ED 426 078 TM 029 302

AUTHOR	Stecher, Brian M.; Klein, Stephen P.; Solano-Flores, Guillermo; McCaffrey, Dan; Robyn, Abby; Shavelson, Richard J.; Haertel, Edward
TITLE	Do Content, Format, and Level of Inquiry Affect Scores on Open-Ended Science Tasks?
SPONS AGENCY	National Science Foundation, Arlington, VA.
PUB DATE	1998-04-00
NOTE	28p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
CONTRACT	MDR-9154406
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Correlation; Grade 8; *Junior High School Students; Junior High Schools; *Performance Based Assessment; *Science Tests; Scores; Test Construction; *Test Content; Test Format; Test Results
IDENTIFIERS	Large Scale Assessment; *Open Ended Questions

ABSTRACT

This study investigated three factors that may contribute to the large variation in student performance across open-ended measures. These factors are content domain, format (whether the task required only pencil and paper or involved a hands-on manipulation of equipment), and level of inquiry (whether the task guided the student toward the solution or required the student to develop a solution strategy). Six similar investigations of acids and bases were developed from a common shell that controlled for format and level of inquiry. More than 1,200 eighth graders completed 2 of these tasks as well as tasks drawn form other content areas and a multiple-choice test of science. Results do not bear out the hypothesis that tasks that were similar to each other in content, level of inquiry, and format would be more highly correlated with each other than with measures that differed on these dimensions. Post-hoc analyses of the tasks revealed unanticipated differences in developers' interpretation of the shell that may have affected student performance. Implications for large-scale use of performance measures are discussed. (Contains 3 tables, 6 figures, and 23 references.) (Author/SLD)

DO CONTENT, FORMAT, AND LEVEL OF INQUIRY AFFECT SCORES ON OPEN-ENDED SCIENCE TASKS?

Brian M. Stecher and Stephen P. Klein, RAND

Guillermo Solano-Flores, WestEd

Dan McCaffrey and Abby Robyn, RAND

Richard J. Shavelson and Edward Haertel, Stanford

Paper presented at the annual meeting of the American Educational Research Association, April 13-17, 1998, San Diego, CA.

This material is based upon work supported by the National Science Foundation under Grant No. MDR-9154406.

DO CONTENT, FORMAT, AND LEVEL OF INQUIRY AFFECT SCORES ON OPEN-ENDED SCIENCE TASKS?

Abstract

This study investigated three factors that may contribute to the large variation in student performance across open-ended measures. These factors are content domain, format (whether the task required only pencil-and-paper or involved a hands-on manipulation of equipment) and level of inquiry (whether the task guided the student toward the solution or required the student to develop a solution strategy). Six similar investigations of acids and bases were developed from a common shell that controlled for format and level of inquiry. Students completed two of these tasks as well as tasks drawn from other content areas and a multiple choice test of science. Results did not bear out the hypothesis that tasks that were similar to each other in content, level of inquiry, and format would correlate higher with each other than with measures that differed on these dimensions. Post-hoc analyses of the tasks revealed unanticipated differences in developers' interpretation of the shell that may have affected student performance. Implications for large-scale use of performance measures are discussed.

Background. The past few years have witnessed a rapid growth in the use of performance (open-ended) measures in large-scale state and national testing programs (Aschbacher, 1991; Bond et al., 1996). However, recent research has raised concerns about the reliability and validity of such performance-based assessments (Hambleton et al., 1995; Koretz, et al., 1994). For example, open-ended tasks within a content domain (such as writing, mathematics, or science) usually have only low correlations with each other (Hieronymous, et al., 1987; Burger and Burger, 1994; Baxter, et al., 1993). Similarly, generalizability analyses of open-ended, "hands-on" measures in science (i.e., tasks that require working with materials and/or equipment) find that the person-by-task component of variance is usually greater than the person or task main effects and can account for as much as 80% of the total score variation (Shavelson, et al., 1993). Results such as these have important consequences for the use of open-response measures in large-scale assessment. For example, several tasks and a large amount of testing time may be needed to produce scores that are reliable enough to allow reporting results for individual students or even for classrooms or schools (Cronbach, Bradburn and Horvitz, 1994).

Researchers have suggested several factors that may contribute to the large variation in student performance across open-ended measures, including the format of the task and the content area being measured. Shavelson, Baxter, and Pine (1992) found that the relative standings of students in a well defined content area (such as basic understanding of electric circuits) was very sensitive to the measurement method employed (e.g., direct observation, multiple choice test, laboratory notebook, or computer administered test). Research on multiple choice tests shows that scores on measures of similar content are more alike than scores on tests in different content areas. For example, Hoover, et al. (1995) report higher correlation among the subtests of the English Language portion of the ITBS (e.g., spelling, punctuation, etc.) than between these subtests and the Reading Comprehension test. Although it would be natural to expect to find similar format and content effects on performance assessments, the evidence is mixed. Klein, et al. (in press), for example, found that two hands-on science tasks that used virtually identical formats and assessed what appeared to be very similar skills and knowledge often correlated no higher with each other than they did with hands-on tasks in other science content areas or even with a standardized multiple choice test.

