Please cite the source as:

Solano-Flores, G., Shavelson, R. J. & Schneider, S. A.(2001). Expanding the notion of assessment shell: from task development tool to instrument for guiding the process of science assessment development. *Revista Electrónica de Investigación Educativa, 3* (1). Retrieved month day, year from: http://redie.uabc.mx/vol3no1/contents-solano.html

Revista Electrónica de Investigación Educativa

Vol. 3. No. 1, 2001

Expanding the Notion of Assessment Shell: From Task Development Tool to Instrument for Guiding the Process of Science Assessment Development[1]

Una ampliación del concepto de Templete: de herramienta para desarrollar ejercicios a instrumento para regular el proceso de desarrollo de los exámenes de ciencias

Guillermo Solano-Flores (1)
wsolano@WestEd.org
Cultural Validity in Assessment Project
West Ed

Richard J. Shavelson (2)
rich@stanford.edu
DEPENDENCIA
Stanford University

Steven A. Schneider (3)
sschnei@WestEd.org
DEPENDENCIA
West Ed

(1) 4200 Farm Hill Boulevard
94061
Redwood City, Califronia,
United States of America

(2) CALLE Y NO.

ZIP CODE
Standford, California,
United States of America

(Recibido: 1o. de febrero de2001; aceptado para su publicación: 22 de marzo de 2001)

Resumen

En este artículo discutimos las limitaciones y ventajas del uso de templetes. Un templete es un conjunto de instrucciones para desarrollar ejercicios; su fin es abatir el costo y el tiempo de desarrollo de los exámenes. Los templetes no permiten generar ejercicios intercambiables desde el punto de vista estadístico. Sin embargo, cuando sus instrucciones son precisas y se entrena a los autores de pruebas para usarlos adecuadamente, los templetes permiten generar ejercicios de estructura y apariencia similares. Basados en nuestra experiencia y en nuestro trabajo de investigación, discutimos las ventajas de usar templetes como: (a) herramientas para desarrollar pruebas de respuesta construida, (b) documentos que formalizan las propiedades estructurales de los ejercicios; (c) ambientes para la creación de ejercicios que permiten estandarizar y simplificar los formatos de respuesta para los estudiantes; y (d) herramientas conceptuales que regulan el proceso de desarrollo de exámenes. En este artículo también advertimos de posibles usos inapropiados de los templetes.

Palabras clave: Enseñanza de las ciencias, construcción de pruebas, prueba de ejecución.

Abstract

We discuss the limitations and possibilities of shells (blueprints with directions for test developers intended to reduce test development costs and time). Although shells cannot be expected to generate statistically exchangeable exercises, they can generate exercises with similar structures and appearances when they are highly specific and test developers are properly trained to use them. Based on our research and experience developing a wide variety of assessments, we discuss the advantages of conceiving shells as: (a) tools for effective development of constructed-response items, (b) formal specifications of the structural properties of items; (c) task-authoring environments that help test developers standardize and simplify user (examinee) interfaces; and (d) conceptual tools that guide the process of assessment development by enabling test developers to work systematically. We also caution against possible misuses of shells.

Key words: Science education, test construction, performance assessment.

Introduction

Our knowledge of the limitations and possibilities of assessment shells (blueprints for assessment development) has increased considerably in the last few years. As we have used these tools to develop a wide variety of assessments, we have identified several ways in which they can address two issues: (1) the high cost in dollars and time of developing alternative assessments (Aschbacher, 1991; General Accounting Office, 1993; Nuttall,

1992; O'Neil, 1992; Shavelson, Baxter, & Pine, 1992; Stecher & Klein, 1997; Solano-Flores & Shavelson, 1997); and (2) the need to document content representativeness as evidence of assessment validity (Downing & Haladyna, 1997; Crocker, 1997).

In this paper we discuss how shells can help assessment systems keep up with these new demands. We discuss the possibilities and limitations of shells as: tools for developing alternative assessments, including hands-on tasks, simulations, portfolios, and other constructed-response items[2]; documents that specify the structural properties of assessments; programming environments for generating computer-administered items and simplifying the user's (examinee's) interface; and conceptual tools that formalize and regulate the process of assessment development. Our ideas build on experience gained over the last seven years from constructing shells for generating science performance assessments for grades 4 through 8 and evaluating the psychometric properties of these assessments and from using shells to develop portfolios and different sorts of constructed-response tasks for the certification of visual arts and science teachers.

Shells as tools for developing constructed-response assessments

The notion of "item shell" originated from the need for formal procedures for writing short-answer and multiple-choice items. Shells can be thought of as "hollow" frameworks or templates whose syntactic structures generate sets of similar items (Haladyna and Shindoll, 1989), devices that allow test developers to systematically and efficiently generate items.

