Solano-Flores, G., Jovanovic, R. J., & Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education, 21*(3), 293-315. This is an electronic version of an article published in the International Journal of Science Education Copyright 1999 Taylor and Francis; International Journal of Science Education is available online at http://www.informaworld.com.

On the development and evaluation of a shell for generating science performance assessments

Guillermo Solano-Flores, WestEd, Jasna Jovanovic, University of Illinois, Urbana-Champaign, Richard J. Shavelson, Stanford University, and Marilyn Bachman, Montecito Union School, US

We constructed a shell (blueprint) for generating science performance assessments, and evaluated the characteristics of the assessments produced with it. The shell addressed four tasks: Planning, Hands-On, Analysis, and Application. Two parallel assessments were developed, *Inclines* (IN) and *Friction* (FR). Two groups of fifth graders who differed in both science curriculum experience and socioeconomic status took the assessments consecutively in either of two sequences, IN → FR or FR → IN. We obtained high interrater reliabilities for both assessments, statistically significant score differences due to assessment administration sequence, and a considerable task-sampling measurement error. For both assessments, the magnitude of score variation due to the hands-on task indicated that it tapped a kind of knowledge not addressed by the other three tasks. Although IN and FR were similar in difficulty, they correlated differently with an external measure of science achievement. Moreover, measurement error differed depending on assessment administration sequence. The results indicate that shells can produce reliable assessments, but do not solve the task-sampling variability problem or insure assessment exchangeability. We conclude that future shell research should focus on: (a) increasing shell precision, (b) improving shell usability, and (c) determining what specifications must be provided by the shell to ensure that the assessments generated by different developers are comparable.

Introduction

As policy makers and practitioners push for alternative assessments that promote and evaluate higher-order thinking, the need for effective approaches to assessment development becomes increasingly evident. For example, large-scale assessment programs need effective ways to insure the comparability of performance measures, but their standardization is weaker than with traditional measures of academic achievement (Haertel and Linn 1996). School districts (in the US) also need a means of generating assessments that is similar to those used by their states, but developing high-quality performance assessments (PAs) is a lengthy and costly process (Aschbacher 1991, Nuttall 1992, O'Neil 1992, Shavelson *et al.* 1992, General Accounting Office 1993, Solano-Flores and Shavelson 1997, Stecher and Klein 1997).

Although the need for new test construction techniques (Shavelson *et al.* 1990) and tests that assess procedural skills (Frederiksen 1990) is well recognized, test designers have not been able to construct PAs in a reasonable time period at a reasonable cost. Moreover, published descriptions of assessment development methods are general and do not adequately guide developers (e.g., Baron 1991, Shavelson *et al.* 1991, Wiggins 1992, Stiggins 1994, Brown and Shavelson 1996).

0950–0693/99 $12·00

To address the need for effective assessment development, we have extended the notion of 'item shell', originally created for systematically writing paper-and-pencil items, to PAs. Shells for test items are 'hollow' frameworks whose syntactic structures generate sets of similar items (Haladyna and Shindoll 1989) or templates that specify the characteristics of 'families' or types of problems (Hively *et al.* 1968). In the context of performance assessment, shells can be thought of as blue-prints that provide directions for assessment developers to generate reliable, valid PAs in a short time (Solano-Flores and Shavelson 1997, Shavelson *et al.* 1998). In addition, assuming the same content knowledge, two or more assessments generated with the same shell should be comparable – they should be similar in both appearance and psychometric properties.

However, developing and using PA shells is not as simple as developing and using paper-and-pencil item shells. First, PAs address more complex skills than those usually addressed by paper-and-pencil items – the scores obtained by students are intended to reflect the quality of their approaches to solving problems and the quality of their reasoning and conceptual understanding (Baxter *et al.* 1994). Second, the administration and scoring of PAs is complex – it involves pieces of equipment and directions provided to students (see Alberts *et al.* 1986, Alberts *et al.* 1986).

In this paper we describe how we constructed a shell for developing science performance assessments (SPAs) and present findings on the psychometric qualities of the assessments generated with it. In addition to reliability and validity, we examine the comparability of the assessments generated with the shell.

Method

Knowledge domain specification

We used a construct-driven approach for test construction (see Messick 1994). First, we specified a knowledge domain that would allow us to sample a set of tasks. To do so, we created a Guttman-like *mapping sentence* that formalized facets (variables) that are relevant to science assessment, such as type of science task, curriculum, level of inquiry, assessment structure, task sampling, assessment administration, and assessment method.[1]

Any facet may be potentially relevant, depending on assessment purposes. In our case, we were interested in *inquiry level*, which involves 'higher-order' thinking skills (cf. Quellmalz 1985, Raizen and Kaser 1989, Wiggins 1989a,b, Shavelson *et al.* 1990, Shavelson *et al.* 1991), and *task*, which involves the process skills that are common to scientific investigations (Tamir and Glassman 1970, 1971, Tamir 1974). We held constant the other facets in the shell by selecting only one category from each.

As a second step in knowledge domain specification, we limited the scope of the shell to comparative-investigation assessments (Shavelson 1995) in which students conduct an experiment to determine a relationship between objects or variables. *Paper Towels* (Baxter *et al.* 1992) illustrates this type of assessment. It examines the relationship between different brands of paper towels and how much water each towel holds. More specifically, students conduct an investigation to discover which of three kinds of paper towels holds the most water and which holds the least. Performance is scored on the scientific soundness of the procedure used to manipulate, control, and measure variables.

Shell facets used in the study

A. *Task.* A task

> simultaneously requires the use of knowledge, skills and values that are recognized as important in a domain of study and is qualitatively consistent with tasks that members of discipline-based communities might conceivably engage in. (Gitomer 1993: 244).

To recreate activities performed by scientists when they investigate functional relationships, we constructed a shell that would generate science assessments composed of four tasks. Those tasks would be administered in two sections in the following order:

Section 1:

- *Planning and Design* (*Planning*, for short): students are provided with equipment that could be used to investigate a functional relationship and asked to describe how they would do an experiment with the equipment to solve a problem or test a hypothesis.
- *Hands-On Investigation* (*Hands-On*, for short): students are asked to use the equipment provided and conduct an experiment to solve a problem or test a hypothesis.

Section 2 (given to students upon completion of Section 1):

- *Analysis and Interpretation* (*Analysis*, for short): students are given accurate data on the functional relationship and asked to organize the data in a table, graph, or diagram and to draw a conclusion about the relationship investigated.
- *Application*: students are provided with a concrete, meaningful problem context and asked to propose a solution by using part or all of the knowledge generated in the previous tasks.

