DOCUMENT RESUME

ED 350 353 TM 019 134

AUTHOR Semple, Brian McLean
TITLE Performance Assessment: An International Experiment.
INSTITUTION Educational Testing Service, Princeton, NJ. Center for the Assessment of Educational Progress.
SPONS AGENCY Carnegie Corp. of New York, N.Y.; National Center for Education Statistics (ED), Washington, DC.; National Science Foundation, Washington, D.C.
REPORT NO ETS-R-22-CAEP-06; ISBN-0-88685-127-0
PUB DATE Jul 92
CONTRACT IAD-91-0222; SDE-8955070
NOTE 61p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Tests/Evaluation Instruments (160) -- Statistical Data (110)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Achievement; Adolescents; Comparative Testing; Educational Assessment; Foreign Countries; *International Studies; Junior High Schools; *Junior High School Students; *Mathematics Achievement; *Problem Solving; *Sciences; Scores; Student Evaluation; Test Construction; Test Results
IDENTIFIERS *International Assessment of Educational Progress; *Performance Based Evaluation

ABSTRACT

The second International Assessment of Educational Progress focused on the mathematics and science achievement of 13-year-olds. Performance assessments were used as part of the overall assessment in four countries (England, Scotland, Soviet Union, and Taiwan) and five Canadian provinces. The performance assessment approach drew heavily on the experience of the United Kingdom in such assessments, but added features to meet the needs of an international study. The performance tasks required students to apply concepts, observe, measure, manipulate equipment and materials, and record and interpret data. Approximately 3,000 students participated in the 1991 mathematics and science assessments. Scores varied widely from task to task and from country to country. Relative performance of countries and provinces generally differed from those identified by written curriculum-based tests. A major lesson learned from the experimental test administration is that this form of performance assessment can be used reliably in international comparative studies although at an estimated cost three to four times greater than that for an equivalent number of written test questions. Sample mathematics and science tasks are included, and there is an appendix on problem solving in mathematics that describes outcomes produced by Scottish students. (SLD)

PERFORMANCE ASSESSMENT:

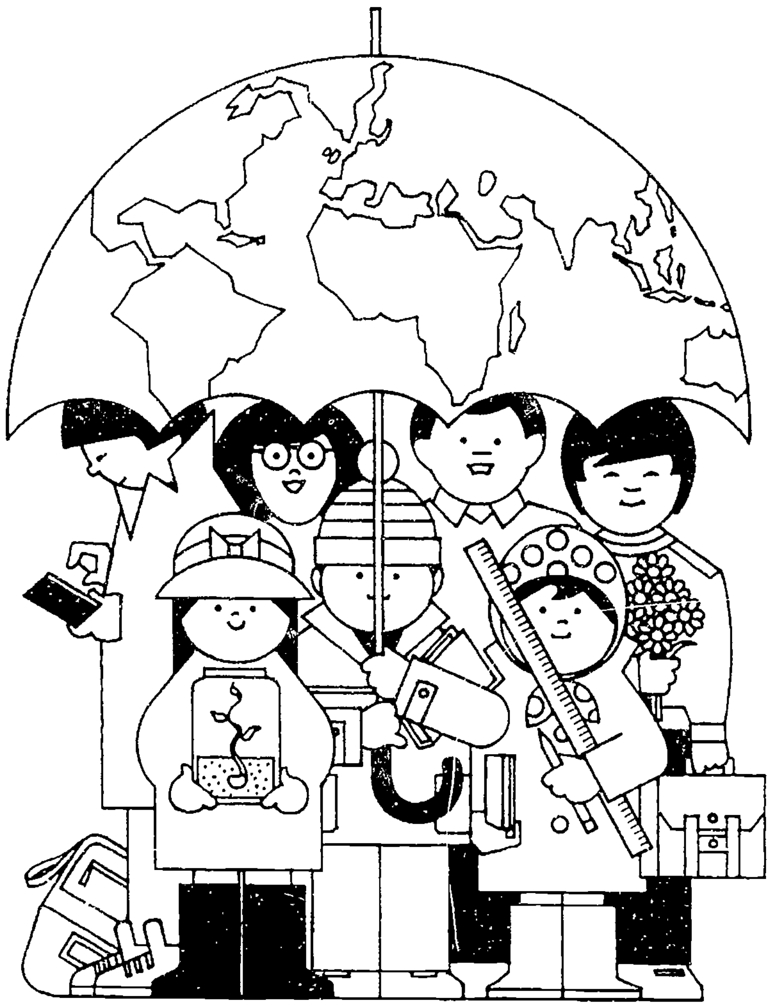
The International Assessment of Educational Progress

IAEP

EDUCATIONAL TESTING SERVICE

PERFORMANCE ASSESSMENT:

AN INTERNATIONAL EXPERIMENT



Brian McLean Semple
The Scottish Office
Education Department

Prepared for the National Center for Education Statistics
U.S. Department of Education and the National Science Foundation

July 1992 Report No. 22-CAEP-06

The International Assessment of Educational Progress

IAEP

EDUCATIONAL TESTING SERVICE

Educational Testing Service (ETS) is a private, nonprofit corporation devoted to measurement and research, primarily in the field of education. It was founded in 1947 by the American Council on Education, the Carnegie Foundation for the Advancement of Teaching, and the College Entrance Examination Board.

The Center for the Assessment of Educational Progress (CAEP) is a division of ETS devoted to innovative approaches to the measurement and evaluation of educational progress. The present core activity of CAEP is the administration of the National Assessment of Educational Progress (NAEP), under contract from the U.S. Department of Education. CAEP also carries out related activities, including the International Assessment of Educational Progress (IAEP), state assessments, and special studies such as the National Science Foundation-supported Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics.

