., mean pretest scores)?

To illustrate the consequences of a set of decisions that need to be made in value-added modeling, colleagues and I (Shavelson et al., 2016) drew on the performance of over 64,000 students at 168 higher-education institutions in 19 clusters of majors called reference groups (e.g., engineering, law, education). All had taken the SABER 11 with scores on language, mathematics, chemistry and social science. All had taken the SABER PRO with scores on quantitative reasoning (QR), critical reading (CR), writing and English (plus many subject-specific examinations).

Here I focus on the QR scores in value-added modeling (see Shavelson et al., 2016, for additional measures) using SABER 11 mathematics and SABER PRO QR. We estimated value added with a two-level, mixed effects model: Level 1—student within reference group (engineering); Level 2—engineering school model. The individual-level covariate was SABER 11 mathematics; the reference-group covariate was either a measure of mean social-economic status (INSE) or mean SABER 11 mathematics. We estimated three different models. Each model defines value-added somewhat differently:

1. Model 1 is the simplest—the predicted SABER PRO QR score is based only on the SABER 11 mathematics scores. This means that only this student-level covariate is used and context effects are ignored.
2. Model 2 adds mean INSE to Model 1's predictor. Colleges with low INSE are compared with one another and colleges with high INSE are compared with one another.
3. Model 3 adds mean SABER 11 mathematics to Model 1's predictor. Colleges with low-scoring students are compared with one another and colleges with high-scoring students are compared with one another.

The value-added results from these three models are portrayed in Figure 1. In panel A, we see a high correlation between mean SABER 11 mathematics scores and mean SABER PRO QR scores (0.94). To make the impact of this correlation clearer (hopefully) the black dot represents a high intake school and the gray dot represents an average intake school. Colleges that recruit lower mathematics achieving students graduate students with lower QR scores (on average) and colleges that recruit higher mathematics achieving students graduate students with higher QR scores (on average); no surprise. The correlation between mean socio-economic status (INSE) and mean SABER PRO QR, not shown in the figure, was moderate, 0.40. Given this pattern of correlations we would expect a much bigger impact when controlling for mean SABER 11 mathematics than controlling for mean INSE on value-added estimates of college performance. Panels B and C show the relationship between Model 1 and the two different context-effects models: Model –controlling for mean INSE; Model 3—controlling for mean SABER 11 mathematics.