Administering the assessment in two sections ensures the independence of Hands-On and Analysis (students cannot carry forward errors from Hands-On to Analysis; nor they can go back to change their responses to Hands-On prompts after they have seen the accurate data provided in Analysis). It also ensures standardization – the data used in Analysis are the same for all students. Whereas Planning, Analysis, and Application can be considered conceptual tasks that can be completed with written responses, Hands-On involves conducting an investigation and manipulating equipment. Depending on the inquiry level used (see below) Hands-On may involve the use of procedures according to a set of highly-structured, pre-established directions or may elicit the construction of complex problem solving strategies.

B. *Inquiry Level.* We defined inquiry level by the characteristics of equipment provided, the number of variables to be considered, the amount of conceptual information provided, and the directions given to the student on how to use the equipment. We devised four inquiry levels for each task – no inquiry, low, medium, and high (see example in table 1). The assessments used in this study were developed at the medium inquiry level.[2]

Table 1. Shell: Hands-On Task.

Step	*Inquiry level* — *None*	*Low*	*Medium*	*High*
1	Provide preparatory knowledge in one of three ways: –Written instruction –Illustration with related task –Illustration with embedded task.	Provide preparatory knowledge in one of three ways: –Written instruction –Illustration with related task –Illustration with embedded task.	Provide preparatory knowledge in one of three ways: –Written instruction –Illustration with related task –Illustration with embedded task.	Introduce the concepts that will be used in the assessment.
2	Pose a problem or a hypothesis involving one relevant independent variable.	Pose a problem or a hypothesis involving one relevant independent variable.	Pose a problem or a hypothesis involving one relevant independent variable.	Pose a problem or a hypothesis involving one relevant independent variable (A) and one irrelevant independent variable (B).
3	Do and explain manipulations and measurements.	Provide equipment – include independent variable. Introduce variable name.	Provide equipment – include independent variable. Introduce variable name.	Provide equipment – include independent variable A and independent variable B. Introduce variable names.
4	Ask students to watch.	Tell students which manipulations should be done and how they should be done.	Ask students to solve the problem or test the hypothesis.	Ask students to solve the problem or test the hypothesis.
5	Ask students to report manipulations, measurements, and results. Provide table/chart.	Ask students to solve the problem or test the hypothesis.	Ask students to report manipulations, measurements, and results. Provide table/chart.	Ask students to report manipulations, measurements, and results.
6	END	Ask students to report manipulations, measurements, and results. Provide table/chart.	END	END
7		END		

Shell representation

The shell consisted of a table with four columns, each prescribing a sequence of actions assessment developers should take to create the tasks and response formats to assess the four skill areas at one of four levels of inquiry. The shell for the Hands-On task is shown in table 1.

Our research team used the shell to generate two standardized PAs in physics for use across the State of California. The Science Framework for California Public Schools (California Department of Education 1990) identifies six 'major themes of science' – Energy, Evolution, Patterns of Change, Scale and Structure, Stability, and Systems and Interactions – as the principal focus of a science curriculum. These major themes are unifying constructs that 'link the theoretical structures of the various scientific disciplines' and show the interrelationships of different facts and ideas in science. From the theme titled, *Patterns of Change and Systems and Interactions*, we randomly selected the concept, 'Force and Motion' for fifth-grade physical science, which addresses the notion of force and its relation to motion from the standpoint of classical mechanics. The topics, Inclines and Friction, were sampled as representative of 'Force and Motion'. We used the shell to generate two assessments, *Inclines* (IN) and *Friction* (FR), at a medium level of inquiry.

Despite the differences in the equipment used (figure 1), the Planning, Hands-On, and Analysis tasks had remarkably parallel structures. IN addressed the relation between the inclination of a plane and the force needed to move an object to the top of that plane; FR addressed the relation between surface texture and the force needed to move an object across that surface (table 2). This parallelism, however, did not hold for Application. For FR, this task consisted of proposing a solution to a problem, whereas for IN it consisted of choosing between two conflicting situations. An examination of the shell revealed that this lack of similarity occurred because the directions provided to assessment developers allowed them to choose among a variety of application problem types, with no guidelines on when to use one or another type of problem.[3]

The IN and FR response formats, called 'notebooks', posed problems and provided instructions similarly. For Planning, they provided information and directions on how to set up the equipment. Each task had one or several items related to the task carried out by the students that consisted of: (a) open-ended questions that asked the students to describe a procedure (Planning), explain their reasoning regarding the relationship investigated (Planning), or propose a solution to a practical problem (Application); (b) tables in which the students had to enter the data they obtained (Hands-On); or (c) graphs the students had to complete using a data set provided (Analysis: see figure 2). The notebooks were piloted with students who had completed a 'Force and Motion' unit the previous year. This allowed us to address reading comprehension problems.

Although the shell focused on tasks and response formats, the parallelism was also reflected in the IN and FR scoring forms (figure 3). Indeed, once the scoring system for one assessment was developed, the scoring system for the other was readily constructed.