Another factor that may affect student performance is a dimension we call "level of inquiry." Shulman and Tamir (1973) argued that students will perform differently depending on who sets the problem, the procedures for addressing it, and the nature of the solution. Tamir, Nussinovitz, and Firedler (1982) echoed this distinction in later work to develop an assessment inventory for inquiry-oriented practical laboratory examinations. Along the "level of inquiry" dimension we distinguish between guided tasks and unguided tasks. In a guided task, the student is presented with a problem and appropriate tools and is given a step-by-step procedure for solving it. In an unguided task, the student is presented with a problem and appropriate tools, but must figure out a method for solving it.

Purpose. Our research examined whether differences on three dimensions--format, content area, and inquiry level--contributed to the variability in student scores across open-ended science tasks. Specifically, did two open-ended science tasks that were similar on all three of these dimensions correlate higher with each other than they did with tasks that were different on one or two of these dimensions? In addition, we compared scores on open-ended science tasks with scores on a broad gauge standardized multiple choice science test to see whether the performance assessments correlated higher with one another than with the multiple choice test. Finally, we explored whether scores were affected by the sequence in which students took the tasks, because there is some evidence of practice effects on hands-on science tasks (Hamilton, et al., 1998).

Methods. We used eight open-ended science tasks that varied in terms of format (hands-on or paper-and-pencil), level of inquiry (guided or unguided) and content. (See Table 1.) Six of the open-ended tasks addressed the same core content, acids and bases. They were developed for this project using a single "shell" or blueprint to construct conceptually similar tasks (Hively, Patterson and Page, 1968; Solano-Flores and Shavelson, 1997). The shell defined a comparative investigation task in which students had to compare objects (e.g., samples of vinegar) with respect to a certain attribute (e.g., pH level). Each task lasted one class period and comprised three stages: (a) performing an experiment, (b) interpreting the results, and (c) applying the principles learned to a new, but related, situation. These stages were consistent with the science process skills described by Tamir and Doran (no date).

The shell indicated how to create the following three versions of the tasks: Discovery, Recipe and Text (See Table 1). The shell specification allowed the developer to vary two dimensions: level of inquiry (guided or unguided) and format (hands-on or paper-and-pencil). In the paper-and-pencil versions, students read descriptions of experiments and saw drawings of scientific measurements being taken but they had no actual equipment and did not make scientific measurements themselves. In the hands-on versions, students used scientific equipment and measuring tools to solve a problem.

Two teams of researchers from the University of California, Santa Barbara (Team 1) and from Stanford University-WestEd (Team 2) used the common shell to develop Discovery, Recipe and Text versions of a task on acids and bases. The tasks developed by Team 1 involved determining which of three solutions would cure an imbalance in the pH of an imaginary space alien's blood (using pH paper as a measuring tool). The tasks developed by Team 2 involved measuring the strength of three concentrations of cooking vinegar (using universal indicator solution to measure pH). The teams worked independently to develop their tasks.[1] Figure 1 shows the portion of the shell that defined the Recipe version.

In each task, students were introduced to the topic of acids and bases and to scientific tools for testing pH. Figures 2 and 3 show the introductory stages (A-G) of the Recipe versions of both tasks. The next portion of each task presented the students with a problem (e.g., determine which of the three bottles contained the solution that would save the alien's life). At this point, the three tasks in each set began to differ in format and/or level of inquiry. The Discovery task was an unguided, hands-on measure. It asked students to design and carry out their own experiment using the materials provided to solve the problem. The Recipe task provided explicit directions for using the equipment and materials to solve the problem. Students followed the steps, recorded their own data to reach a conclusion. Therefore, the Recipe tasks was a guided, hands-on measure on the topic of acids and bases. The Text task described an experiment that examined the properties of each solution and provided the data obtained from each step in that experiment. Students were shown drawings of the experimental set up, but did not work with the materials. Thus, the Text task could be classified as a guided, paper-and-pencil measure on the topic of acids and bases.

All three versions (Discovery, Recipe, and Text) used the same concluding section. In other words, all three had the same interpretation and application questions. These questions asked students to apply what they learned from the experiment to a new problem. For example, all three Blood tasks asked students to use the relationships they observed in their experiment to answer questions about an acid/base imbalance that was affecting the health of fish in a lake. Figures 4 and 5 show the application questions of the Blood and Vinegar tasks, respectively. Each of the six acid/base tasks required one classroom-period of testing.

[1] The UCSB researchers also developed the shell used in this study, so they had more experience with it than the other team.