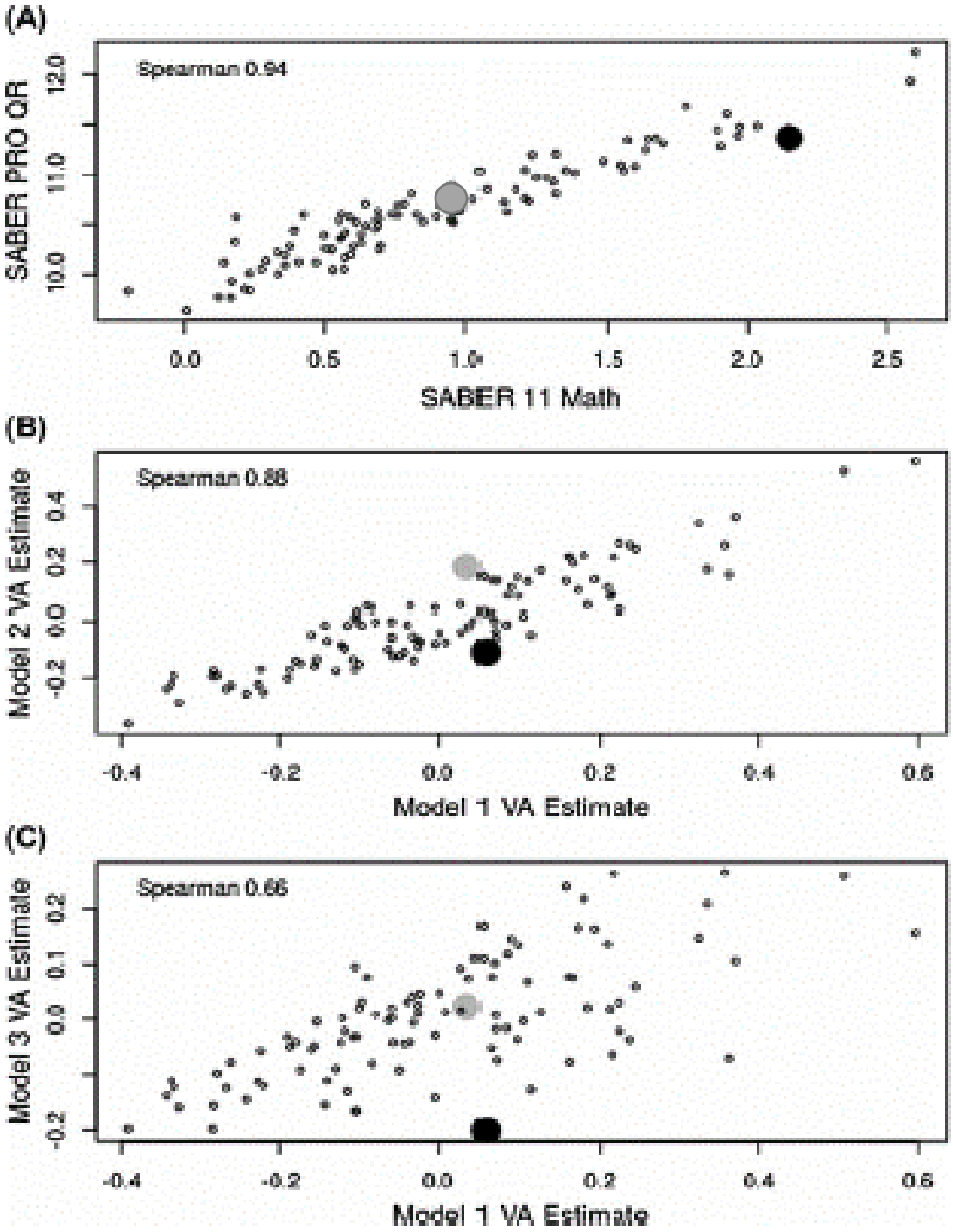

Figure 1. Impact of model specification on the estimate of value added: Panel A—relationship between SABER 11 high-school-leaving mathematics scores and college-leaving quantitative reasoning scores; Panel B—relationship between value added estimates for Model 1 controlling for mathematics scores and Model 2 controlling for both individual mathematics scores and mean socio-economic status; Panel C—relationship between value added estimates for Model 1 controlling for mathematics scores and Model 3 controlling for both individual mathematics scores and mean mathematics scores.

Now watch the dots—they have reversed themselves. The gray dot institution, when compared with its SES peers performs higher than the black dot institution when compared with its peers! In Panel B the correlation between models 1 and 2 is 0.88. We see that controlling for INSE context impacts the value-added estimate for colleges even though the correlation between mean INSE and QR is moderate. This matters especially because the black-dot institution admitted high scoring SABER 11 students and produced high-scoring SABER QR students! While the black dot institution excelled in status (unadjusted outcome), this was not so when context was controlled in value added-modeling. A similar but more drastic outcome is portrayed in Panel C, as expected, when mean SABER 11 mathematics in added as a control.

Models matter! One might say that Model 2 is "fairest." It adjusts for more and less socio-economically advantaged students. However, from a policy perspective, how high should the bar be set for any school? Using a lower bar for low SES schools than high SES schools raises important policy questions.

OK, so one might say that Model 3 is "fairest" because it controls for cognitive "intake quality" and compares like with like. However, colleges are or would like to be selective in student intake. They carefully put together intake cohorts recognizing that peers are important in teaching and learning. Should these schools be penalized for this policy?

Measurements also matter. In the example above, we used the generic skill, QR, as an outcome in the value-added models. If we evaluated institutions on critical reading, for example, a somewhat different picture of value added would emerge (see Shavelson et al., 2016).

If we turn to more domain-specific measures the findings are only slightly different in Colombia (but not for teacher-evaluation in precollege). For example, in using examinations in law and education as outcomes we found that the value-added estimates differed little from the QR results. But the domain-specific measures produced greater variation among colleges.