END

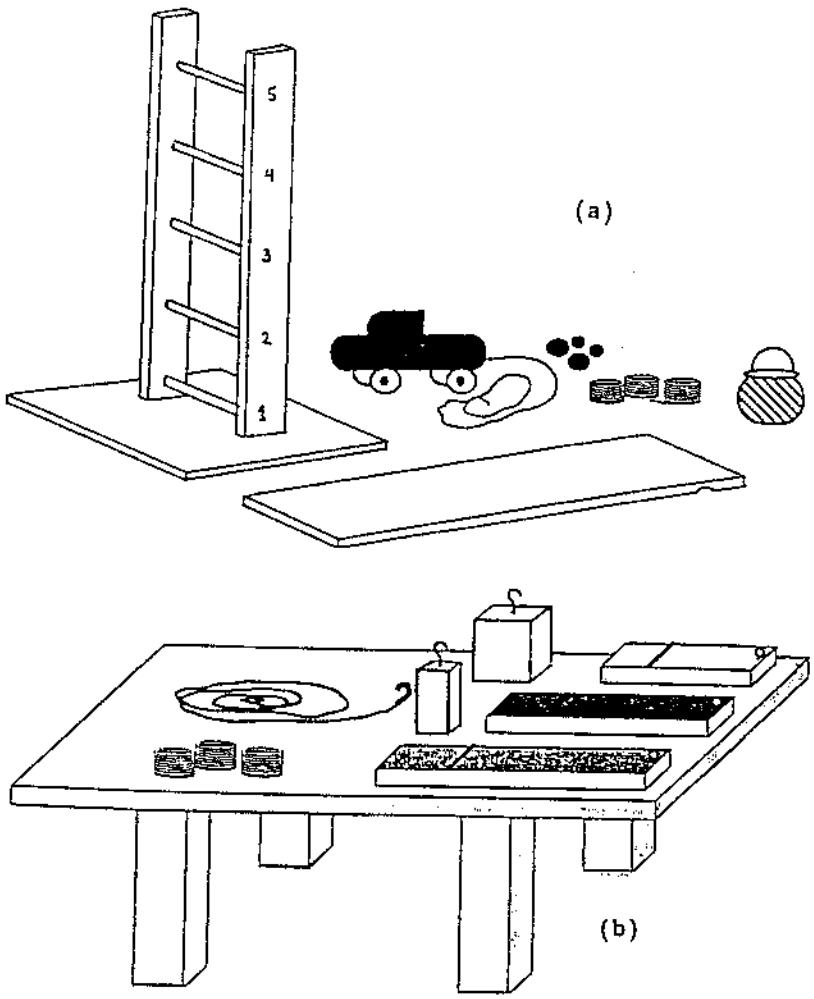

Figure 1. (a) Equipment used in *Inclines*. Students manipulated the weight of the truck by putting marbles in it, and the plane inclination by placing the board at different levels of the ladder. Force was measured by counting the number of washers placed in the bucket to move the truck up the ramp. (b) Equipment used in *Friction*. Students manipulated surface texture (plain wood, felt, sand paper), by varying the boards and selected between two blocks of different weight. Force was measured by counting the number of washers on the big hook needed to move the block across a line on the surface.

Design and Participants

Students from two public schools in California participated in this study. The Science Experienced (SE) students were randomly selected from three classes in a middle to high-income school that emphasized science and used mainly a textbook approach. They had completed a three-week unit, 'Force and Motion', before IN and FR were administered. The Occasional Enrichment (OE) students were randomly selected from two classes in a low-income school that had studied force and

Table 2. Planning task for the assessments developed with the shell.

Inclines	*Friction*
Frank thinks that it will take more force to pull the truck up the incline plane when the incline plane is at a high slope than when it is at a low slope.	Sue thinks that the amount of force needed to pull the block depends on the surface texture of the board. The rougher the surface, the more force she will need to pull the block.
BUT	BUT
Al doesn't think that changing the slope of the incline plane matters. He thinks that it will take the same amount of force to pull the truck up the incline plane at any slope.	Maria doesn't think that changing the surface texture matters. She thinks that the amount of force needed to pull the block will be the same for each board.
Can you think of an experiment you could do to test who is right? You can use the equipment in front of you to figure out how you could design an experiment.	Can you think of an experiment you could do to test who is right? You can use the equipment in front of you to figure out how you could design an experiment.
BELOW, write down the steps you would follow to do your experiment.	BELOW, write down the steps you would follow to do your experiment.

(a)

Look at the results Frank and Al got. How did the amount of force needed to pull the truck change when the slope changed?

(b)

Think about the results Sue and Maria got. How did the amount of force needed to pull the block change when the surface texture changed?

Figure 2. Portions of the *Inclines* (a) and *Friction* (b) notebooks.

(a)

Includes 2 levels of the ladder	
Includes more than 2 levels of the ladder	
Provides a number of washers for every level of the ladder included	
The number of washers increases as the level increases	
Repeats experiment or makes more than one observation per experimental condition	

(b)

Includes all the boards	
Provides a number of washers on the hook for every board	
The number of washers increases as surface roughness increases	
Makes more than one observation for each board with the same block	

Figure 3. Scoring form examples of the *Inclines* (a) and *Friction* (b) assessments.

motion in a curriculum that treated science only occasionally. All students took the assessments on two consecutive days; a randomly determined group completed IN on the first day and FR on the second day; the other group completed the assessments in the opposite order.

For ease of interpretation, we evened the sizes of the cells resulting from the combination of two grouping factors, school (SE vs OE) and assessment sequence (IN → FR vs FR → IN) by randomly discarding cases from the original sample of 109 students until we attained a balanced 2 × 2 design with 64 students – 16 in each of the four cells. Although this approach eliminated almost 40% of the students from the original sample, all the analyses reported here were performed with the original sample of 109 students and produced consistent results.[4]

Administration

To prevent students from carrying forward errors or going back to previous pages to change their answers after seeing accurate data, both IN and FR were administered in two sections: in Section 1, students completed a notebook for Planning and Hands-On; in Section 2, the students completed a notebook for Analysis (where new, accurate data for the investigation were provided) and Application. Students were not given Section 2 until they returned the notebooks for Section 1.

All students completed their investigations and notebooks individually and their scores were computed individually. Due to each school's schedule and space and material constraints, students were tested simultaneously in the same classroom. Although they could see what others did, the tasks were engaging enough to keep them focused on their own investigation.

Prior to the PAs, the students took a 15-item multiple-choice (MC) test on topics related to the concept, 'Force and Motion'. The topics included: energy, force, speed, acceleration, and gravity. The items were scored dichotomously and the test score for each student was computed as the number of items correct. The SE group took the MC test first, and the OE group a few days later (the groups attended different schools; the students could not communicate about the test content). Because the internal consistency of the test taken by the SE group was low (< 0.60), we changed and rewrote some items. Thus, the OE group took a revised version of the test that produced a reasonable internal consistency (0.79). Because of the low internal consistency obtained for the SE group, and due to the fact that the groups took versions of the MC test that differed considerably, our analysis of MC scores will be limited to the OE group (mean = 6.5, s.d. = 9.22). We will not compare the groups as to their performance on the MC test – which was not the intent of the investigation anyway.

Scoring

Scoring was based on students' notebooks, which have proven to be good surrogates for real-time observation (Baxter 1991, Baxter *et al.* 1992). To develop the scoring system, we began by generating a comprehensive set of answers to each 'item' in the notebook. Used as a model of ideal performance, this answer set was divided into a set of essential characteristics. For example, the essential characteristics in the response to an IN Planning item ('Describe the steps you would follow to investigate the relationship between an incline plane's slope and the force needed to pull a truck up the plane') were: placing the truck on the ramp (criterion: using the equipment properly), counting the number of washers in the bucket needed to move the truck (criterion: measuring force), and doing the same operations for at least two levels of the ladder (criterion: manipulating the minimum values of the independent variable necessary to investigate a relationship).