The work upon which this publication is based was performed pursuant to Grant No. SDE-8955070 of the National Science Foundation, and additional funding was provided by the U.S. Department of Education through Interagency Agreement No. IAD-91-0222. Supplementary funds were provided by the Carnegie Corporation of New York.

This report, No. 22-CAEP-06, can be ordered from the Center for the Assessment of Educational Progress at Educational Testing Service, Rosedale Road, Princeton, New Jersey 08541-0001.

Library of Congress Card Number: 92-71732

ISBN: 0-88685-127-0

Educational Testing Service is an equal opportunity/affirmative action employer.

Contents

Overview 1

- Why performance assessment? 1
- What type of performance assessment? 2
- What was assessed? 3
- Who was tested? 5
- What information was provided? 6
- How well did students perform? 6
- What lessons did we learn? 9
- How did teachers and students react? 12
- What can students do in performance assessment? 13

Mathematics Tasks 15

Science Tasks 35

Appendix: Problem Solving in Mathematics 53

Acknowledgements 59

Overview

Why performance assessment?

The main curricular emphasis in the second International Assessment of Educational Progress (IAEP) was on mathematics and science. The extensive use of pencil-and-paper tests in the main IAEP assessments made it possible to achieve good coverage of the knowledge and skills which could be assessed using such test instruments.[1] However, an analysis of the mathematics and science curricula in most of the countries involved in IAEP showed that they included at least some skills and processes which could not be assessed adequately with pencil-and-paper tests alone.

The potential value of performance assessment as a supplement to pencil-and-paper tests was recognized by the IAEP assessment developers. Experience in the United Kingdom with national educational monitoring by the Assessment of Performance Unit (APU) in England and Wales and the Assessment of Achievement Programme (AAP) in Scotland had demonstrated that some types of performance assessment were feasible, in practical and cost terms, in national surveys of student attainment.

[1]Archie E. Lapointe, Nancy A. Mead, and Janice M. Askew. *Learning Mathematics*. Princeton, NJ: Educational Testing Service, 1992.

Archie E. Lapointe, Janice M. Askew, and Nancy A. Mead. *Learning Science*. Princeton, NJ: Educational Testing Service, 1992.

Given the UK experience and the desirability of extending the curriculum coverage in IAEP, it was decided to include a limited, optional component of performance assessment in the 1991 survey. The assessment was developed for 13-year-old students only and included mathematics and science tasks to enable IAEP participants to experiment with performance assessment in an international context.

What type of performance assessment?

The experimental nature of the performance assessment in IAEP required that the approach used be based on existing best practice. Consequently, the approach and the test materials used drew heavily on the UK experience but added essential amendments to meet the needs of an international study. For instance, they had to be robust enough to be valid in a variety of different curricular contexts and to be capable of operation with limited equipment and materials by staff who had no prior experience in this type of assessment.

The approach decided on was a series (or circuit) of stations, each involving a short task (or tasks), which students would carry out under the supervision of a trained assessor. The tasks required students to demonstrate practical skills, such as measurement or observation, and provided a more realistic context than a written test for assessing cognitive skills, such as inferring or hypothesizing. The activity at each station was designed to be completed by students in about five to eight minutes. Kits of standard pieces of equipment and materials, including master copies of diagrams, were supplied to all of the countries participating in the performance assessment. Standardized instruction manuals and scoring guides were supplied for administering and scoring the tasks.

Procedures and tasks for the performance assessment were pilot-tested in May 1990. The final design of the performance assessment had two circuits, each containing eight stations. One circuit consisted of mathematics tasks and the other of science tasks. The two circuits could be administered by one assessor in parallel or in series, each accommodating separate samples of six students. Students were allowed to spend as much time at each station as they needed. Assessors only asked individual students to move on if they appeared to have completed all that they could. This flexibility and the availability of more stations in a circuit than students working on it was intended to avoid queuing for stations while providing time for students to deliver their best performance.

The outcomes included artifacts, drawings, and written responses, a total of about 36 outcomes from the 16 stations. Only two outcomes, both in the science circuit, were marked on the spot by the assessor. The remainder were scored later by trained scorers.

What was assessed?

The performance tasks required students to apply concepts, observe, measure, manipulate equipment and materials, and record and interpret data. The tasks were designed to correspond with aspects of the main assessment framewc.k and the curricula of the participating countries. However, classification by content and process was complicated because many of the tasks required a series of steps and could be solved in a number of ways. The outcomes of the performance assessment tasks reflected only a portion of the overall assessment framework, as do all tests of broad curricular areas. However, it must be borne in mind that only a subset of the elements in the overall framework included practical skills or skills that benefit from a practical context.

The mathematics tasks tended to concentrate on aspects of Measurement and Geometry, which together constituted 35 percent of the agreed-upon IAEP assessment framework. This was for two main reasons: elements of these two aspects of mathematics could be assessed only by using performance tasks; and focussing on these elements made it possible to obtain measures of performance on sets of related skills. The tasks also assessed conceptual understanding, procedural knowledge and problem solving, which represented the other dimension of the assessment framework.[2]

In science, the tasks were drawn mainly from aspects of the Physical Sciences and Nature of Science, which together constituted 50 percent of the overall assessment framework. This was partly because of restrictions on sending biological material, such as potting soils or seeds, between countries and partly because few good tasks in the Life or Earth and Space Sciences were available. The tasks did, however, assess the other dimensions of the assessment framework: knows facts, concepts and principles; uses knowledge to solve simple problems; and integrates knowledge to solve more complex problems.