In addition to the acids and bases tasks, all the participants completed two other guided, hands-on tasks that were drawn from other content areas (Stecher & Klein, 1996).[2] One of these tasks was called Levers, and the other was called Materials. In the Levers task, students examined the relationship between the length of a lever, the location of its fulcrum, and its ability to lift a fishing weight. In the Materials task, students were given a set of eight natural materials (rock, fur, shell, etc.) and asked to create a two-way classification system, so that each material fit in only one cell and each cell had at least one material. The Levers and Materials tasks each required twenty-five minutes to complete, and they were administered during one class period. In addition, all the participants took the Science portion of the multiple-choice Iowa Tests of Basic Skills (ITBS). This test also lasted about one class period.

Over 1,200 eighth graders participated in this research. Each student completed four periods of testing over a four or five day period, including two acid/base tasks (one version of Vinegar and one version of Blood), Levers, Materials, and the ITBS Science subtest.[3] Classrooms were assigned randomly to one of the 18 possible permutations of sequences of the acid/base tasks (e.g., Vinegar-Text and then Blood-Recipe), but more classrooms were assigned to some permutations than others to increase power for certain planned comparisons. All students completed the Levers and Materials tasks on the second day and the ITBS test on the last day (see Table 2).

Student responses on all the open-ended measures were scored by science teachers using task-specific analytic rubrics. The rubrics were designed by the task development teams and were reviewed and revised by other members of the research team, so the scoring procedures would be similar and the criteria for judgment would be comparable. Each task contained about a dozen scorable items, and a different

[2] Because their format was hands-on their inquiry level was guided, these two tasks could be classified as "recipe."

[3] In most schools the tests were administered during science class over four consecutive days. Scheduling conflicts at other schools required that the testing period be extended by one or two days to accommodate other activities or a weekend.

number of points were awarded to each item based on its complexity. For example, a question that offered a dichotomous choice ("Which acid is stronger, Acid A or Acid B?) was worth one point. A more complex question ("How do you know this?") was worth up to two points depending on the number of key features mentioned in the student's explanation. The data table in which students recorded the results of four experimental trials was scored for accuracy on a scale from zero to four. Almost all responses were scored by two independent readers, and many were scored by three readers. Inter-reader correlations were extremely high, ranging from 0.93 to 0.97 across the eight open-ended tasks used in this study. For the analyses below, each student's score was the average of the scores assigned by the first two readers.

We also conducted three post-hoc reviews of the similarity of the assessments generated by the two task development teams. The first review compared the tasks in terms of reading level. The Text versions of the tasks were reviewed using word processing software that computed three estimates of reading level: Flesh-Kincaid, Coleman-Liau, and Bormuth. These indices use word length, sentence length, and structural features of the text to estimate the grade level at which a typical student could read the material. We computed the average of these three estimates. The second review compared the two recipe versions of the tasks in terms of the logic required to solve them. We outlined the most direct sequence of steps required to complete each task and looked for differences in the reasoning at each step in the process.

The third review assessed the compliance of the discovery and recipe tasks generated by the two teams with the task generation rules contained in the shell. We inspected the task materials to determine how faithfully the development teams had interpreted and followed the directions provided by the shell. Each paragraph in the task was examined to determine the shell action or actions that it seemed to be intended to address. A group of researchers composed of members of both development teams used this approach to compare the discovery and recipe versions of the Vinegar and Blood tasks with each other and with the shell.

Results. We anticipated that tasks that were similar to each other in content, level of inquiry, and format (such as those that use hands-on activities) would correlate higher with each other than they would with other measures. This did not happen (see Table 3). With few exceptions, all the correlations among the tasks fell in the relatively narrow range of 0.50 to 0.66.[4] None of the correlations were significantly different from each other.[5] For example, the correlation between two unguided, hands-on, acid/base tasks (e.g., Vinegar-Discovery and Blood-Discovery) was .50. This was not significantly different than their respective correlations with the ITBS (a multiple-choice test of general science) or their correlations with the Levers and Materials tasks (guided, hands-on tasks in different content areas). Similarly, the Recipe tasks correlated no higher with each other than they did with the ITBS. In general, the average correlation among tasks was not related to the number of dimensions on which the tasks were similar.[6]

We recognize that the relationships among measures might change if it were possible to correct the observed correlations for attenuation due to unreliability in the individual scores. However, there is no truly appropriate indicator of task reliability that could be used for this purpose. Using inter-rater reliability would adjust for errors in the scoring process. However, because the inter-rater reliabilities for the performance

[4] The correlations in Table 3 pool all students who took the task regardless of order. The results are not substantially different if sequence is considered.

[5] Because these correlations coefficients were based on relatively small numbers of students nested in small numbers of classrooms, we computed three sets of standard errors and 95% confidence intervals. Between-classroom standard errors were estimated using the jackknife procedure and within-classroom standard errors were estimated using a z-transformation and by bootstrapping. All three methods produced confidence intervals with a width of 0.20 to 0.30 correlation units. The only consistently significant differences in Table 3 are between the two extreme values (0.74 and 0.42).