The characteristics of the students' performances, as represented in their notebooks, were scored 1 for 'present', 0 for 'absent' (figure 3); all characteristics were weighted equally. Each task score was computed as the proportion of characteristics rated 1. We also computed a total score by averaging across task scores. The psychometric quality of scores obtained with this simple analytical, compensatory scoring system is comparable to that of scores obtained with other, more sophisticated scoring systems (Solano-Flores 1994, Solano-Flores and Shavelson 1994).

Two raters were trained to use the scoring forms with 20 responses selected randomly from the original sample of 109 students. The raters scored the notebooks independently, then discussed the differences they found, and agreed upon the ways in which the scoring forms should be interpreted. This process was repeated with the same students until an interrater reliability (score correlation) of 0.90, based on independent ratings, was reached. Then, all the student notebooks were scored independently by the two raters before the cells of the design were evened.

Results and discussion

We addressed three questions: (1) How reliable were the assessment generated with the shell? (2) How valid were the interpretations of scores obtained with them? (3) How comparable were they as to reliability and validity? As a part of our analyses, we examined the effects of science curriculum experience and assessment administration sequence on the PA scores. We used $\alpha = 0.05$ in all statistical tests.

Reliability

To estimate the reliability of the scores obtained with IN and FR we used generalizability (G) theory (Cronbach *et al.* 1972, Shavelson and Webb 1991). For each task, we examined the generalizability of scores with a student $(p) \times$ rater(r) design, which enabled us to estimate the relative magnitudes of two sources of measurements error – raters and the residual, and to determine the relative, ρ^2 (norm-referenced), and absolute, ϕ (domain-referenced), generalizability coefficients for IN and FR.

On average, student and rater and the residual accounted, respectively, for 88%, less than 1%, and 12% of the total score variation for IN task scores, and for 80%, less than 1%, and 19% of the total score variation for FR task scores. Also on average, the ρ^2 and ϕ coefficients were 0.93 and 0.92 for IN, respectively, and both were 0.94 for FR. This pattern of variation was similar for total scores; 93, 0, and 7% of the variation of the total IN scores, and 89, 0, and 11% of the variation of total FR scores. As with other SPAs (see Shavelson and Baxter 1992), interrater reliability was not a problem with the shell-generated assessments. Both IN and FR produced similar, dependable scores.

Validity

To examine validity, we asked two questions: (1) Do the tasks within each assessment specify somewhat different aspects of a subject-matter domain? (2) Do the scores exhibit convergent and discriminant validity in a task-by-assessment design?

Knowledge domain

We examined whether the four tasks addressed different kinds of knowledge by examining the score variability due to the task facet using G theory. Rater and assessment were considered random facets, whereas task was considered fixed. While different raters and assessments could have been 'sampled', the four tasks exhausted the types of knowledge addressed by the shell. Following the approach suggested by Shavelson and Webb (1991), we first performed an ANOVA in a student $(p) \times$ rater $(r) \times$ assessment $(a) \times$ task (t) design, treating all sources of variation as random. Then, since the main effect for task was moderate (13%), we examined the scores for each task separately to see whether the patterns of score variability due to rater and assessment differed across tasks. Although these patterns were consistent across tasks, considerable score variability due to the interaction of student and assessment was observed for Hands-On and

Table 3. Estimate variance components (EVC) and percentage of score variation (%SV) in the random student × rater × assessment model for task and total scores.

Source of variation	*Task*									
	Planning		*Hands-On*		*Analysis*		*Application*		*Total*	
	EVC	*%SV*	*EVC*	*%SV*	*EVC*	*%SV*	*EVC*	*%SV*	*EVC*	*%SV*
student (p)	0.025 12	42	0*	0	0.080 97	59	0.017 49	22	0.023 06	64
rater (r)	0.000 40	1	0*	0	0.000 75	1	0*	0	0*	0
assessment (a)	0*	0	0.000 10	0	0*	0	0.000 29	0	0.000 09	0
pr	0.002 05	3	0.000 51	1	0.006 63	5	0.003 72	5	0.000 63	2
pa	0.024 59	41	0.030 53	88	0.037 76	27	0.045 65	57	0.009 41	26
ra	0.000 26	0	0.000 18	0	0*	0	0.000 33	0	0.000 09	0
pra,e	0.008 05	13	0.003 37	9	0.011 65	8	0.012 68	16	0.002 55	7
ρ^2	0.62		0		0.76		0.38		0.80	
ϕ	0.61		0		0.76		0.38		0.80	
1 rater, 1 assessment										
ρ^2	0.37		0		0.59		0.21		0.64	
ϕ	0.36		0		0.58		0.21		0.64	

Note: Some percentages do not sum exactly to 100 due to rounding.
* Small negative value treated as zero.

Application, whereas the same interaction was moderate for Planning and Analysis (table 3). The null main-effect for the facet, student, on Hands-On is due to the small score variance. The considerable student × assessment interaction obtained for Hands-On confirms previous findings that student performance on hands-on assessments varies considerably from one task to another (Shavelson *et al.* 1993). When two raters and two tasks are used, only total scores are reasonably reliable ($\rho^2 = \phi = 0.80$). In a decision (D) study, we found that, if only one rater and one assessment are used, the reliabilities for tasks and the total score are low, ranging from 0 for Hands-On to 0.64 for total scores. Thus, dependable performance scores are obtained only if the four tasks are considered together, using several raters and assessments.

Convergent and discriminant validity

An examination of the correlations between task scores revealed that Hands-On correlated consistently low with the conceptual tasks. The correlations between Planning, Analysis and Application ranged from 0.29 to 0.65, whereas the correlations between conceptual tasks and Hands-On or between Hands-On across assessments ranged from −0.06 to 0.17 (table 4). Based on these findings, we combined the Planning, Analysis, and Application scores into a 'Conceptual' category and constructed a multi-assessment (IN, FR)-multi-score (Conceptual, Hands-On) correlation matrix to examine the validity of interpretations of Conceptual and Hands-On task scores as representing distinguishable aspects of the knowledge domain (table 5). Regarding convergent validity, the correlation between scores of the conceptual task using different assessments was moderately high ($r = 0.71$), but not so for the Hands-On task scores ($r = -0.01$). Regarding discri-

Table 4. Multiassessment (inclines, friction)-multitask (planning, analysis, hands-on, application) matrix. Generalizability (interrater reliability) ρ^2 coefficients in parentheses.