It is worth noting here that performance assessment frameworks may vary more from country to country because of differences in the countries' curricula than do assessment frameworks for traditional pencil-and-paper tests. For instance, the main categories for science at age 13 in the UK national monitoring programmes were:

APU - England and Wales	**AAP - Scotland**
Using symbolic representations	Observing*
Using apparatus and measuring instruments*	Measuring*
Using observation*	Handling information*
Interpretation and application	Using knowledge*
Design of investigations	Using simple procedures*
Performance of investigations*	Inferring*
	Investigating*

* Aspects in which some form of performance assessment is conducted.

Such differences have implications for the emphasis given to performance assessment, as opposed to pencil-and-paper tests, and for the type of tasks used.

[2] Center for the Assessment of Educational Progress. *The 1991 IAEP Assessment: Objectives for Mathematics, Science, and Geography*. Princeton, NJ: Educational Testing Service, 1991.

Who was tested?

Nine of the IAEP participants, four countries and five Canadian provinces (two including separate English- and French-speaking populations), administered the performance assessment tasks. After careful consideration, the United States decided not to participate in this international experiment, but rather to evaluate the results of the assessment before taking part in future comparative studies of this type.

A subsample of about one quarter of the schools involved in the IAEP pencil-and-paper testing were randomly selected for the performance assessment. Likewise, the performance tasks were administered to a subsample of the students in those schools who had taken the IAEP pencil-and-paper tests, to ensure some common base of experience and permit comparison of written test and performance task results. About two-thirds of the original students in the selected schools took part in the performance assessment. The participants and their achieved samples were:

Participants	**Number Assessed in Mathematics**	**Number Assessed in Science**
Canada		
Alberta	179	180
British Columbia	261	240
Nova Scotia	258	258
Ontario-English	326	329
Ontario-French	305	319
Saskatchewan-English	381	377
Saskatchewan-French	224	218
England[3]	158	131
Scotland	283	283
Soviet Union[4]	370	380
Taiwan	315	316

[3]The combined school and student participation rate for the original sample was very low (47 percent), so the results for England presented in this report may reflect nonresponse bias.

[4]Since the assessment occurred in 1991, the results are applicable to the former Soviet Union; the country is thus referred to as the Soviet Union.

In addition to their limited size, it should be noted that some countries' school samples were restricted geographically in order to contain costs. For instance, the schools in Scotland were selected from those in the central belt of the country only.

What information can be reported?

The performance assessment experiment provided two types of information. First, although the student samples were limited and, therefore, could not produce accurate measures of performance of total country (or province) populations, the process did provide a rich source of information on how a sample of students perform on practical tasks. Second, the experiment also provided information about the strengths and weaknesses of the particular approach used.

How well did students perform?

There are three general points worth emphasizing before the findings are considered in more detail.

- Scores varied widely from task to task, suggesting that the measures tap a range of skills and knowledge.
- Scores on the various tasks varied significantly from country to country (and from province to province) in systematic ways, indicating real differences in performance between the various populations.
- The relative performances of countries and provinces were generally different from those identified by the written tests covering related curricular areas. This suggests that using "hands-on" methods of assessment allowed students to demonstrate their skills in ways that were not possible with traditional paper-and-pencil tests.

The following describes some of the main findings from the performance assessment. More detailed results are provided in the second and third parts of this report.

- **In measurement skills, the main question was not one of accuracy of measurement but whether decisions on what to measure (or how to calculate answers from measurements) were correct.** Across all participating countries and provinces, about 40 percent of students measuring the perimeter of a rectangle provided correct answers within plus or minus 4mm. Those who were incorrect tended to be a long way out, probably due to over- or under-counting the rectangle's sides or by misreading cm and mm on the ruler. The task of estimating irregular areas using a grid square produced a similar outcome, with reasonably accurate answers produced by most students but wildly inaccurate ones by the rest. **Those giving wrong answers when measuring angles with a 180° protractor generally gave the size of the "reverse" angle** (i.e., in the case of the acute and obtuse angle, 180° minus the size of the angle, and in the case of the reflex angle, 360° minus the size of the angle).

- **In problem-solving tasks based on geometry, most students performed well.** Difficulties in producing and handling shapes were few, although more difficulties were encountered with more complex shapes. However, in a task requiring the use of a pinboard (or geoboard) to identify different-size squares, **the square with a "diamond" orientation was overlooked by many students.**

- One problem-solving task based on weighing showed that students had difficulty weighing accurately. **However, of greater interest was whether students used a precise method for solving the problem or an estimation approach. In most cases, the latter predominated.** For details of the range of strategies adopted for this task, see the Appendix.

- Three of the science tasks required objects to be categorized in terms of electrical conductivity, magnetism, and appearance, respectively. The first two of these required students to carry out tests on the objects. The categorization was accomplished successfully by most students and most also provided a satisfactory explanation for their categorization. **It seems, therefore, that most 13-year-olds can carry out simple tests systematically and that they have at least a basic understanding of conductivity and magnetism.** In the third task, most students could categorize correctly, but only half or fewer could provide a satisfactory explanation.

- Two of the science tasks required students to follow instructions provided mainly in diagrammatic form. In both tasks high success rates were achieved and the required artifacts, an electrical circuit and filtering apparatus, were assembled correctly by most pupils. **It seems that by age 13, most students can utilize scientific apparatus, even when they are not familiar with it, when clear instructions are provided.** In this case the less able students were probably helped by the very limited reading requirements of the tasks.