[6] We also examined the dimensionality of the tasks using factor analytic techniques and found that a single factor solution best fit the data, i.e., there was no evidence that content, format, and inquiry level had separable effects on scores. Note, however, that because no student took more than one of the three task versions created by each team, some content effects may have gone undetected.

tasks were extremely high (ranging from 0.93 to 0.97), disattenuating for rater reliability does not alter the relationships among the measures. Using internal consistency estimates based on the individually scored elements of each task would adjust for heterogeneity of student performance on a task. However, because the internal consistency reliability estimates (Cronbach's alpha) for the six acids and bases tasks were quite uniform (ranging from 0.75 to 0.81),[7] disattenuating these values would not alter the pattern of correlations among the measures.[8]

We anticipated that the order in which students took the tasks would affect their scores. For example, students might learn skills from one task that would improve their performance on a subsequent task. In particular, we expected that a student's score on an unguided Discovery task would be higher if that student had taken a guided Recipe or Text task previously. We also anticipated that a student's score on a hands-on task (Discovery or Recipe) might be higher if the student had already taken a hands-on task rather than a paper-and-pencil task. However, there were no consistent sequence effects. Figure 6 illustrates this finding by showing the adjusted mean scores and 95% confidence intervals on each of the two Discovery tasks when those tasks were taken second categorized by the type of task taken first.[9] The top three score bands in Figure 6a show the mean score on Vinegar-Discovery when taken second following the text-, recipe-, or discovery-version of the Blood task. The bottom score band in Figure 6a shows the mean score on Vinegar-Discovery when it was taken first (which reflects a "no treatment" condition). Figure 6b shows similar results for Blood-Discovery. Taken together, these figures show that there is no significant difference in mean scores

[7] The comparable values for Levers and Materials were 0.74 and 0.84.

[8] Furthermore, the scorable "items" in each performance task are not independent which biases the estimate of internal consistency (e.g., producing the correct answer to one part of the task may be contingent on correctly performing a prior part of the task).

[9] Scores were standardized and then adjusted for differences in student abilities as measured on the ITBS Science test. This adjustment was made because treatment conditions were randomly assigned by class not by student giving greater variability across groups than would normally be expected. Using unadjusted mean scores does not change the pattern of results.

associated with which task was taken first (i.e., there is substantial overlap among the four confidence intervals). In particular, scores on the unguided tasks were not consistently higher when they followed guided tasks, i.e., students did not gain information from a text task or a recipe task that helped them perform a subsequent discovery task. The same results were obtained when recipe and text tasks were taken second. In general, scores on hands-on tasks are not consistently higher when they follow one type of hands-on task than when they follow another type of hands-on task.

There were two exceptions to this pattern. Scores on the text version of the Blood task were significantly higher when this version was taken after the recipe version of the Vinegar task than when it was taken following the text or discovery versions of the Vinegar task. Interestingly, there were contradictory trends when the Recipe versions of the two tasks were taken second. Scores on the recipe version of the Blood task were significantly higher after taking the recipe version of the Vinegar task than after taking any other version of the Vinegar task. However, the opposite was true for the recipe version of the Vinegar task. Scores on the recipe version of the Vinegar task were significantly lower after taking the recipe version of the Blood task than after taking any other version of the Blood task. These results suggest that the recipe versions of the Vinegar and Blood tasks may have been less alike than anticipated in terms of cognitive demands.

The content review of the tasks may partially explain this last result. Despite the effort and care that went into task development, the content reviews suggest that the Blood and Vinegar tasks were not equivalent in a number of potentially important ways. First, the average of three estimates of reading grade level for the text version of a task was 7.3 for Vinegar and 6.2 for Blood. These estimates indicate that the typical eighth grade student should be able to read both tasks, although students would find it more difficult to read the Vinegar task than the Blood task. (We were testing during the eighth and ninth months of the school year.) The comparable readability index for the

ITBS Science test used in this study is 6.8 (using the Dale-Chall estimation method; Hieronymus, et al. 1990). Students who were reading well below grade level might have had difficulty reading the open-response tasks, particularly the Vinegar task, as well as the multiple-choice test.

Second, the tasks were less alike conceptually than we anticipated. For example, the logical path through the Vinegar questions was more complex than the path for the corresponding question on the Blood tasks. The latter required a static comparison of pH readings to solve the problem while the former required a comparison of colors derived from a sequence of steps to solve the problem. The tasks differed in other ways, as well. The Blood task required testing the pH of the solution after the experimental intervention, while in the Vinegar task, the pH testing was coincidental with the experimental action. These differences may have contributed to the contradictory practice effects from taking one task before the other, e.g., learning to neutralize acids using the Vinegar procedure might make it easier to learn the testing method needed in the Blood task.