	Inclines				*Friction*			
	Planning	*Hands-On*	*Analysis*	*Application*	*Planning*	*Hands-on*	*Analysis*	*Application*
Inclines								
Planning	(0.89)							
Hands-On	0.00	(0.95)						
Analysis	0.50	0.17	(0.97)					
Application	0.29	−0.06	0.47	(0.92)				
Friction								
Planning	*0.47*	−0.06	0.45	0.38	(0.92)			
Hands-On	0.08	*−0.01*	0.04	0.22	0.03	(0.92)		
Analysis	0.44	0.07	*0.65*	0.40	0.42	0.30	(0.87)	
Application	0.50	0.11	0.53	*0.27*	0.56	−0.04	0.59	(0.83)

Table 5. Multiassessment (inclines, friction)-multitask (conceptual, hands-on) matrix. Generalizability (interrater reliability) ρ^2 coefficients in parentheses.

	Inclines		*Friction*	
	Conceptual	*Hands-on*	*Conceptual*	*Hands-on*
Inclines				
Conceptual	(0.96)			
Hands-On	0.06	(0.95)		
Friction				
Conceptual	*0.71**	0.05	(0.94)	
Hands-On	0.13	*−0.01*	0.15	(0.92)

* Significant at $\alpha = 0.05$; two-tailed, $df = 62$.

Table 6. Correlations with multiple-choice test scores: science occasional enrichment group.

Assessment and task	*Multiple choice*
Inclines	
Planning	0.53*
Hands-On	0.44*
Analysis	0.30
Application	0.38
Total	0.63*
Conceptual (planning, analysis and application combined)	0.51*
Friction	
Planning	0.40*
Hands-On	0.15
Analysis	0.19
Application	0.31
Total	0.39*
Conceptual (planning, analysis and application combined)	0.38*

Note: Six of the 32 students of the occasional enrichment group did not take the multiple choice test.
* Significant at $\alpha = 0.05$, two-tailed, $df = 24$.

minant validity, the correlations between measures of different tasks using the same assessments were low ($r = 0.06$ for IN, $r = 0.15$ for FR); and the correlations between measures of different tasks using different assessments were similarly low ($r = 0.05$ and $r = 0.13$). Thus, the conceptual tasks on the assessments stand up to the test of convergent and discriminant validity; not so for the hands-on task.

Comparability

A. *Correlation with an external measure of academic achievement.* We compared the correlations of IN and FR scores with MC scores, an external measure of academic achievement. Since a reasonable internal consistency (0.79) was attained only for the revised version of the MC test (see *Administration*, above), our discussion of the correlations will be limited to the OE group.

Correlations of both task and total scores with MC scores were consistently higher for IN than FR; the largest difference observed was for the hands-on task (table 6). Even though IN and FR were sampled from the same core concept and generated with the same shell, their scores correlated differently with an external measure of academic achievement.

B. *Sensitivity to group differences.* We used sensitivity to group differences (Crocker and Algina 1986) as another criterion to compare the psychometric qualities of the PAs developed with the shell. We compared the sensitivity of IN and FR to differences between the SE and OE groups. Since curricular experience and socioeconomic status were confounded, a straightforward curricular experience interpretation of the mean differences is impossible. Moreover, even if covariates were available, adjustment might still be questionable (see Shavelson *et al.* 1991). Therefore, these results should not be taken as an indicator of the sensitivity to curricular differences.

Both task and total scores were higher for the SE group. A series of t tests revealed that, except for Hands-On, all the score differences were statistically significant (table 7). Moreover, a science curriculum experience (s) × assessment (a) × task (t) split-plot ANOVA, with s as the grouping factor, revealed statistically significant differences due to the main effects of science curriculum experience and task. No statistically significant differences due to the main effects of assessment were observed. The interaction of science curriculum experience and task also produced significant differences. A series of pairwise post-hoc tests of the interaction using Tukey's method (Shavelson 1996) revealed that the SE and OE groups differed significantly on all conceptual tasks, but not on the Hands-On task. The differences in favour of the SE group on the conceptual task seem to reflect a difference between groups in opportunity to learning science, whereas the absence of differences on Hands-On may be attributable to the fact that neither group had the opportunity to learn in a Hands-On curriculum.

C. *Effect of assessment sequence on student scores.* We compared the score differences produced by two assessment sequences – IN → FR versus FR → IN – to see whether the experience gained from taking the first assessment influenced performance on the second assessment, and whether the effect was the same for both sequences. IN scores for the IN → FR group and FR scores for the FR → IN group were dubbed 'first-take' scores; IN scores for the FR → IN group and FR scores for the IN → FR group were dubbed 'second-take' scores (table 8).

With the exception of Application – which is not comparable across assessments due to the looseness of the directions provided by the shell (see Note 3) – in both assessment sequences the experience gained from taking one assessment influenced favourably students' performance on the second assessment. A three-way split-plot, fixed-effects ANOVA was used to test for the effects of one between-subjects factor, assessment sequence, and two repeated-measures factors, take (first- versus second-take scores) and task. Statistically significant differences were found for the main effects of take and task but not assessment sequence. (Although the interaction of take and task produced statistically significant differences, pairwise comparisons revealed that significant differences between task

Table 7. Mean task and total scores on the inclines (IN) and friction (FR) assessments by science curriculum experience.

Science curriculum experience	*Task and assessment*									
	Planning		*Hands-on*		*Analysis*		*Application*		*Total*	
	IN	*FR*	*IN*	*FR*	*IN*	*FR*	*IN*	*FR*	*IN*	*FR*
Science experienced ($n = 32$)	0.8008	0.8066	0.7656	0.7227	0.9297	0.8698	0.6594	0.5730	0.7889	0.7435
s.d.	0.1921	0.1826	0.0787	0.1871	0.0949	0.1974	0.1583	0.2155	0.0634	0.1334
Occasional enrichment ($n = 32$)	0.6328	0.6113	0.7031	0.6719	0.4323	0.4844	0.3672	0.3508	0.5339	0.5296
s.d.	0.2437	0.2422	0.2148	0.1951	0.3675	0.3638	0.3020	0.2392	0.1732	0.1715
t test ($\alpha = 0.05$)	yes	yes	no	no	yes	yes	yes	yes	yes	yes

Table 8. Mean scores and standard deviations for two assessment administration sciences

Take	*Task and sequence of administration*									
	Planning		*Hands-on*		*Analysis*		*Application*		*Total*	
	IN → FR	*FR → IN*	*IN → FR*	*FR → IN*	*IN → FR*	*FR → IN*	*IN → FR*	*FR → IN*	*IN → FR*	*FR → IN*
First ($n = 32$)	0.7188	0.6641	0.6719	0.6406	0.6693	0.6276	0.5453	0.4570	0.6513	0.5973
s.d.	0.2574	0.2238	0.2129	0.2262	0.3819	0.3727	0.2928	0.2619	0.2058	0.2022
Second ($n = 32$)	0.7539	0.7148	0.7539	0.7969	0.7266	0.6927	0.4688	0.4812	0.6758	0.6714
s.d.	0.2397	0.2112	0.1289	0.0309	0.3226	0.3545	0.2468	0.2693	0.1636	0.1582

scores across takes occurred only between Application and each of the other three tasks.)