Regardless of knowledge domain, the principles for constructing and using shells are the same. Content standards, flow charts, mapping sentences, set theory, boolean formulas, concept maps, matrices, or other devices are used to represent propositional and/or procedural knowledge (e.g., Bormuth, 1970; Guttman, 1969; Solano-Flores, 1993). A "universe" of item forms is defined from which tests may be drawn. Each item form is defined by a set of characteristics or generation rules (Hively, Patterson, & Page, 1968). For example, the generation rules in the "division universe" specify the characteristics of the dividend (greater than, equal to, or smaller than 0; fractional or whole); the divisor (greater than, equal to, or smaller than 0; fractional or whole); the quotient (with 0 or without 0); the format (non-equation format, equation with missing dividend, equation with missing divisor, or equation with missing quotient); whether the dividend is greater than, equal to, or smaller than the divisor; and whether or not there is a remainder.

To assemble a test, developers use shells as templates that specify characteristics such as structure, format, style, and language that allow them to sample items over the universe of item forms. Depending on the knowledge domain, shells may have different appearances, from sets of generation rules to complex syntactical structures (Figure 1). In the simplest approach, an item is randomly generated for each item form --and it is assumed that two tests assembled with the same procedure from the same item forms are randomly parallel (Cronbach, Rajaratnam, & Gleser, 1963). Another approach consists of focusing on select item forms and discarding others. For example, given a certain instructional context, test

developers may discard item forms in which the dividend or the divisor are negative because they are beyond the assessment scope.

Developing shells for generating constructed-response assessments is more complex than it is for generating short-answer items. First, the accurate knowledge domain specification is especially critical to generating tasks that are representative of the domain, so that valid inferences from assessment performance to domain performance can be made (Wigdor & Green, 1991). Because of the complexity of the performance involved, numerous practical, methodological, and conceptual dimensions may potentially be relevant to identifying item forms (Figure 2). A specific shell can possibly be used to generate only a few item forms from the vast number of possible item forms.

(a) Generation Rules

Division, equation form:

- missing quotient (A ÷ B = __)
- A > 0, whole, three digits
- B > 0, fractional, three digits
- A < B
- quotient with 0

(b) Directions

Comparative, relational hands-on investigation:

1. Introduce the concepts that will be used in the assessment.
2. Pose a problem or a hypothesis involving one relevant independent variable (X) and one irrelevant independent variable.
3. Provide equipment - include independent variable X and independent variable B. Introduce variable names.
4. Ask the students to solve the problem or test the hypothesis.
5. Ask students to report manipulations, measurements, and results.

(c) Syntactical structure

Describing and explaining a phenomenon in science, based on three relevant related phenomena:

A generalized belief about [phenomenon A] *is that* [description of a misconception of phenomenon A]. *Using your knowledge of* [phenomenon A] *and drawing upon your knowledge of* [non-major content area]:

(a) Describe how [phenomenon W, phenomenon X, phenomenon Y, and phenomenon Z] *are related to* [phenomenon A];

(b) Explain why the statement, "[statement of the misconception of phenomenon A]" *is inaccurate; and*

(c) Explain why [concrete situation or fact that reflects or is due to the relationship between phenomenon A and one or more of the phenomena W, X, Y, and Z].

Your response must show accurate, in-depth knowledge of the concepts, principles, and reasonings related to [phenomenon A].

Figure 1. Examples of three types of shells. (a) Generation rules: test developers assign values to certain variables according to a set of specifications. (b) Directions: test developers take the actions prescribed. (c) Syntactical structure: test developers create text and information specific to an item according to a set specifications (indicated with brackets); some portions of text (indicated with italics), must be kept unmodified.

A science performance assessment is a	locally **globally** 1	standardized	observational classificatory component-identification **comparative** other 2	investigation that requires	**no** **low** **medium** **high** 3						
level of inquiry that may be	embedded in **external but linked to** 4	a curriculum and that contains:	no **one or more** 5	prerequisite tasks,	no **one or more** 6						
preparatory tasks,	one **more than one** 7	core task, and	**one** 8	integrative task that is	**concrete** computer simulation paper and pencil 9	and					
purposively **randomly** exchangeably 10	sampled from activities	that are **like those** 11	characteristically used by	teachers **curriculum developers** subject-matter experts 12	to teach a	purposively **randomly** exchangeably 13					
sampled concept within a	purposively- **randomly-** exchangeably- 14	sampled big idea domain that elicits	**planning** **designing** **investigating** **analyzing** **interpreting** **applying** 15	behavior from	**individual** small groups of 16	assesses that					
produce scores based on	direct observation video observation **written responses** oral responses 17	that	are **are not** 18	accompanied by a concrete physical product that reflects both the processes used							
in carrying out the investigation and the outcomes of the investigation and are scored	**analytically** holistically 19	by using a	**compensatory** noncompensatory 20	scoring model.							