- One of the science tasks was a simple investigation in which the starch and/or glucose content of three solutions was determined by using chemical indicators. Across the participating countries and provinces, about two-thirds of the students identified the glucose solution, fewer than 60 percent identified the starch solution and less than half recognized that the third solution contained both. **False positives proved to be a problem, perhaps because some students regarded any change in the indicator, even intensity of colour, as the required change. This probably reflects unfamiliarity with the use of chemical indicators. Problems in following procedures were minimal.**

- A task designed to assess whether students could differentiate between statements of fact and deductions involved two small plastic objects, each in its own jar, one object floating and one submerged in a clear liquid. **Most pupils selected the two factual statements from the five statements provided but more than half also selected a deduction**, that the submerged object was heavier. In fact, the deduction was incorrect, as the objects were identical but the density of the liquids differed. **This result mirrors a tendency 13-year-olds displayed in similar written test questions, where they made plausible deductions when asked to make only observations.**

- A task designed to test visual, auditory and olfactory observations involved dissolving a fruit-flavoured tablet in water and recording what happened. Most students noted changes in the tablet's size and in the colour of the water and that gas (i.e., bubbles) was emitted. Only a minority noted that the tablet moved, that the transparency of the water decreased, that there was a fizzing sound, and that there was a smell of fruit. **The results may reflect students' interpretation of the word "observe" in the question as being a visual cue and p rhaps also a tendency to regard their visual senses as predominant in a scientific context.**

What lessons did we learn?

The main lesson was that this form of performance assessment can be used reliably in international comparative studies, although at an estimated cost three to four times greater than for an equivalent number of written test questions. With careful task development and pilot-testing, tight quality control of equipment and materials, and clear instructions, performance tasks can be designed and administered to provide standardized results in curricular areas which cannot be assessed adequately with written tests. The emphasis educators place on these curricular areas will dictate whether this type of performance assessment justifies the extra resources required.

Equipment and materials. The equipment and materials used in the performance assessment were constrained in a number of ways. Since each participant needed to purchase several kits which included the necessary equipment and materials, it was necessary to use mostly low-cost items. The need to send kits to various countries required them to be reasonably compact, light in weight, and robust. Kits also had to comply with customs and duty regulations. In addition to these practical issues, the equipment had to be either familiar to students or simple enough to be used with minimal instructions. Also, since schools were asked to provide only flat work surfaces and a supply of water, equipment needed to meet these constraints as well.

How well did we do with respect to these factors? By using combinations of short tasks for our assessment, we proved it is possible to produce portable kits of reasonably robust equipment. The kits were assembled centrally and distributed in appropriate numbers to each participating country, to ensure that the equipment used in the assessment was standardized and identical in all countries and provinces.

Even with careful preparation, three major problems arose with equipment and materials used in the performance assessment. As a result, a few results had to be discarded and some others had to be re-analyzed.

The first of these problems arose during the reproduction of students' booklets in the various countries. This was done by photocopying processes, which tend to produce small changes in scale, a feature warned about in the performance assessment guidelines. The resulting nonstandard sizes of diagrams (in two of the stations) in some countries and provinces required reanalysis of their results, using the actual rather than the intended sizes. This problem could be resolved in future surveys by supplying all copies of diagrams from a central source.

The second problem was similar in nature. Plastic containers with specific volumes marked on them were used in one station and some of these proved to be nonstandard, a few by a substantial amount. It was possible to take into account the equipment variation in the analysis of the first task in this station, but results from the second had to be discarded. In future surveys, all such apparatus should be checked, as significant production variations can occur even in something as simple as rulers.

The third problem concerned the task which required the categorization of different types of seeds. As biological material could not be included in the kits, each country and province was required to obtain the specified seeds. In some cases this proved impossible and, despite some attempts at improvisation, it became clear that the results from different testing locations for one group of seeds were not comparable. The results were subsequently discarded.

Two minor difficulties arose related to students' understanding of the instructions that accompanied the equipment and material. First, there was a translation problem with the students' instructions in the Soviet Union for the task called *Leaves*. This was corrected during the assessment, but the results are not comparable with those obtained from other countries. Second, one set of instructions used the word "enable," which was misinterpreted by some students in one population as "unable" (despite the fact that the sentence involved did not then make sense). In this situation, students who provided the opposite response along with an appropriate explanation were counted as giving correct answers.

Administration and organization. The performance assessment tasks were designed to be administered by assessors running the two circuits either in parallel or in sequence and, in practice, they were administered in both ways.

The assessors in some countries and provinces were members of the teaching staff in the sample schools and in other locations, they were not, although most were trained teachers. Each assessor received training about the performance assessment tasks, their administration and their scoring.

The administration of the tasks raised few problems beyond those already mentioned. In some smaller schools space was limited and the circuits had to be set up wherever spac' was available. The lack of a water supply in some schools had been anticipated, so testing needs could be met with a large container.

How did teachers and students react?

The message here was clear and unequivocal. Teachers involved in the performance assessment and others who only observed it in progress expressed enthusiasm for the approach and interest in the equipment and materials used. Many said they were impressed by the tasks and some suggested that such activities could and should be integrated into classroom work. In fact, this is being done in at least one Canadian province as well as in the UK. In Taiwan the performance assessment tasks have since been used with students other than those involved in IAEP.

Teachers who observed students they knew doing the tasks were sometimes surprised at how the students tackled them and what they were able to achieve. **Generally, students were enthusiastic and willing to try all of the tasks, regardless of their individual abilities, and in some cases, students regarded as low attainers achieved more than was expected.** The fact that all students could, at least, attempt the tasks and achieve something seemed to eliminate the feeling of failure that low-attaining students sometimes experience in taking written tests. Secondary analysis of the written test and performance assessment data sets should enable us to clarify whether there are differences in the relative performance of low-attaining students.