We carried out a series of Tukey's post-hoc tests for split-plot designs (see Shavelson 1996) to compare task scores within the same sequence (e.g. Planning-IN and Planning-FR for the IN → FR group) and found that the score gains were statistically significant only for Hands-On. We also carried out a series of post-hoc tests to compare task scores between the IN → FR and FR → IN groups (e.g., Planning-IN, IN → FR group versus Planning-IN, FR → IN group). Again, second-take scores were significantly higher only for Hands-On. Apparently, the hands-on task distinguishes itself from the other, conceptual tasks, in the amount of knowledge students construct in taking the task the first time and use when they take an equivalent task.

D. *Effect of assessment sequence on score variability due to difference sources of measurement error.* We compared the assessment sequences as to the patterns of score variability due to student (p), rater (r), and assessment (a) for each task. These patterns were similar across assessment sequences, except for Hands-On, in which

Table 9. Estimate variance components (EVC) and percentage of score variation (%SV) in the mixed model (t fixed) for different three-task combined scores by assessment administration sequence.

Sequence and source of variation	*Combination of tasks*							
	Hands-On analysis, application		*Planning, analysis, application*		*Planning, Hands-On application*		*Planning, Hands-On analysis*	
	EVC	*%SV*	*EVC*	*%SV*	*EVC*	*%SV*	*EVC*	*%SV*
IN → FR								
student (*p*)	0.027 35	65	0.041 09	71	0.015 81	59	0.025 23	66
rater (*r*)	0.000 14	0	0*	0	0*	0	0.000 04	0
assessment (*a*)	0*	0	0*	0	0*	0	0.001 29	3
pr	0.001 49	4	0.001 68	3	0.001 15	4	0.001 13	3
pa	0.010 41	25	0.012 36	21	0.007 78	29	0.008 16	21
ra	0.000 08	0	0.000 23	0	0.000 28	1	0.000 23	1
pra, e	0.002 41	6	0.002 25	4	0.001 65	6	0.001 98	5
ρ^2	0.80		0.84		0.76		0.83	
ϕ	0.80		0.84		0.76		0.81	
FR → IN								
student (*p*)	0.021 17	47	0.035 99	63	0.011 06	43	0.020 46	50
rater (*r*)	0.000 01	0	0	0	0*	0	0.000 16	0
assessment (*a*)	0.002 86	6	0.000 65	1	0.002 54	10	0.003 71	9
pr	0.000 79	2	0.000 76	1	0*	0	0.000 15	0
pa	0.015 38	34	0.013 97	25	0.007 13	28	0.012 95	32
ra	0*	0	0*	0	0.000 27	1	0*	0
pra, e	0.004 59	10	0.005 41	10	0.004 67	18	0.003 22	8
ρ^2	0.69		0.80		0.70		0.73	
ϕ	0.66		0.79		0.64		0.68	

Note: some percentages do not sum exactly to 100 due to rounding.
* Small negative value treated as zero.

the amounts of score variation due to assessment were different, 29% for the FR → IN group and only 7% for the IN → FR group.

To evaluate the extent to which each task contributed to score variability in each sequence, we ran a series of student (p) × rater (r) × assessment (a) G studies in which one task was excluded at a time. Two facts stand out. First, the highest generalizability coefficients were obtained when Hands-On scores were excluded from the analyses. Second, although the patterns of score variability produced by both assessment sequences were similar, the generalizability coefficients were consistently higher for the IN → FR sequence (table 9).

Conclusions

As with assessments developed with other procedures, we found high interrater reliabilities with shell-generated assessments. We also found that the interaction of student and assessment was the largest source of measurement error. Moreover, the conceptual tasks of the assessments distinguished between students with different characteristics (socioeconomic status and science curriculum experience confounded).

We have learned three important lessons. First, to be capable of generating comparable assessments, the directions provided by shells need to be quite specific. The Application shell provided too much room for interpretation by assessment developers; consequently, the Application tasks produced were not comparable across assessments. The present shell, then, provides a starting place for successive refinements with further experience and research.

Second, even in roughly parallel assessments developed with the same shell, the hands-on task clearly distinguishes itself from the other, conceptual tasks. The sources of score variability affecting the Hands-On task differed from those affecting the conceptual tasks; Hands-On was especially influenced by the student × assessment interaction.

Third, just because two assessments are drawn as samples from the same core concepts and developed with the same shell, they are not necessarily exchangeable. IN and FR posed equivalent problems and had parallel tasks, response formats, and scoring systems; they were developed by the same team and administered on consecutive days. They also had comparable difficulties. However, they correlated differently with an external measure of science achievement and the sequence in which the students took them produced different patterns of score variability for the hands-on task.

The assessments' contextual factors – problems, equipment, variables, wording (see Baxter *et al.* 1992) – and the level of conceptual knowledge and reasoning used – which can vary substantially within the same student (see Hodson 1992), may account for those differences. Moreover, each assessment poses intrinsic cognitive demands: counting the number of rungs on the ladder of Inclines is very different from selecting one board from three boards with different surface textures. Thus, content cannot be dissociated from cognitive processes, even when "sister," shell-generated assessments are used: taking Inclines, then Friction, is not the same as taking Friction, then Inclines. Each assessment sequence entails a different process of knowledge utilization.

Shells, then, can generate reliable science performance assessments with similar looking tasks, response formats, and scoring systems, but they do not ensure assessment exchangeability and, alone, cannot solve the old problem of task sampling variability.

With some knowledge of what can and cannot be expected from shell-generated assessments, we should focus on revising shells to make science performance assessments increasingly exchangeable and more widely accessible to educators. The importance of shell usability, then, should not be underestimated.

By *usability* we mean the shell's ability to guide developers in generating an assessment. The importance of shell usability is illustrated by our experience in an early stage of this investigation, when we asked a fifth-grade science teacher who was unfamiliar with the project to use the shell to generate an assessment on a topic of her choice. We provided her with the shell in the form of flowcharts, which she found difficult to use perhaps because of the practice and special interpretation skills involved in using flowcharts (Krohn 1983). That is the reason why, to make the shell user-friendly, we had to translate it into tables that contained the same information but were easier to interpret.