Figure 2. Guttman-like mapping sentence that formalizes the complex types, uses, and characteristics of science performance assessments. A science performance assessment is defined in terms of various dimensions (indicated by numbers) that refer to issues relevant to performance assessment, such as curriculum, assessment structure, task sampling, knowledge domain specification, assessment administration, assessment method, and scoring approach. Bold letters indicate the categories selected by a team of researchers to construct shells for generating specific science performance assessments (see Solano-Flores, Jovanovic, Shavelson, & Bachman 1999; Stecher, Klein, Solano-Flores, McCaffrey, Robbyn, Shavelson, & Haertel, 1999).

Second, interpreting shells for generating constructed-response items poses considerable demands on test developers. To "sample" over the universe of item forms, they must use their expert knowledge on a specific subject matter and make complex decisions to properly meet item form specifications. In addition, because of their complexity, scoring rubrics are defining components of constructed-response items (Solano-Flores & Shavelson, 1997). Classical literature on shells (e.g., Roid & Haladyna, 1982) is not concerned with scoring rubric development because interrater reliability is not an issue for multiple-choice items. For constructed-response exercises, however, the viability of judging performance is critical (see Fitzpatrick & Morrison, 1971) and a significant proportion of assessment development time has to be invested in scoring rubric development.

We recently began using shells to develop scoring rubrics for open-response items and portfolios (Schneider, Daehler, Hershbell, McCarthy, Shaw, & Solano-Flores, 1999; Solano-Flores, Raymond, Schneider, & Timms, 1999). Along with the shell for generating a task of a given type, we provide developers with a generic description of the characteristics of the response at different levels of proficiency. As a part of developing a specific exercise, developers must elaborate these generic scoring rubrics by writing *bullets* that are specific to it. Although systematic research is yet to be done, we have observed informally that, in addition to reducing assessment development time, assessment developers attain in a short time proper alignment between exercise content and scoring rubrics content.

The psychometric properties of performance assessments generated with shells are similar to those for assessments developed with other procedures (Solano-Flores, Jovanovic, Shavelson, & Bachman, 1999; Stecher & Klein, 1997; Stecher, Klein, Solano-Flores, McCaffrey, Robbyn, Shavelson, & Haertel, 1999). First, high interrater reliabilities can be attained with shell-generated items. Second, the interaction of student and task reveals a considerable source of measurement error (Shavelson, Baxter, & Gao, 1993). Although tasks generated with the same shell may correlate slightly higher with each other than they do with tasks from other shells (Klein, Shavelson, Stecher, McCaffrey, & Haertel, 1997), in general, students who perform well on one task do not necessarily perform well on another task within the same knowledge domain.[3]

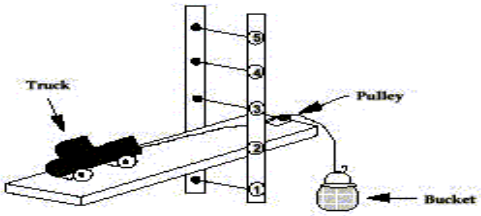
Our experience with Inclines and Friction, two physics assessments for fifth grade, is a case in point (Solano-Flores, Jovanovic, Shavelson, & Bachman, 1999). Inclines and Friction were drawn as samples from the same core concept, Forces and Motion, and developed with the same shell. They posed equivalent problems, had similar appearances, were developed by the same team, and were given to students on consecutive days (the effects of sequence of administration were controlled properly). In addition, the response formats and scoring rubrics were remarkably similar. Notwithstanding, the student-assessment interaction was a considerable source of measurement error. Inclines and Friction scores correlated differently with an external measure of science achievement, and the sequence in which the students took these assessments produced different patterns of score variability.

Contextual factors such as the problem posed, the variables involved, the equipment and wording used for each assessment (see Baxter, Shavelson, Goldman, & Pine, 1992), and the cognitive demands intrinsic to each assessment (see Hodson, 1992) may account for those differences. Each task entails a different process of knowledge utilization. Student performance on two tasks with common content, format, and level of inquiry is no more alike than student performance on two tasks that differ along the same dimensions (Stecher, Klein, Solano-Flores, McCaffrey, Robbyn, Shavelson, & Haertel, 1999). Thus, shells do not ensure assessment exchangeability and should not be thought of as a solution for the old problem of task sampling variability.

Shells as documents that specify item structural properties

Although shells may not be able to generate statistically exchangeable items, they can generate items that appear comparable (i.e., they have similar structures, complexities, formats, and styles). This capability is critical to the success of assessment programs, which need effective ways to insure the similarity of performance measures, but their standardization is weaker than with traditional measures of academic achievement (Haertel & Linn, 1996). For example, school districts need a means of generating assessments that are similar to those used by their states.