Finally, the results from the analysis of compliance with the shell show that the two task development teams were not entirely faithful to the shell and did not operationalize the shell in comparable ways. The most common discrepancies between the shell and the two sets of tasks involved including actions not prescribed by the shell, omitting actions contained in the shell, repeating actions not called for by the shell, and interpreting the directions from the shell "too liberally." These differences suggest that the shell was not prescriptive enough to eliminate important differences between the development teams (such as style or personal preferences) or to control differences derived from the characteristics of the specific task or equipment with which a team chose to work.

Some of the deviations result from the fact that the shell included actions that were not clear enough or were difficult to implement. For example, stage C ("Provide irrelevant variables") does not appear in either the Blood or Vinegar assessments, and stage G ("Let students practice with equipment") was not included in the Vinegar task. (See Figures 2 and 3.) Maybe the teams decided that some of these actions were trivial for the assessments or were covered by other actions of the shell. Another type of deviation suggests that the shell allows too much room for choice. For example, in stage R, students are supposed to be asked to "suggest possible alternative solutions." The developers of the Blood task gave students some choices, but the developers of the Vinegar task chose to ask students to explain the steps that led them to their solution (see Figures 4 and 5). These differences in interpretation of the shell may have affected the correlations among tasks. Since no student took two tasks created by the same team, there was no way to assess possible developer effects in the analysis.

Discussion. The increasing use of open-ended measures in large-scale (and often high-stakes) testing programs has made it imperative that we develop a better understanding of just what these tasks measure and the factors that affect scores on them. The good news for those directing testing programs is that the mean correlation among tasks (.60) is substantially greater than what was expected on the basis of several previous studies. This finding suggests that reliable scores can be obtained with hands-on measures in far less testing time and at lower cost than what had been anticipated. This is especially important given the resources needed to construct, administer, and score these tasks relative to the resources needed for multiple choice exams (Stecher & Klein, 1997).

The fact that the correlations among tasks are not related to subject matter area further suggests that test developers will not have to be overly concerned about having to sample tasks systematically from several discrete areas within a content domain; i.e., because all of the tasks behave pretty much alike regardless of their design specifications and subject matter area focus.

Content, Format and Inquiry Level

There are a number of plausible explanations for our finding that performance tasks that appear to be very similar in format, content, and inquiry level generally correlate no higher with each other than they do with measures that use quite different formats, are designed to assess different skills and knowledge, and provide different amounts of guidance to their solution. It may be that the students who mastered the specific skills and knowledge required for one task also tend to acquire the skills and knowledge needed for the other tasks. These may co-exist in students' instruction, although it seems unlikely that the eighth grade students we tested had comparable exposure to the diverse topics measured.

A more likely explanation for the limited effects of content, format, and level of inquiry is that the scores on all the measures were highly influenced by a common underlying general academic ability. It is certainly true that all the measures we used relied heavily on a student's ability to read and understand the directions and questions. If reading is the source of the relatively high inter-task correlations, then it would suggest that task developers need to pay more attention to the verbal demands of tasks, possibly placing greater emphasis on diagrams and pictorial representations where those are appropriate. Whatever the reason, the lack of differentiation in scores due to format, content or level of inquiry in our research raises important questions about what open-ended tasks in science truly measure.

Another complication comes from the fact that the task generation directions provided by the shells were not specific enough to insure that they would be interpreted in the same way by different task development teams. While this variation did not change the emphases of the tasks in terms of content, format or level of inquiry, it did create other unintended differences of unknown importance. Klein, et al. (1997) found that shells accounted for less than 20% of the variability in scores on hands-on tasks. In that study, the same team developed all the tasks from a given shell. This research suggests that using different teams to develop tasks from a shell would lead to even greater variability among tasks and a lower estimate of shell effects.

In our study, the interpretation and preferences of the assessment developers as well as the characteristics of the approach they adopted (such as the equipment they chose to use) played a significant role in the creation of each task. These factors also imposed a limit on the capabilities of the shell to prescribe what had to be done and how it was to be done. We think that before shells can be used efficiently to generate assessments *en masse*, several improvements must be implemented. Solano-Flores and Shavelson (1997) suggest a number of such changes, including refining the conceptual framework for developing science performance assessments and increasing the specificity of the directions to test developers. Nevertheless, test development is as much an art as a science. We suspect that no set of rules, however well documented, will fully capture all of the elements of a well-constructed task.