Since the same team developed Inclines and Friction, it is not a surprise that these assessments were remarkably parallel in appearance. In contrast, in an investigation by the RAND corporation (Stecher *et al.* 1998), two teams of assessment developers worked independently with a common shell to generate assessments on the same topic, Acids and Bases. The assessments generated differed considerably as to the equipment and activities used in the tasks, the layout and complexity of the response formats, and the administration procedures. Clearly, the directions provided by shells need to be very specific so their interpretation is consistent across teams of assessment developers.

Based on that experience, we have increased the specificity of the directions provided by shells. As a part of a series of projects for the assessment and certification for teachers (see Solano-Flores *et al.* 1998), the Science Assessment Development Laboratory at WestEd is currently developing several types of exercises (e.g. problem-solving, pedagogical content knowledge, conceptual knowledge) for each of four science content areas: biology, chemistry, Earth and space science, and physics[5]. We use a shell for each type of exercise with very specific directions for developers. That is, in addition to meeting strict content specifications, all exercises of the same type must have the same structure and comparable complexities both within content area and across content areas. We have observed that exercise comparability across teams of assessment developers and across content areas can be attained if two conditions are met. First, the shell must be highly structured; it must provide not only directions, but also a model of the characteristics of the exercises to be developed (see figure 4). Second, the training of assessment developers must help them realize that content-rich exercises can be developed with shells despite their strict specifications, and given them the opportunity to develop a number of exercises under the guidance of an experienced colleague or a staff member. As a part of the training provided, assessment developers examine examples of exercises that were generated with the same shell and have the same structure, regardless of content area.

To a great extent, when we train assessment developers to use shells, we train them to translate their ideas into the structure specified by those shells. The learning process involved cannot occur overnight. Shells can be used to generate science exercises of comparable characteristics in a short time, but considerable

Shell

The table below shows [information and variables provided] for [description of objects]. Three [objects] [belong to class K1] and the other [objects] [belong to class K2].

(Provide additional information that can be used as a clue to solving the problem.)

	Characteristics			
Objects	A	B	C	• • •
1				
2				
3				
4				
•				
•				
•				

Based on the information provided:
a. Determine which [objects] [belong to class K1].
b. Justify your answer. Describe the procedures and reasonings you used to solve the problem.
c. Discuss the limitations and advantages of your solution.

Exercise Developed with the Shell

The table below shows [the distribution of 16 hypothetical species (A through P)] for [eight cells of a landscape]. Three [cells] (must be selected to be used as reserves for protecting biodiversity] and the other [five cells] {will be used for urban development].

(B, C, E, and F are very rare species.)

	Species															
Cell	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	x	x	x				x					x	x			x
2	x		x				x	x			x			x		x
3	x	x							x			x	x			x
4	x		x		x		x								x	x
5	x									x	x	x	x	x	x	x
6			x	x			x	x			x		x		x	x
7				x	x	x	x			x					x	x
8	x		x		x		x						x		x	

Based on the information provided:
a. Determine which [three cells] [should be used as reserves to protect biodiversity].
b. Justify your answer. Describe the procedures and reasonings you used to solve the problem.
c. Discuss the limitations and advantages of your solution.

Figure 4. A highly-structured shell for developing problem-solving exercises and an exercise developed with that shell. *Objects* are unknowns, phenomena, groups, or entities of any kind. *Classes* are categories to which the objects belong; the examinee must identify to which of two classes objects belong based on the characteristics of the objects. Brackets indicate portions of text that must be replaced by the assessment developer with information specific to the exercise, according to certain content specifications. Parentheses indicate directions for assessment developers. For the purpose of illustration, brackets and parentheses have been kept in the exercise to show how its structure reflects the structure prescribed by the shell. The content of the exercise is an adaptation of a problem described by Pimm and Lawton (1998).

effort and training time must be invested before a reasonable level of efficiency can be attained.

It we improve the design of future shells, large scale assessment systems with teams of assessment developers, working independently, might generate assessments of comparable qualities and characteristics. School districts might use shells to generate performance assessments that suit their assessment needs. If carefully designed shells were provided by a state, assessment comparability aligned with the state's assessment system might be possible across school districts. Schools must develop assessments similar to those used in the accountability system, so students are not surprised by the annual, on-demand assessments.

Acknowledgements

We wish to thank Ed Haertel, Lee Cronbach, Pinchas Tamir, Steve Klein, Brian Stecher, Steve Schneider, the participants at our occasional Stanford sessions, and four anonymous reviewers for their valuable comments on previous versions of this paper.

Notes

1. We are indebted to the RAND project team – Steve Klein, Brian Stecher, and their colleagues – for their contribution to the construction of this mapping sentence. The shells for all the tasks described and complete information on the assessments are available from the first author upon request.
2. Since the rationale for the switch to PAs is the need to address higher-order skills and critical thinking (Wiggins, 1989b), we were not interested in developing assessments with a level of no inquiry. However, we specified a level of no inquiry only for formal, methodological completeness.
3. For example, the shell provided the following directions for the Application task: 'Ask the students to show a product for the solution of the problem, *or* give the steps that led them to the solution, *or* identify the advantages and disadvantages of the solution, *or* suggest possible alternative solutions'.
4. All the analyses reported here were performed with the original sample of 109 students in the unbalanced design and produced consistent results. The results of the unbalanced design can be obtained from the first author.
5. The first author is indebted to his colleagues at WestEd – Steven Schneider, Stan Ogren, Kirsten Daehler, Kristin Hershbell, Jerome Shaw, and Jody McCarthy – for their enthusiastic participation in the development and use of these shells. Also, he is indebted to all the teachers who have acted as assessment developers in the AYA/Science Assessment Development Laboratory and have used the shells to generate science exercises.

References

Alberts, R. V. J., van Beuzekom, P. J. and Hellingham, C. (1986) The development of standardized tests for experimental work in schools in the Netherlands. *European Journal of Science Education,* 8 (2), 135–143.

Alberts, R. V. J., van Beuzekom, P. J. and de Roo, I. (1986) The assessment of practical work: a choice of options. *European Journal of Science Education,* 8 (4), 361–369.

Aschbacher, P. R. (1991) Performance assessment: state activity, interest, and concerns. *Applied Measurement in Education,* 4 (4), 275–288.