Assessment comparability is not difficult to attain when assessments are developed by the same team (Solano-Flores, Jovanovic, Shavelson, & Bachman, 1999) (Figure 3). However, when two or more teams work independently, the shell must be very specific so that all teams interpret it consistently. In an investigation on this matter (Stecher, Klein, Solano-Flores, McCaffrey, Robbyn, Shavelson, & Haertel, 1999), two teams independently developed a performance assessment on the topic, Acids and Bases, with a common shell that provided a sequence of general directions for test developers with only vague guidance on the desired structure and appearance of the exercises. The assessments developed with this shell were considerably different, to the extent that each seemed to reflect the idiosyncrasies of the team that had generated it rather than a set of common directions. For example, the assessments differed considerably on the kind of contextual and ancillary information provided to students, the complexity and sequence of the tasks the students needed to complete, and the format, style, readability, and length of the paragraphs (Figure 4).



Practice pulling the truck up the incline plane by placing washers in the bucket. ***The truck should move to the top of the ramp.***

.....

Frank and **Al** are wondering if they will need *more* force or *less* force to pull the truck up the incline plane if they change the weight of the truck by adding marbles to it.

Suppose you add two large marbles to the truck. Will you have to put *more* washers or *less* washers in the bucket to pull the truck up the incline plane?

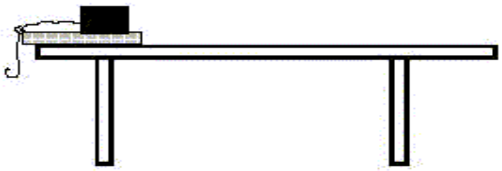
_____

_____

_____

How does the amount of force needed to pull the truck <u>change</u> when the weight of the truck changes?

_____



Notice that the blocks are different **weights** and the BOARDS have different **surface textures**.

Practice pulling one of the blocks along the board by putting washers on the hook.

The back end of the block should cross the starting line.

.....

Sue and **Maria** want to know whether they will need different amounts of force to pull the blocks of different weights along the plain wood board.

Suppose it takes 4 washers to pull the lighter block. Will you have to add *more* washers or *less* washers to the hook to pull the heavier block?

_____

_____

_____

How does the amount of force needed to pull the block <u>change</u> when the weight of the block changes?

_____

Figure 3. Portions of the student notebooks for two assessments on *different* concepts, inclines and friction, generated by the *same* team of developers with a shell whose directions were *vague* (Solano-Flores, Jovanovic, Shavelson, & Bachman, 1999).

Every solution is an acid, a base, or neutral. Acids and bases are chemical opposites of each other. Solutions that are neither acids or bases are neutral. Chemists use numbers to indicate the strengths of acids and bases. The numbers go from 1 to 14. Strong acids have low numbers and strong bases have high numbers. Neutral solutions are in the middle.

Chemists use a solution called Universal Indicator to identify acids and bases. Universal Indicator changes color when mixed with an acid or base. The Universal Indicator Color Guide shows that Universal Indicator turns red when it is added to a strong acid, it turns purple when it is added to a strong base, and it turns yellowish-green when it is added to a neutral solution.

UNIVERSAL INDICATOR COLOR GUIDE

Strong Acid		Weak Acid				Neutral				Weak Base		Strong Base	
1	2	3	4	5	6	7	8	9	10	11	12	13	14
RED	RED	ORANGE	YELLOW	GREEN	YELLOWISH GREEN	BLUE	PURPLE	PURPLE					

All acids in the range of 1 to 4 turn the indicator red. All bases in the range of 11 to 14 turn the indicator purple. Today you will learn how to test if one acid is stronger than another even if they both turn the indicator the same color.

PART 1: READING THE SCALE

1a. Which acid is stronger -- one that turns Universal Indicator orange or one that turns Universal Indicator yellow?

_____

1b. Which base is stronger -- one that turns Universal Indicator blue or one that turns Universal Indicator purple?

_____

All solutions are acids, bases, or neutral. You can use pH paper and a pH Color Chart to test whether a solution is an acid, a base or neutral.

Part 1: **READING THE pH SCALE**

To practice using the pH paper:

- Squeeze 6 drops of Solution X into one of the measuring cups. Gently swirl the cup.
- Take one strip of pH paper out of the bag, and dip it into Solution X.
- Remove the strip from the cup and quickly observe the color of the pH paper. Be **sure to look at the color right away, because it will change quickly. The first color shows the correct pH.**

1a. What is the **color** of the pH paper right after you dipped it into Solution X? _____

1b. What **number** on the pH Color Chart goes with this color? _____

1c. Look at the chart below. Is Solution X an acid, a base, or neutral? _____

pH levels

1	3	5	**7**	9	11
Pink	Rust/ Brown	Golden-Orange	**Greenish-Yellow**	Dark Green	

ACIDS NEUTRAL BASES

Blue

Figure 4. Portions of the student notebooks for two assessments on the *same* concept, acids and bases, generated by *independent* teams of developers with a shell whose directions were *vague*. (Stecher, Klein, Solano-Flores, McCaffrey, Robbyn, Shavelson, & Haertel, 1999). The notebooks reflect different interpretations for the same set of directions prescribed by the shell.