The reaction of students was almost universally favourable. Many assessors reported that students said they had enjoyed doing the tasks and remarked on how well motivated they appeared to be in undertaking them. **The lack of any significant problems in completing tasks, when they were essentially untimed, may well be attributable to the positive attitudes of the students and to their self-discipline in this type of activity, as well as to the work of the assessors.**

What can students do in performance assessment?

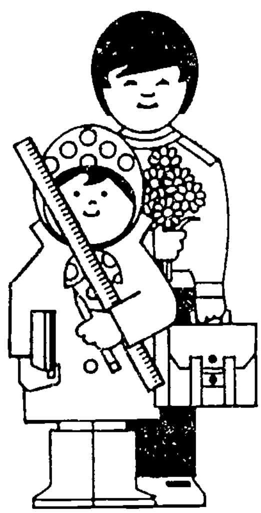
The following section of the report describes what students achieved in the performance assessment tasks. In order to ensure that the results can be understood readily, full details of each task are provided. It should be noted, however, that the text of the tasks has been abbreviated and the diagrams have been altered in size and presentation. In keeping with the experimental nature of the performance assessment, information is also provided on the few tasks which did not work as expected.

The tasks are described in the following order:

Content Area	Name of Task	Skills Assessed
Mathematics	Perimeter	Measurement skills
	Ticket	Measurement skills and problem solving
	Angles	Measurement skills
	Leaves	Measurement skills
	Pinboard	Geometry concepts and problem solving
	Triangles	Geometry concepts
	Scissors	Geometry problem solving
	Clay	Measurement skills and problem solving
	Water	Measurement skills
Science	Light-up	Physical science concepts and skills
	Circuit	Physical science concepts and skills
	Filter	Nature of science skills
	Magnet	Physical science concepts and skills
	Indicators	Physical science skills
	Float	Nature of science skills
	Tablet in Water	Nature of science skills
	Seeds	Life science problem solving

The results presented for each task that follow are weighted percentages of correct responses or responses of a particular type. Next to each printed statistic in parentheses is an estimate of sampling error (standard error). It is especially important to consider the imprecision in the estimates when comparing two populations with similar results.

Mathematics Tasks



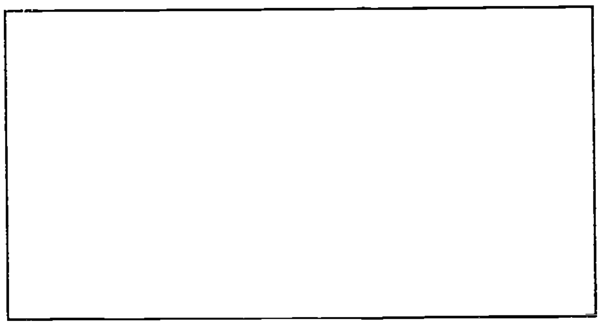
PERIMETER

Task Descriptor

To measure the perimeter of a rectangle.

Equipment/Material

A centimetre/mm ruler and a rectangle (119 mm by 72 mm) printed in the student's booklet.



Student Instructions

Measure the distance around the rectangle to the nearest mm.

Scoring Scheme

Credit for correct measurement ± 4mm.

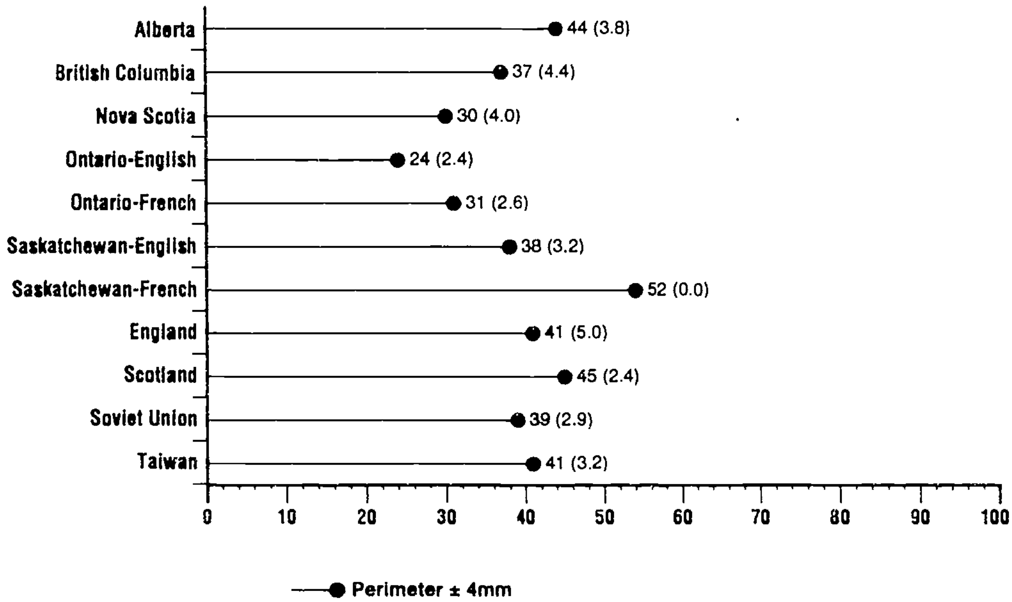
Problems

In some countries photocopying expanded the size of the rectangle to be measured. To allow for this, all answers within ± 4mm of the actual size were given credit.