Baron, J. B. (1991) Strategies for the development of effective performance exercises. *Applied Measurement in Education,* 4 (4), 305-318.

Baxter, G. P. (1991) Exchangeability of science performance assessments. Unpublished doctoral dissertation. University of California, Santa Barbara.

Baxter, G. P., Glaser, R. and Raghavan, K. (1994) Analysis of cognitive demand in selected alternative science assessments. CSE Technical Report 382 (August). National Center for Research on Evaluation, Standards, and Student Testing.

Baxter, G. P., Shavelson, R. J., Goldman, S. R. and Pine, J. (1992) Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement*, 29(1), 1–17.

Brown, J. H. and Shavelson, R. J. (1996) *Assessing Hands-On Science* (Thousand Oaks, CA: Corwin Press).

California Department of Education (1990) *Science Framework for Public Schools* (Sacramento, CA: CDE).

Crocker, L. and Algina, J. (1986) *Introduction to Classical and Modern Test Theory* (Fort Worth: Holt, Rinehart, and Winston).

Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. (1972) *The Dependability of Behavioral Measurements* (New York: Wiley).

Frederiksen, N. (1990) Introduction. In N. Frederiksen, R. Glaser, A. Lesgold and M. G. Shafto (eds), *Diagnostic Monitoring of Skill and Knowledge Acquisition* (Hillsdale, NJ: Erlbaum), ix-xvii.

General Accounting Office (1993) *Student Testing: Current Extent and Expenditures, With Cost Estimates for a National Examination* (Washington, DC: GAC).

Gitomer, D. H. (1993) Performance assessment and educational measurement. In R. E. Bennett and W. C. Ward (eds), *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment* (Hillsdale, NJ: Erlbaum), 341–263).

Haladyna, T. M. and Shindoll, R. R. (1989) Shells: a method for writing effective multiple-choice test items. *Evaluation and the Health Professions*, 12, 97–104.

Haertel, E. H. and Linn, R. L. (1996) Comparability. In Phillips, G. W. (ed.), *Technical Issues in Large-Scale Performance Assessment* (Washington, DC: National Center for Education Statistics, Department of Education, Office of Educational Research and Improvement).

Hively, W., Patterson, H. L. and Page, S. H. (1968) A 'universe-defined' system of arithmetic achievement tests. *Journal of Educational Measurement*, 5(4), 275–290.

Hodson, D. (1992) Assessment of practical work: Some considerations in philosophy of science. *Science & Education*, 1(2), 115–144.

Krohn, G. S. (1983) Flowcharts used for procedural instructions. *Human Factors*, 25(5), 573–581.

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessments. *Educational Researchers*, 23(2), 13–23.

Nuttall, D. L. (1992) Performance assessment: the message from England. *Educational Leadership*, 49(8), 54–57.

O'Neil, J. (1992) Putting performance assessment to the test. *Educational Leadership*, 48(8) 14–19.

Pimm, S. L. and Lawton, J. H. (1998) Planning for biodiversity. *Science*, 279, 2068–2069.

Quellmalz, E. S. (1985) Developing criteria for performance assessments: the missing link. *Applied Measurement in Education*, 4(4), 319–331.

Raizen, S. A. and Kaser, J. S. (1989) Assessing science learning in elementary school: Why, what, and how? *Phi Delta Kappan*, 70(9), 718–722.

Shavelson, R. J. (1995) On the development of a science performance assessment technology. NAEP paper.

Shavelson, R. J. (1996) *Statistical Reasoning for the Behavioral Sciences* (Boston: Allyn & Bacon).

Shavelson, R. J. and Baxter G. P. (1992) What we've learned about assessing hands-on science. *Educational Leadership*, 49(8), 20–25.

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

Shavelson, R. J., Baxter, G. P. and Pine, J. (1991) Performance assessment in science. *Applied Measurement in Education*, 4(4), 347–362.

Shavelson, R. J., Baxter, G. P. and Pine, J. (1992) Performance assessments: political rhetoric and measurement reality. *Educational Researcher*, 21 (4), 22–27.

Shavelson, R. J., Baxter, G. P., Pine, J., Yure, J., Goldman, S. R. and Smith, (1991) Alternative technologies for large scale science assessment: instrument of education reform. *School Effectiveness and School Improvement*, 2 (2), 97–114.

Shavelson, R. J., Carey, N. B. and Webb, N. M. (1990) Indicators of science achievement: options for a powerful policy instrument. *Phi Delta Kappan*, 71 (9), 692–697.

Shavelson, R. J. and Webb, N. M. (1991) *Generalizability Theory: A Primer* (Newbury Park, CA: Sage).

Shavelson, R. J., Solano-Flores, G. and Ruiz-Primo, M. A. (1998) Toward a science performance assessment technology. *Evaluation and Program Planning* 21, 171–184.

Solano-Flores, G. (1994) A logical model for the development of *science performance assessments*. Unpublished doctoral dissertation. University of California, Santa Barbara.

Solano-Flores, G., Raymond, B., Schneider, S. A. and Timms, M. (1998) Management of scoring sessions in alternative assessment: the computer-assisted scoring approach. (under review)

Solano-Flores, G. and Shavelson, R. J. (1994) Binary-based versus weight-based scoring in Science perfomance assessments. Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA, April, 5–7.

Solano-Flores, G. and Shavelson, R. J. (1997) Development of performance assessments in science: conceptual, practical and logistical issues. (Eng.) *Educational Measurement: Issues and Practice*, 16 (3), 16–25.

Stecher, B. M. and Klein, S. P. (1997) The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis*, 11 (1), 1–14.

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J. and Haertel, E. (1998) The effects of content, format, and inquiry level on performance on science performance assessment scores. (under review)

Stiggins, R. J. (1994) *Student-Centred Classroom Assessment* (New York: Macmillan)

Tamir, P. (1974) An inquiry oriented laboratory examination. *Journal of Educational Measurement*, 11, 25–33.

Tamir, P. and Glassman, F. (1970) A practical examination for BSCS students. *Journal of Research in Science Teaching*, 7, 107–112.

Tamir, P. and Glassman, F. (1971) A practical examination for BSCS students: A progress report. *Journal of Research in Science Teaching*, 8 (4), 307–315.

Wiggins, G. (1989a) Teaching to the (authentic) test. *Educational Leadership*, 46 (7), 41–47.

Wiggins, G. (1989b) A true test: toward more authentic and equitable assessment. *Phi Delta Kappan*, 70 (9), 703–713.

Wiggins, G. (1992) Creating tests worth taking. *Educational Leadership*, 49 (8), 26–33.