Based on that experience, we have moved away from assuming that shells can be interpreted in the same way by all developers. We now know that, to attain exercise comparability across teams, three conditions must be met: (a) shells must be constructed based on a clear idea of the structure of the exercises they are intended to generate; (b) the directions for assessment developers must be highly specific; and (c) developers must be carefully and extensively trained to use them.

The challenges posed by the assessment for the certification of science teachers (National Board for Professional Teaching Standards, 1996) underscore the importance of task structure, specificity of directions, and developer training for proper shell construction and use. Assessment comparability across four science content areas --biology, chemistry, Earth and space science, and physics-- was critical because, whereas each candidate had to choose one area as his or her specialty (and, depending on that selection, complete a specific set of exercises), the certificate of accomplished science teacher made no reference to any science content area. Therefore, we needed a means to ensure that the same assessments for the four content areas were composed by the same combination of types of exercises and that each type of exercise was represented in all content areas.

To ensure assessment comparability across content areas, we first identified basic science process skills (e.g., data interpretation, use of procedures, construction of strategies for problem solution). Second, we constructed a shell intended to generate exercises for each of these process skills. Third, we generated an exercise for each content area and each science process skill by sampling content from the content standards (National Board for Professional Teaching Standards, 1996; National Research Council, 1996) and presenting this content according to the structure specified by the shell.

Developers who were unfamiliar with shells reported that they felt "confined" by all the specifications, probably because they were used to generating tasks that met content specifications, not content *and* task structure specifications. Training, then, involved helping them realize that content-rich exercises could be developed with shells despite their strict specifications and helping them translate their ideas into the structures specified by those shells. As a part of the training, we provided assessment developers with samples that illustrated what the exercises developed using the shells should look like. Also, we had them develop a number of exercises under the guidance of experienced colleagues or staff members. Depending on exercise complexity, this training took from as little as a few hours to as much as three days before a reasonable level of efficiency could be attained.

Thanks to this training and the specificity of the directions provided by our “illustrated” shells, we were able to produce exercises that were comparable both across development teams and content areas.

Shells as authoring environments for test developers

Shells can be used as authoring tools that contribute to ensuring the usability of assessments. The design of these components must consider users' characteristics and preferences, so that examinees have no difficulty in following the directions provided by an assessment, manipulating equipment, or understanding how to provide their responses. This ultimately contributes to reducing measurement error produced by factors not relevant to the knowledge or skills measured.

Our experience using Science Explorer 3.0™, a software package developed and published by LOGAL™ (1995), illustrates how shells can contribute to ensure usability. We used this software experimentally, as a part of our search for performance-based tasks that could be administered efficiently to assess scientific process skills. Explorer 3.0™ is a science curriculum based on simulations of phenomena in physics, chemistry, and biology. To operate the simulations, the user must perform actions such as mouse dragging and clicking on screen buttons. Whereas some tools and objects are common to all simulations (e.g., start the simulation, stop, reset), others were specific to each simulation (e.g., manipulate the value in ohms of a resistor or select either the parent or F1 cross, respectively in the Electricity and the Modes of Inheritance assessments).

We adapted LOGAL™'s simulations to create tasks in which examinees conducted investigations by manipulating variables, observing the results of their actions, and reporting their investigations. In addition to using the simulation engine to model specific, contextualized problems, we took advantage of the software's scriptability to control features such as the numerical and visual information displayed on the screen, the format in which that information was presented, and the simulations' presentation time.

Based on feedback from users and our own observations, we realized that no computer skills should be assumed in the examinees. Therefore, we constructed a shell that intended to both allow task developers to generate simulations efficiently and minimize the computer skills needed to be able to operate the simulations. This shell acted as both a programming environment and a user's (examinee's) interface. For example, it specified the screen layout and the set of characteristics of tools, objects, graphs, and input and output variables that should be used across all simulations; it also restricted the actions needed to operate the simulations to mouse pointing, mouse clicking, and typing. Mouse dragging-and-clicking or double-clicking were not needed (Figure 5). This allowed us to address two major issues in computer-based assessment: the convenience of using task-authoring environments that facilitate the creation of tasks according to a set of format specifications (see Katz, 1998), and the need for user's interfaces that prevent user's inexperience with computers from interfering with performance (see Bejar, 1995).