Table 1
FORMAT, CONTENT AND INQUIRY LEVEL
OF OPEN-ENDED SCIENCE TASKS

Task	Format	Content	Inquiry Level
Vinegar Discovery	Hands-on	Acids and Bases	Unguided
Vinegar Recipe	Hands-on	Acids and Bases	Guided
Vinegar Text	Paper-and-pencil	Acids and Bases	Guided
Alien Blood Discovery	Hands-on	Acids and Bases	Unguided
Alien Blood Recipe	Hands-on	Acids and Bases	Guided
Alien Blood Text	Paper-and-pencil	Acids and Bases	Guided
Lever	Hands-on	Force and Motion	Guided
Materials	Hands-on	Properties of Materials	Guided

Table 2
NUMBER OF STUDENTS IN EACH COMBINATION OF A VINEGAR AND A BLOOD TASK

First Acid/Base Task	Second Acid/Base Task	Number of Classes	Number of Students
Vinegar Discovery	Blood Discovery	3	97
	Blood Recipe	2	69
	Blood Text	2	68
Blood Discovery	Vinegar Discovery	3	90
	Vinegar Recipe	2	73
	Vinegar Text	2	55
Vinegar Recipe	Blood Discovery	3	98
	Blood Recipe	2	57
	Blood Text	2	61
Blood Recipe	Vinegar Discovery	3	94
	Vinegar Recipe	2	63
	Vinegar Text	2	51
Vinegar Text	Blood Discovery	2	58
	Blood Recipe	2	68
	Blood Text	2	59
Blood Text	Vinegar Discovery	4	136
	Vinegar Recipe	2	65
	Vinegar Text	2	51
TOTAL		42	1,493

Table 3
CORRELATIONS AMONG ALL MEASURES

	Blood					
Measure	Discovery	Recipe	Text	Levers	Materials	ITBS
Vinegar Discovery	.50	.51	.56	.62	.50	.54
Vinegar Recipe	.42	.63	.58	.49	.57	.54
Vinegar Text	.58	.66	.74	.59	.51	.66
Levers	.49	.53	.51	1.00	.50	.56
Materials	.46	.51	.52	.50	1.00	.53
ITBS	.51	.55	.58	.56	.53	1.00

The Ns for correlations among acid/base tasks ranged from 87 to 170. The correlation of one acid/base task (e.g., Vinegar-Discovery) with Levers, Materials, or the ITBS was based on about 250 students. The correlations between the latter three measures were based on over 1,000 students. The correlations are based on all students who took each pair of tasks, regardless of the order in which the tasks were completed.

Figure 1
SHELL FOR RECIPE VERSION OF COMPARATIVE INVESTIGATION

Step Action

PERFORMING:

A Provide equipment
B Provide independent variable
C Provide irrelevant variables
D Describe how equipment is used
E Introduce variable names
F Include diagrams
G Let students practice with equipment
H Provide a problem/hypothesis involving independent variable
I Ask students to describe what they will be looking for to solve the problem/hypothesis involving independent variable
J Provide step-by-step instructions on how to conduct experiment to solve/test the problem/hypothesis involving independent variable
K Ask students to take notes as they conduct their experiment

INTERPRETING:

L Ask students to (1) rearrange, or (2) transform, or (3) collapse, or (4) compute, or (5) synthesize their results in a labeled table/graph/diagram given to them in order to show the relationship between the independent variable and the outcome
M Ask students to draw a conclusion about the experiment and the relationship found
N Ask students to draw a general conclusion about the relationship involving the independent variable

APPLYING:

O Provide a concrete, meaningful context
P Create a scenario that involves the scientific concept of interest
Q Provide either a "pure science" problem (e.g., description, measurement, classification) or a problem of social or practical interest (e.g., water pollution) whose solution can be accomplished by using part or all of the knowledge previously taught on the same domain of science knowledge
R Ask students to

1. show a product for the solution of the problem, or
2. give the steps that led them to the solution, or
3. identify the advantages of the solution
4. suggest possible alternative solutions

Figure 2
BLOOD TASK: STEPS A-G (Team 1)

INSTRUCTIONS

In this activity you will be working by yourself. You can write your answers directly on these pages. If you have a question, please raise your hand and we will come to help you.

Please take the materials out of the bag in front of you. Put the materials on your placemat. Raise your hand if you are missing any of these materials:

MATERIALS

1 bag pH indicator paper strips
1 pH Color Chart
8 plastic measuring cups
1 dropper bottle **Solution** X

1 dropper bottle **Alien Blood**
1 dropper bottle **Medicine A**
1 dropper bottle **Medicine B**
1 dropper bottle **Medicine C**

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	Blue
ACIDS			NEUTRAL	BASES	

Figure 3
VINEGAR TASK: STEPS A-G (Team 2)

Acids and Bases - Form R

EQUIPMENT: You will need the following materials. Raise your hand if you are missing any of these materials:

1 bottle labeled INDICATOR
1 bottle labeled BASE X
1 bottle labeled ACID A
1 bottle labeled ACID B
1 bottle labeled ACID C

3 plastic cups
1 placemat
Safety goggles
Paper towels for spills

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

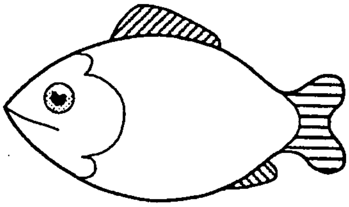
____

GO TO NEXT PAGE

Figure 4
BLOOD TASK: STEPS Q AND R (Team 1)

Part 5: USING WHAT YOU LEARNED

The people of Spring City were concerned because the fish in their pond were dying. They hired an environmental scientist who measured the pond's pH and found that it was too acidic. Pond fish need neutral water to survive. The people followed the specialist's advice and added Pro-Base, (a strong base) to the pond. After two days, the fish stopped dying. The people kept adding Pro-Base to the water and after three more days, the fish started dying again. In fact, the more Pro-Base they added, the more fish died.