Comments

- There was a wide range of performance among the Canadian provinces, with scores ranging from 24 to 52 percent, whereas the scores in the other countries were more similar.
- In Ontario and Saskatchewan, differences in the performance of the English- and French-speaking populations favoured the latter.
- A sizable number of students (17 percent across all participating countries and provinces) gave the measurement in cm although mm was specified in the instructions.

Percentage of Correct Responses (with Standard Errors)



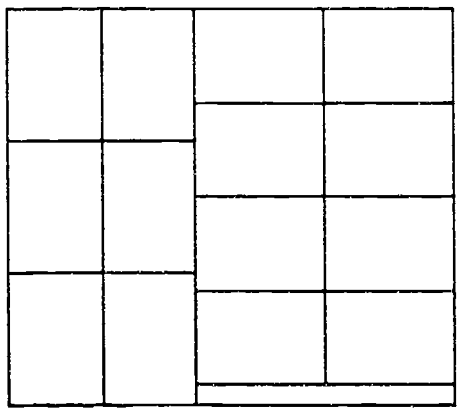
TICKET

Task Descriptor

To determine what is the greatest number of tickets (rectangles) that can be cut from a sheet of paper.

Equipment/Material

One ticket (5 cm by 7 cm) and a blank sheet of paper (24 cm by 21 cm). (Lines have been added to show the solution.)



Student Instructions

Paul got 12 tickets by cutting up his sheet of paper. Julie managed to get 13 tickets from her sheet.

Find the greatest number of tickets that can be made from a sheet of paper. Draw lines on the sheet of paper to show how it would be divided up to make this number of tickets.

Scoring Scheme

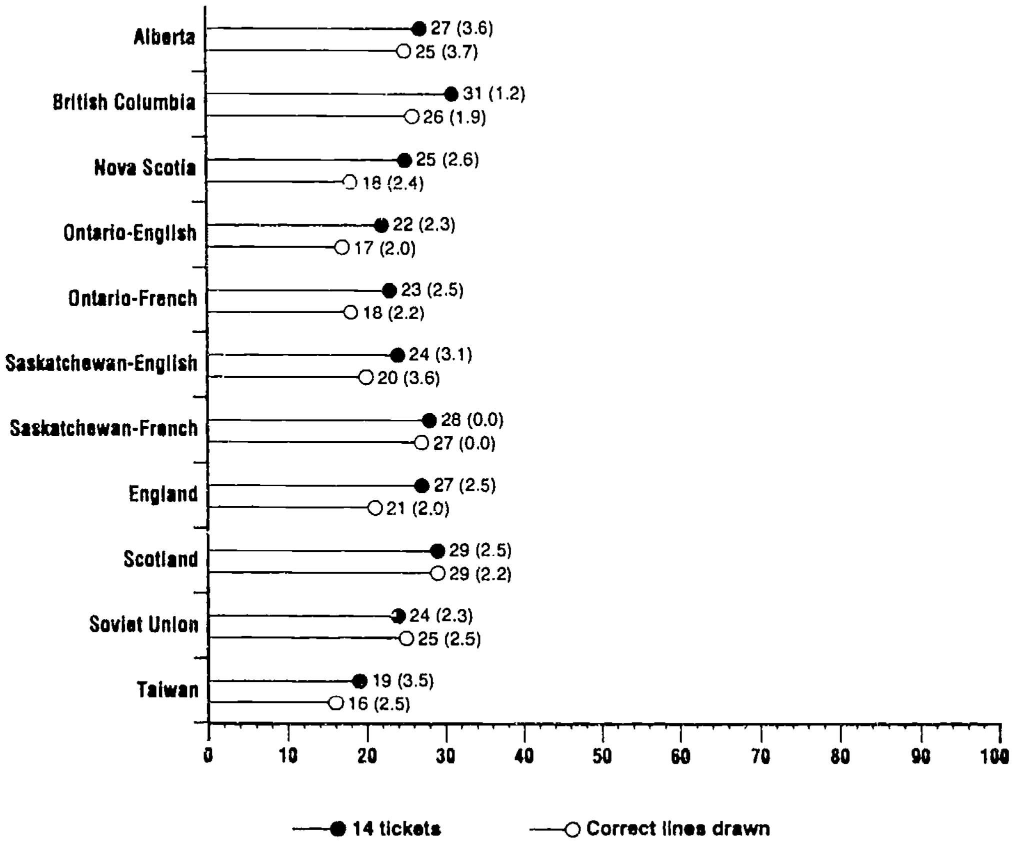
Credit given for answer 14 and for drawing lines as shown above.

Comments

- Scores on this task were quite low -- the highest being just over 30 percent indicating that 14 tickets could be made. Fewer were able to draw the lines showing the correct solution.

- Across all participating countries and provinces, 9 percent of students thought the maximum number of tickets that could be made was 13 and 47 percent thought only 12 could be made. (Both of these numbers were mentioned in the task instructions.)

Percentage of Correct Responses (with Standard Errors)



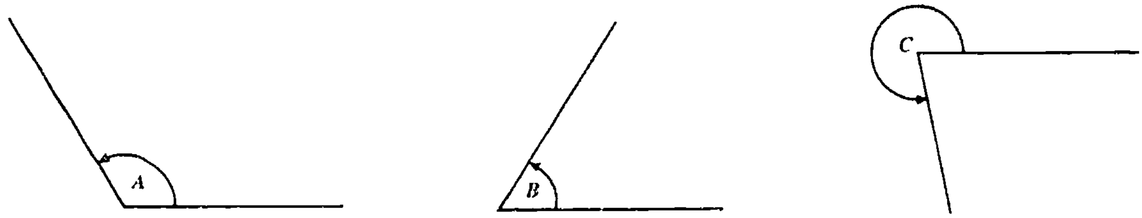
ANGLES

Task Descriptor

To measure three angles using a 180° protractor.