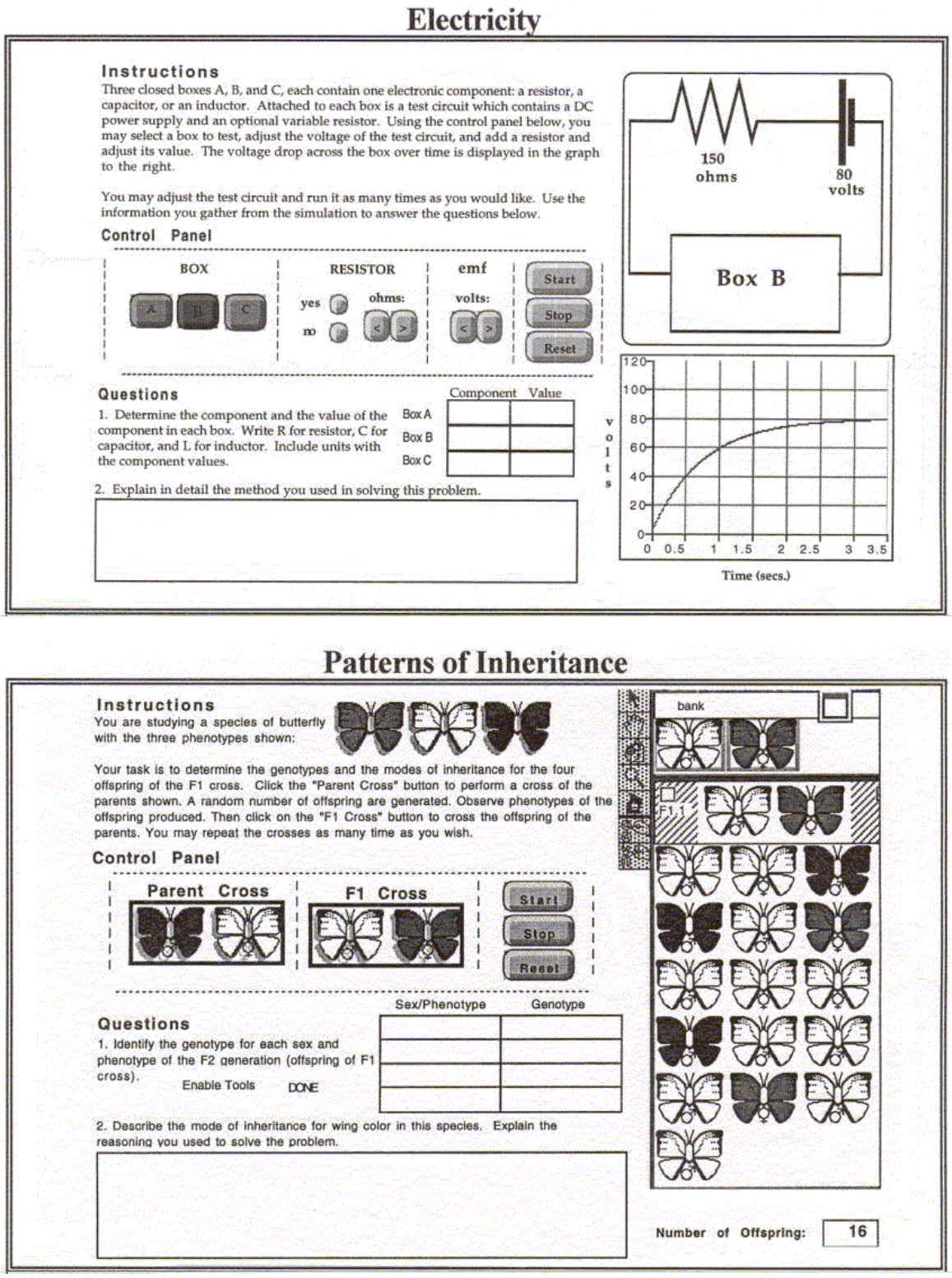


Figure 5. Two simulations developed with LOGAL™. The shell acts as both a programming environment and an interface that standardizes the response formats across simulations to minimize the required computer skills. Candidates select the values of variables and run, stop, and reset simulations by clicking on Control Panel buttons and enter text in the boxes

provided. The right side of the screen displays the process resulting from the examinee's manipulations. In "Electricity," examinees test the contents of three mystery boxes, A, B, and C, by building circuits attached to them; the simulation displays the selected box and circuit and a time-voltage graph. In "Butterflies," examinees investigate patterns of inheritance based on the phenotypes of two generations of a hypothetical species of butterflies; since the total number of individuals and the number of individuals with a specific phenotype varies with each offspring, to solve the problem correctly the examinee needs to run the simulation several times.

Shells as conceptual tools for regulating the process of assessment development

Because of the formality with which they may specify the intended structure of exercises, shells can be used as tools that regulate the cyclical process of assessment development (Solano-Flores & Shavelson, 1997). In each iteration of this cyclical process, assessment developers discuss and negotiate any change made on an exercise's task, response format, or scoring rubrics. Shells can be used as tools that contribute to making those discussions systematic.