In a review of research on the use of value added to evaluate precollege teachers, Baker et al. (2015) found that:

- Value-added model estimates are unstable across statistical models and the particular achievement measure used, from one year to the next, and across the classes that a teacher teaches.
- Multiple factors impact student learning gain scores *within* schools that cannot adequately be disentangled:
    - Current teacher effects depended on students' previous teachers
    - School conditions influenced estimates (e.g., peers, leadership, teacher support, curricular quality, tutoring, class size)
    - Out-of-school conditions influenced estimates (e.g., neighborhoods, social capital)
- Multiple factors impact student learning gains across schools even more.

To sum up, what I have attempted to show is that measurement and models matter. What is measured and how it is measured impacts, in significant part, what is found; change the measurement and findings may change. Moreover, the choice of statistical model impacts what is found. Models come with a host of assumptions and critical decisions in the modeling process. The assumptions of the model may be problematic (causal claims may not be warranted); the decision about what variables to include in the model impact the meaning of the results (e.g., the definition of value added). No model is the "right" model; some are more useful than others in specific contexts. Be careful.

Conclusions

The question of whether to use standardized (summative, quantitative) or contextualized (formative, qualitative) evaluation in education was simply put at the outset of this paper. The question turned out to be quite complicated when context is taken into account, as it should be in most all evaluations of education. The admonition to beware of false dichotomies still holds but context matters. Evaluation methods should not drive the evaluation. Rather the questions that gave rise to the evaluation should drive the design and conduct of the evaluation. Moreover, and stated again, the *evaluation must be sensitive to context.*

Politics, measurement methods and modeling all matter in conducting an evaluation. Ignore the politics and context surrounding an evaluation at your peril. Measurement methods matter. What you measure and how you measure it will have a huge impact on what you find as "answers" to evaluation questions. Moreover, models matter. Seemingly simple and reasonable models come with many unseen decisions and assumptions. Changing the model, as we saw, changes the outcome and conclusions drawn. Be careful and transparent in using indicators such as value added from statistical (and other) models.

References

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L.A. (2015). *Problems with the use of student test scores to evaluate teachers.* Washington, DC: Economic Policy Institute (www.epi.org).

Fu, A. C., Kannan, A., Shavelson, R. J., Peterson, L., & Kurpius, A. (2016). Room for rigor: designs and methods in informal science education evaluation. *Visitor Studies, 19*(1), 12-38.

Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81*, 945–960.

Holland, D. C., & Eisenhart, M. A. (1990). *Educated in romance: Women, achievement and college culture.* Chicago: University of Chicago Press.

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results. Evidence from Project STAR. *Economic Journal,* 111, 1-28.

McDonnell, L. M., & Weatherford, M. S. (2016). Recognizing the political in implementation research. *Educational Researcher,* 45(4), 233-242.

Reardon, S. F., & Raudenbush, S. W. 2009. Assumptions of value-added models for estimating school effects. *Education Finance and Policy 4*(4): 492–519.

Shaddish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* NY: Houghton Mifflin.

Shavelson, R. J., Domingue, B. W., Mariño, J. P., Molina-Mantilla, A., Morales, J. A., & Wiley, E. E. (2016). On the practices and challenges of measuring higher education value added: The case of Colombia. *Assessment and Evaluation in Higher Education, 41*(5), 695-720.

Shavelson, R. J., & Towne, L. (Eds.) (2002). *Scientific research in education.* Washington, DC: National Academy Press.

About the Author

Richard J. Shavelson
Stanford University (Emeritus)
richs@stanford.edu
Richard J. Shavelson is the Emeritus Margaret Jacks Professor and I. James Quillen Dean of the Graduate School of Education, Professor of Psychology (Courtesy), and Senior Fellow in the Woods Institute for the Environment, at Stanford University. He served as president of the American Educational Research Association, is a member of the National Academy of Education and the International Academy of Education, and is a fellow of the American Association for the Advancement of Science, American Educational Research Association, American Psychological Association, American Psychological Society and the Humboldt Society (Germany). His current work includes assessment of undergraduates' learning including the Collegiate Learning Assessment, accountability in higher education, and international performance assessment of learning. His publications include *Statistical Reasoning for the Behavioral Sciences, Generalizability Theory: A Primer* (with Noreen Webb), *Scientific Research in Education* (edited with Lisa Towne), and *Assessing College Learning Responsibly: Accountability in a New Era* (2010, Stanford University Press).