Equipment/Material

A 180° protractor and three angles provided in the student's booklet (120°, 58° and 280° respectively).



Student Instructions

Measure the angles to the nearest degree.

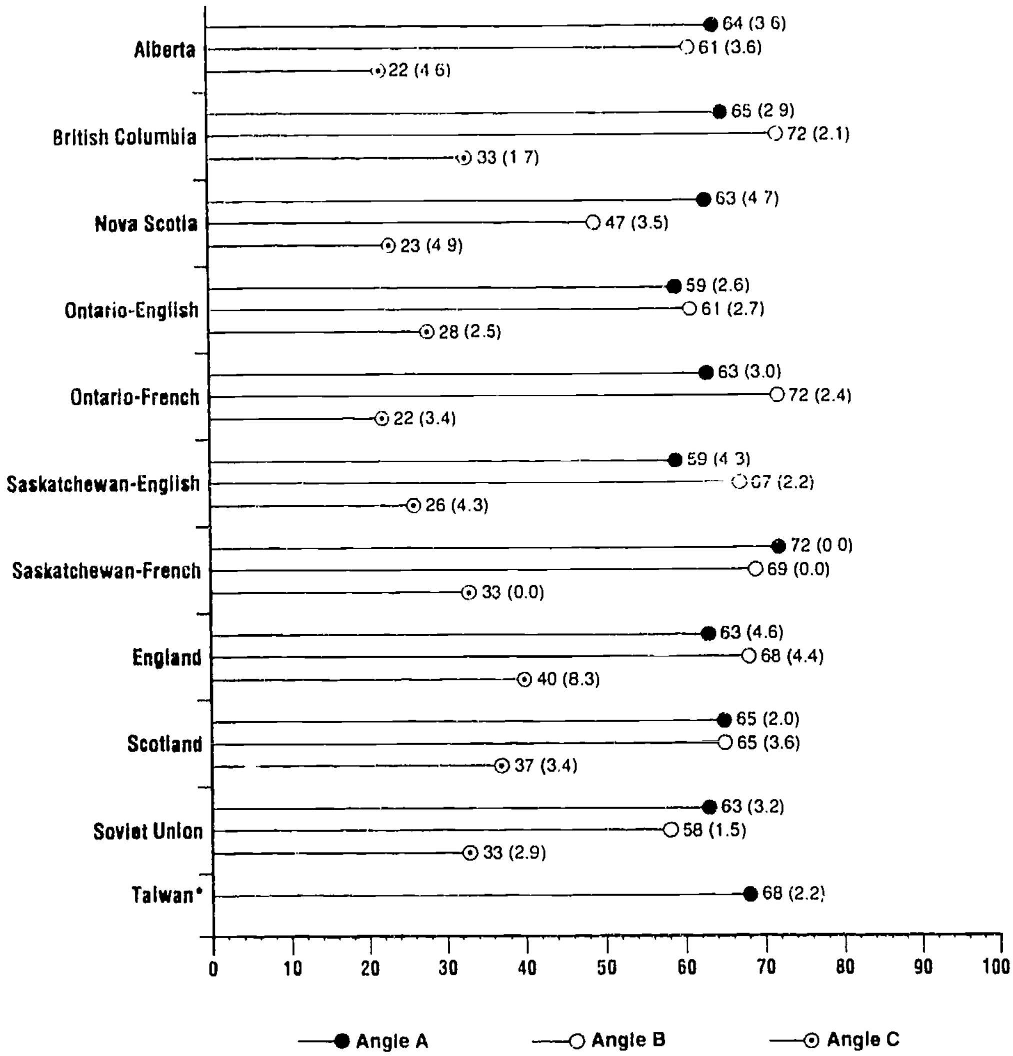
Scoring Scheme

Credit for correct measurement ± 2°.

Comments

- Results were similar on the obtuse and acute angles, ranging from 47 to 72 percent correct. Performance was much lower on the reflex angle, ranging from 22 to 40 percent correct.
- Obtuse Angle A: The most common wrong answer, given by 21 percent of students across participating countries and provinces, was 60° (± 2°), presumably because they measured the "reverse" acute angle.
- Acute Angle B: The most common wrong answer, given by 11 percent of all students, was 122° (± 2°), presumably because they measured the "reverse" obtuse angle.
- Reflex Angle C: The most common wrong answer, given by 28 percent of all students, was 80° (± 2°), presumably because they measured the "reverse" acute angle. Twelve percent of students answered 100° (± 2°), perhaps because they measured the acute angle (80°) and subtracted from 180° instead of 360°.

Percentage of Correct Responses (with Standard Errors)



*No information for Angle B and Angle C

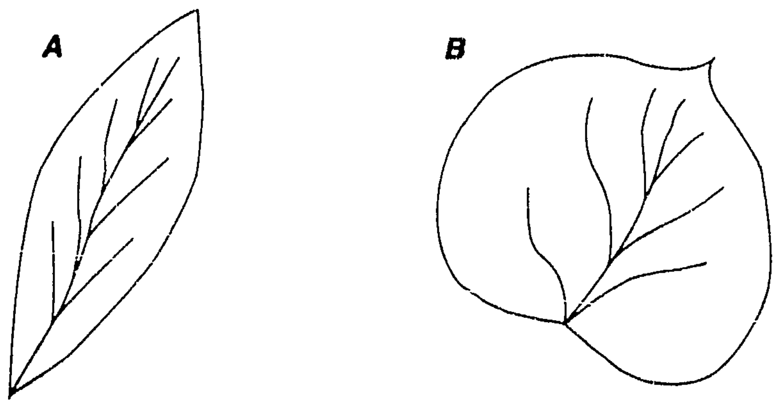
LEAVES

Task Descriptor

To find the area of two irregular shapes using a grid square.