The development of the portfolios for the certification of science teachers illustrates this. Candidates complete these portfolios throughout a school year (see Schneider, Daehler, Hershbell, McCarthy, Shaw, & Solano-Flores, 1999). Each entry (portfolio exercise) intends to capture a major aspect of science teaching (e.g., assessment, conveying a major idea in science) and requires candidates to submit a specific set of materials, such as videotape footage of their teaching, samples of student work, or narratives. To ensure standardization and clarity, the directions for completing each entry must specify, among other things, the required format and length of the responses (e.g., page limits, font size, specifications for videotape footage), the actions that should be taken to complete the responses (e.g., how to select the students whose work samples would be discussed by the candidate); and possible actions that should not be taken (e.g., submission of not more than five pages because assessors cannot read more than the text length specified).

We used shells because we wanted all these portfolio entries to have similar appearances. We reasoned that candidates could readily understand the complex actions they had to take to complete their portfolios if the directions were provided in the same sequence and style across entries. In addition, we needed to develop the entries concurrently because they intended to address aspects of teaching that are mutually complementary.

As we gained experience from pilot testing, the shell, together with the entries, evolved and became more detailed --to the extent that it specified, for example, usage of font styles and cases, and where tables, illustrations, and even page breaks should be inserted. During this process, we frequently found ourselves using the shell as a frame of reference in our discussions. Any modification made on one entry (e.g., providing a planner to help candidates organize the activities they needed to complete that entry), was also made on the shell; any modification made on the shell was also made on all the entries in the next version of the portfolio.

Although the shell could not possibly tell much about the content of an exercise, it certainly contributed to making the discussions of the assessment development team more systematic. Thus, a shell that evolves along with the exercises it generates becomes a succinct description that summarizes the intended task structure, a document that formalizes the assessment developers' current thinking.

Summary and concluding remarks

In this paper we have shared knowledge gained from investigating and using shells in the last years. Provided that shells are carefully designed and test developers are properly trained to use them, shells can be effective assessment development tools. Because of their formal properties, using shells promotes the use of a construct-driven approach to assessment development, in which assessment content is determined by specifying a knowledge domain (see Messick, 1994), as opposed to a task-driven approach (e.g., Lindquist, 1951; Thorndike, 1971) in which assessment content is determined by identifying tasks for a specific assessment. This construct-driven approach can contribute to documenting content representativeness as a form of content validity.

The psychometric quality of assessments generated with shells is similar to that of assessments generated without them. In addition, shells make the process of assessment development more systematic and may potentially reduce development time and costs. At the present time, however, only anecdotal information on shell cost-effectiveness is available, in part because, when we started using shells to develop alternative assessments, we focused on the psychometric properties of the assessments generated. In addition, only recently have we become aware of how precise shells must be and how assessment developers must be trained to use them.

Now that we know how shells must be designed and used to effectively generate assessments, a word of caution about possible misuse becomes necessary. In order not to oversample a few particular aspects of a knowledge domain and undersample many others, the exercises that compose an assessment must be generated with a good number of shells. For example, using a shell for each type of exercises such as multiple-choice, short-answer, or hands-on, does not produce enough exercise variety. A good number of shells must be used to generate different types of multiple-choice, different types of short-answer, and different types of hands-on exercises in the same assessment. We followed this principle to generate the assessment center exercises for the certification of science teachers. In as much as we wanted to ensure comparability across content areas; we also wanted to have good diversity of exercises within each content area. Therefore, each of the twelve constructed-response exercises included in the assessment for each content area was developed with a different shell.

Originally conceived as tools for efficient item generation, we now think about shells in multiple ways that make assessment development a more systematic process: as generic descriptions of the desired structure and appearance of a wide range of types of items; as

programming environments for developers; and as conceptual tools that help assessment developers communicate effectively. We hope that future assessment programs will adopt these multiple perspectives into their procedures for assessment development.

References

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

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

Bejar, I.I. (1995). From adaptive testing to automated scoring of architectural simulations. In E. L. Mancall & P.G. Bashook (Eds.), *Assessing clinical reasoning: The oral examination and alternative methods* (pp. 115-127). Evanston: American Board of Medical Specialties.

Bormuth, J. R. (1970). *On the theory of achievement test items*. Chicago: University of Chicago Press.

Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education, 10* (1), 83-95.

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

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*, 137-163.

Downing, S. M. & Haladyna, T. M. (1997). Test item development: Validity evidence form quality assurance procedures. *Applied Measurement in Education, 10* (1), 61-82.

Fitzpatrick, R. & Morrison, E.J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 237-270). Washington: American Council on Education.