About the Guest Editors

Lorin W. Anderson
University of South Carolina (Emeritus)
anderson.lorinw@gmail.com
Lorin W. Anderson is a Carolina Distinguished Professor Emeritus at the University of South Carolina, where he served on the faculty from August, 1973, until his retirement in August, 2006. During his tenure at the University he taught graduate courses in research design, classroom assessment, curriculum studies, and teacher effectiveness. He received his Ph.D. in Measurement, Evaluation, and Statistical Analysis from the University of Chicago, where he was a student of Benjamin S. Bloom. He holds a master's degree from the University of Minnesota and a bachelor's degree from Macalester College. Professor Anderson has authored and/or edited 18 books and has had 40 journal articles published. His most recognized and impactful works are *Increasing Teacher Effectiveness, Second Edition*, published by UNESCO in 2004, and *A Taxonomy of Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, published by Pearson in 2001. He is a co-founder of the Center of Excellence for Preparing Teachers of Children of Poverty, which is celebrating its 14th anniversary this year. In addition, he has established a scholarship program for first-generation college students who plan to become teachers.

Maria de Ibarrola
Department of Educational Research, Center for Research and Advanced Studies
mdeibarrola@gmail.com
Maria de Ibarrola is a Professor and high-ranking National Researcher in Mexico, where since 1977 she has been a faculty-member in the Department of Educational Research at the Center for Research and Advanced Studies. Her undergraduate training was in sociology at the National Autonomous University of Mexico, and she also holds a master's degree in sociology from the University of Montreal (Canada) and a doctorate from the Center for Research and Advanced Studies in Mexico. At the Center she leads a research program in the politics, institutions and actors

that shape the relations between education and work; and with the agreement of her Center and the National Union of Educational Workers, for the years 1989-1998 she served as General Director of the Union's Foundation for the improvement of teachers' culture and training. Maria has served as President of the Mexican Council of Educational Research, and as an adviser to UNESCO and various regional and national bodies. She has published more than 50 research papers, 35 book chapters, and 20 books; and she is a Past-President of the International Academy of Education.

D. C. Phillips
Stanford University
d.c.phillips@gmail.com
D. C. Phillips was born, educated, and began his professional life in Australia; he holds a B.Sc., B.Ed., M. Ed., and Ph.D. from the University of Melbourne. After teaching in high schools and at Monash University, he moved to Stanford University in the USA in 1974, where for a period he served as Associate Dean and later as Interim Dean of the School of Education, and where he is currently Professor Emeritus of Education and Philosophy. He is a philosopher of education and of social science, and has taught courses and also has published widely on the philosophers of science Popper, Kuhn and Lakatos; on philosophical issues in educational research and in program evaluation; on John Dewey and William James; and on social and psychological constructivism. For several years at Stanford he directed the Evaluation Training Program, and he also chaired a national Task Force representing eleven prominent Schools of Education that had received Spencer Foundation grants to make innovations to their doctoral-level research training programs. He is a Fellow of the IAE, and a member of the U.S. National Academy of Education, and has been a Fellow at the Center for Advanced Study in the Behavioral Sciences. Among his most recent publications are the *Encyclopedia of Educational Theory and Philosophy* (Sage; editor) and *A Companion to John Dewey's "Democracy and Education"* (University of Chicago Press).

SPECIAL ISSUE
Historical and Contemporary Perspectives on Educational Evaluation

education policy analysis archives

Volume 26 Number 48 April 16, 2018 ISSN 1068-2341

Please send errata notes to Audrey Amrein-Beardsley at Audrey.beardsley@asu.edu

Join EPAA's Facebook community at https://www.facebook.com/EPAAAAPE and **Twitter feed** @epaa_aape.