5a. Why did ProBase work at first, but not continue to work?

____

____

____

____

____

5b. What should the people in Spring City do now to save the fish in their pond? (Circle the best choice).

A) Add no more chemicals

B) Add an acidic substance

C) Add a neutral substance

D) Add more Pro-Base

5c. Why did you choose this answer?

____

____

____

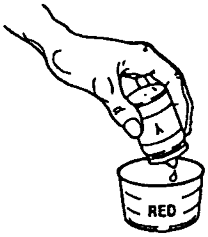
Figure 5
VINEGAR TASK: STEPS Q AND R (Team 2)

4e. Sally has a bottle of Base Y and a bottle of Base Z. To find out which base is stronger:

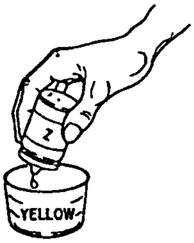
- She puts 7 drops of Base Y, 7 drops of Indicator, and 10 drops of Acid B into a cup. The solution in the cup turns red.

- Into a new cup she puts 7 drops of Base Z, 7 drops of Indicator, and 10 drops of Acid B. The solution in this cup turns yellow.

Which base is stronger--Base Y or Base Z?



Base Y +
Indicator + Acid B



Base Z +
Indicator + Acid B

4f. How do you know this?