Gelman, R. & Greeno, J. G. (1989). On the nature of competence: Principles for understanding in a domain. In L. B. Resnick (Ed.), *Knowing, learning, and instruction. Essays in honor of Robert Glaser* (pp. 125-186). Hillsdale: Lawrence Erlbaum Associates, Publishers.

General Accounting Office (1993). *Student testing: Current extent and expenditures, with cost estimates for a national examination*. Washington: Author.

Guttman, L. (1969). Integration of test design and analysis. In *Proceedings of the 1969 invitational conference on testing problems* (pp. 53-65). Princeton: Educational Testing Service.

Haertel, E. H. & Linn, R. L. (1996). Comparability. In Phillips, G. W. (Ed.), *Technical issues in large-scale performance assessment* (pp. 59-78). Washington: National Center for Education Statistics, U. S. Department of Education, Office of Educational Research and Improvement.

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

Hively, W., Patterson, H. L., & 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, 115-144.

Katz, I. R. (1998). *A software tool for rapidly prototyping new forms of computer-based assessments* (GRE Board Professional Report No. 91-06aP). Princeton: Educational Testing Service.

Klein, S. P., Shavelson, R. J., Stecher, B. M., McCaffrey, D., & Haertel, E. (1997). *The effect of using shells to develop hands-on tasks*. Unpublished manuscript. Santa Monica: The RAND Corporation.

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. Washington: American Council of Education.

LOGAL™ (1996). *Science Explorer 3.0* [Computer program]. Cambridge, MA: LOGAL Software, Inc.

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

National Board for Professional Teaching Standards (1996). *Adolescence and Young Adulthood/Science Standards for National Board Certification*. Detroit: Author.

National Research Council (1996). *National Science Education Standards.* Washington: National Academy Press.

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, 49* (8), 14-19.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing*. New York: Academic Press.

Schneider, S., Daehler, K. R., Hershbell, K., McCarthy, J., Shaw, J., & Solano-Flores, G. (In press). Developing a national science assessment for teacher certification: Practical lessons learned. In L. Ingvarson (Ed.), *Assessing teachers for professional certification: The first ten years of the National Board for Professional Teaching Standards*. Greenewich: JAI Press, Inc.

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

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

Shavelson, R. J., Ruiz-Primo, M. A. & Wiley, E. (1999). Notes on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36* (1), 61-71.

Solano-Flores, G. (1993). Item structural properties as predictors of item difficulty and item association. *Educational and Psychological Measurement, 53* (1), 19-31.

Solano-Flores, G., Jovanovic, J, Shavelson, R. J., & Bachman, M. (1994, April). Development of an item shell for the generation of performance assessments in physics. Paper presented at the *Annual Meeting of the American Educational Research Association*, New Orleans.

Solano-Flores, G., Jovanovic, J., Shavelson, 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.

Solano-Flores, G., Raymond, B., Schneider, S. A., & Timms, M. (1999). Management of scoring sessions in alternative assessment: The computer-assisted scoring approach. *Computers & Education, 33*, 47-63.

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

Stecher, B. M., & 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., Robbyn, A., Shavelson, R. J., & Haertel, E. (In press). The effects of content, format, and inquiry level on performance on science performance assessment scores. *Educational Evaluation and Policy Analysis*.

Thorndike, R. L. (Ed.). (1971). *Educational Measurement* (2nd. Ed.). Washington: American Council on Education.

Wigdor, A. K., & Green, B. F., Jr. (Eds.). (1991). *Performance assessment for the workplace* (Vol. 1). Washington: National Academy Press.

---

[1] The research and experience discussed in this paper comes from different projects with the RAND Corporation, the National Board for Professional Teaching Standards, Stanford University, University of California, Santa Barbara, and West Ed. The ideas presented in this article are not necessarily endorsed by the supporting or funding organizations or our colleagues from those organizations. Steve Klein, Brian Stecher, and Ed Haertel contributed significantly in the construction of the mapping sentence presented in Figure 1. Special thanks to Stan Ogren, Kirsten Daehler, Jasna Jovanovic, Kristin Hershbell, and Jerome Shaw, whose brilliant comments and hard work made it possible to use and refine some of the shells described.

[2] The terms, item, task, prompt, and exercise are used interchangeably throughout this paper.

[3] In the designs of the investigations here reported, the facets occasion and task are inevitably confounded (see Cronbach, Linn, Brennan, & Haertel, 1997). Appropriate analyses to estimate the effect of this confounding have revealed a strong interaction of task and occasion (Shavelson, Ruiz-Primo, & Wiley, 1999). However, the instability of performance across occasions seems to be due at least to some extent to the students' partial knowledge of the domain assessed. Indeed, it has been suggested that granting achievement in a domain should be based on considering how consistently students perform correctly on a variety of tasks that are representative of that domain (Gelman & Greeno, 1989).