Figure 6
95% CONFIDENCE INTERVAL AND MEAN SCORES* ON DISCOVERY TASK TAKEN SECOND BY FORMAT OF TASK TAKEN FIRST

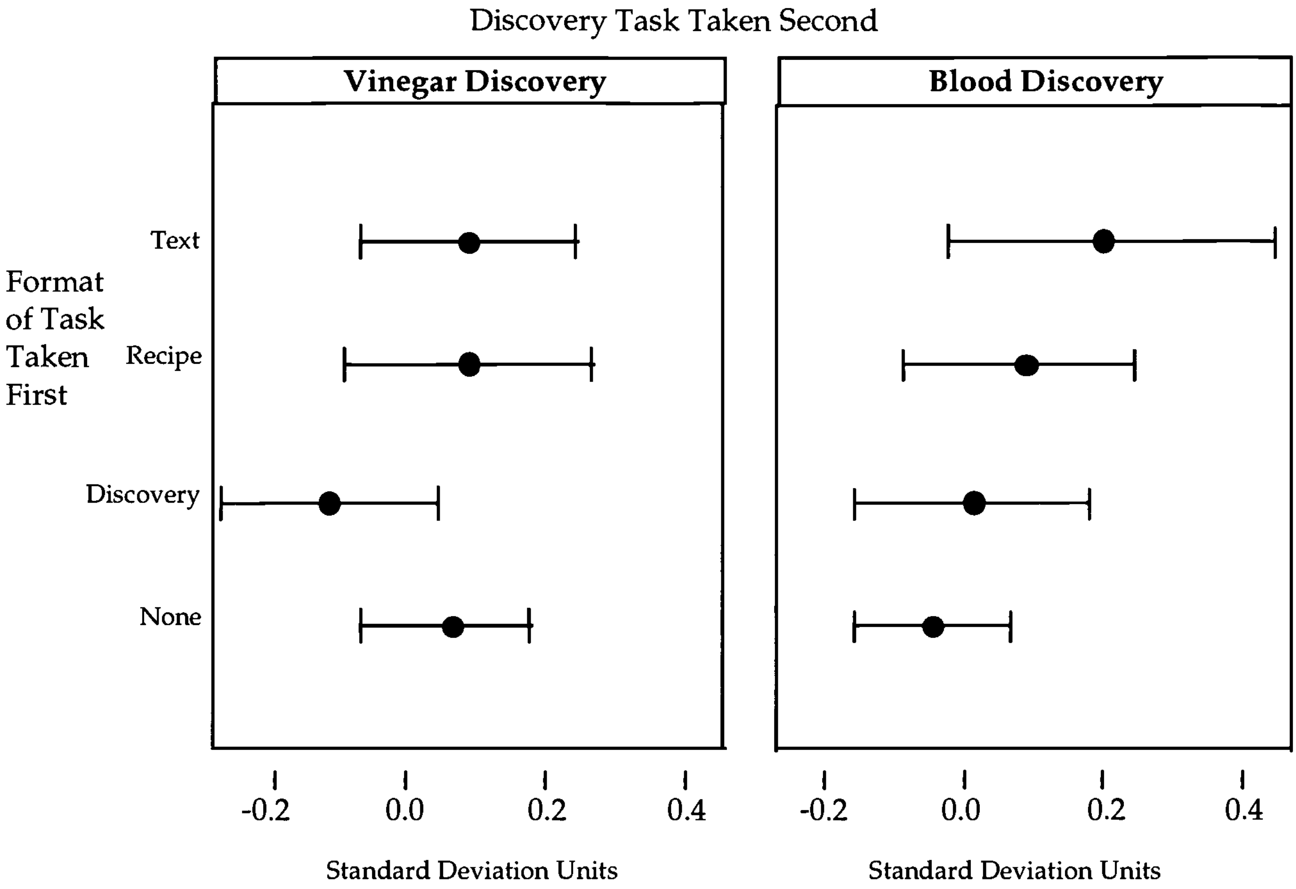


Fig. 6A

Fig. 6B

Note: *Adjusted for ITBS Science score.

References

Aschbacher, P. R. (1991). *Alternative assessment: State activity, interest and concerns.* CSE Technical Report 322. Los Angeles: UCLA Center for Research on Evaluation, Standards and Student Testing.

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

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A., and Valadez, J. R. (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal for Research in Mathematics Education*, 24(3), 190-216.

Bond, L. A., Braskamp, D., van der Ploeg, A., and Roeber, E. (1996). *State student assessment programs database: School year 1994-95.* Oak Brook, IL: North Central Regional Educational Laboratory.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement*, 32(4), 385-396

Burger, S. E. and Burger, D. L. (1994, Spring). Determining the validity of performance-based assessment. *Educational Measurement: Issues and Practice*, 13(1), 9-15.

Cronbach, L. J., Bradburn, N. M., and Horvitz, D. G. (1994, July). *Sampling and statistical procedures used in the California Learning Assessment System: Report of the Select Committee*. Stanford: Author.

Dunbar, S. B., Koretz, D. M., and Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4(4), 289-303.

Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System, 1991-1994. Final Report.* Frankfort: Office of Educational Accountability.

Hamilton, L. Klein, S. Stecher, B. M., McCaffrey, D. & Comfort, K. (1998). "Effects of Practice on Hands-On Science Assessments." Paper to be presented at the 1998 annual meeting of the American Educational Research Association, San Diego.

Hieronymous, A. N., Hoover, H. D, Cantor, N. K., and Oberly, K. R. (1987). *Writing Teacher's guide, Iowa Tests of Basic Skills, Levels 9-14, Forms G/H.* Chicago: Riverside.

Hieronymous, A. N., Hoover, H. D., Frisbie, D. A., and Dunbar, S. B. (1990). *Manual for School Administrators: Supplement, Iowa Test of Basic Skills, Levels 5-14, Form J.* Chicago: Riverside

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

Hoover, H. D., Hieronymous, A. N., Frisbie, D. A., and Dunbar, S. (1995). *Integrated Writing Skills Test: Score conversions and technical summary.* Chicago: Riverside.

Klein, S. P., Shavelson, R. J., Stecher, B. M., McCaffrey, D., and Haertel, E. (in press). Shell effects on hands-on tasks.

Koretz, D., Stecher, B. M., Klein, S. and McCaffrey, D. (1994, Fall). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practices*, 13(3), 5–16.

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

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

Shulman, L. S., and Tamir, P. Research on teaching in the natural sciences. In R. Travers (Ed.), *Second handbook of research on teaching* (pp. 1098-1148.) Chicago: Rand McNally.

Solano-Flores, G. and Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical and logistical issues. *Educational Measurement: Issues and Practices*, 16(3), 16-24.

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

Tamir, P. R. and Doran, R. L. (no date). Science process skills in six countries: Second IEA science study. International Association for the Evaluation of Educational Achievement.

Tamir, P. R., Nussinovitz, R. and Firedler, Y. (1982). The design and use of a practical tests assessment inventory. *Journal of Biological Education*, 10(1), 42-50.

TM029302

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: DO CONTENT, FORMAT, AND LEVEL OF INQUIRY AFFECT SCORES ON OPEN-ENDED SCIENCE TASKS?

Author(s): B. Stecher, S. Klein, G.Solano-Flores, D. McCaffrey, A. Robyn, R. Shavelson,E Haertel

Corporate Source: The RAND Corporation

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Shirley Hall

Printed Name/Position/Title: Shirley Hall, Contract Administrator

Organization/Address:
The RAND Corporation, P.O. Box 2138,
Santa Monica, CA 90407-2138

Telephone: 310-393-0411 x6291

FAX: 310-451-6973

E-Mail Address: shall@rand.org

Date: 10/21/98

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com