Equipment/Materials

Two drawings of leaves (with areas of $21cm^2$ and $48cm^2$, respectively), tracing paper and a transparent grid marked in square centimetres.



Student Instructions

Find the area of the leaves, using the grid. You may use the tracing paper on top of the grid to mark off the boxes that cover the leaf totally or partially.

Scoring Scheme

Credit given for correct areas $\pm 4cm^2$.

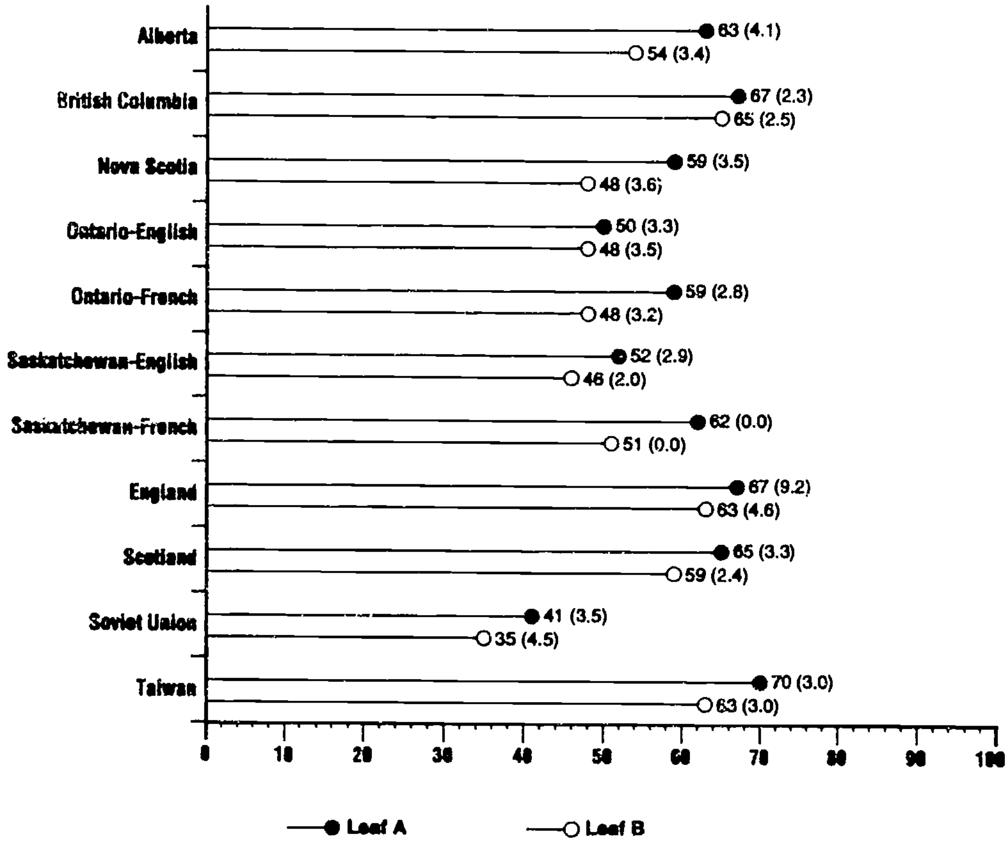
Problems

When master copies of the leaf drawings were photocopied, there were some minor variations in the sizes of the copies produced. Because of this, credit has been given for answers within a wider range than originally intended. Due to a translation problem, the student instructions in the Soviet Union omitted some relevant information. Although this was corrected during the assessment, the results from the former Soviet Union are not comparable to those from other countries and provinces.

Comments

- The students' scores tended to be higher for the smaller leaf.
- Performance differences among countries and provinces were small, in general, with the exception of the Soviet Union (see above), with scores ranging from 46 to 70 percent correct.

Percentage of Correct Responses (with Standard Errors)



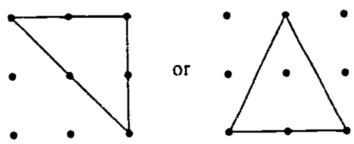
PINBOARD

Task Descriptors

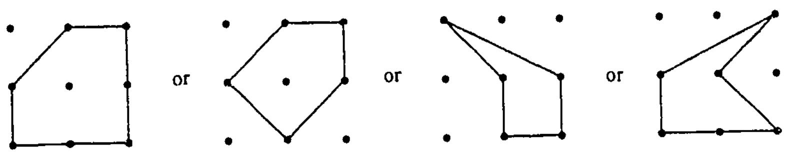
To construct on a nine-point pinboard (or geoboard): (1) a triangle with the largest possible area, (2) a five-sided shape, and (3) three squares of different sizes.

Equipment/Material

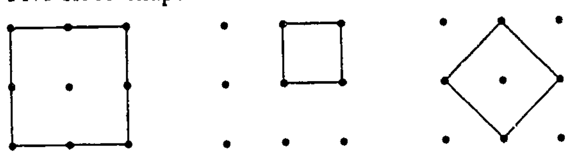
A nine-point pinboard (or geoboard), rubber bands, and three nine-point grids printed in the student's book. (Lines have been added to show examples of correct answers.)



Triangle



Five-sided shape



Large square | Small square | Diamond

Student Instructions

Use the rubber bands to make the required shapes on the pinboard, then draw the shapes on the three grids in your booklet.

Scoring Scheme

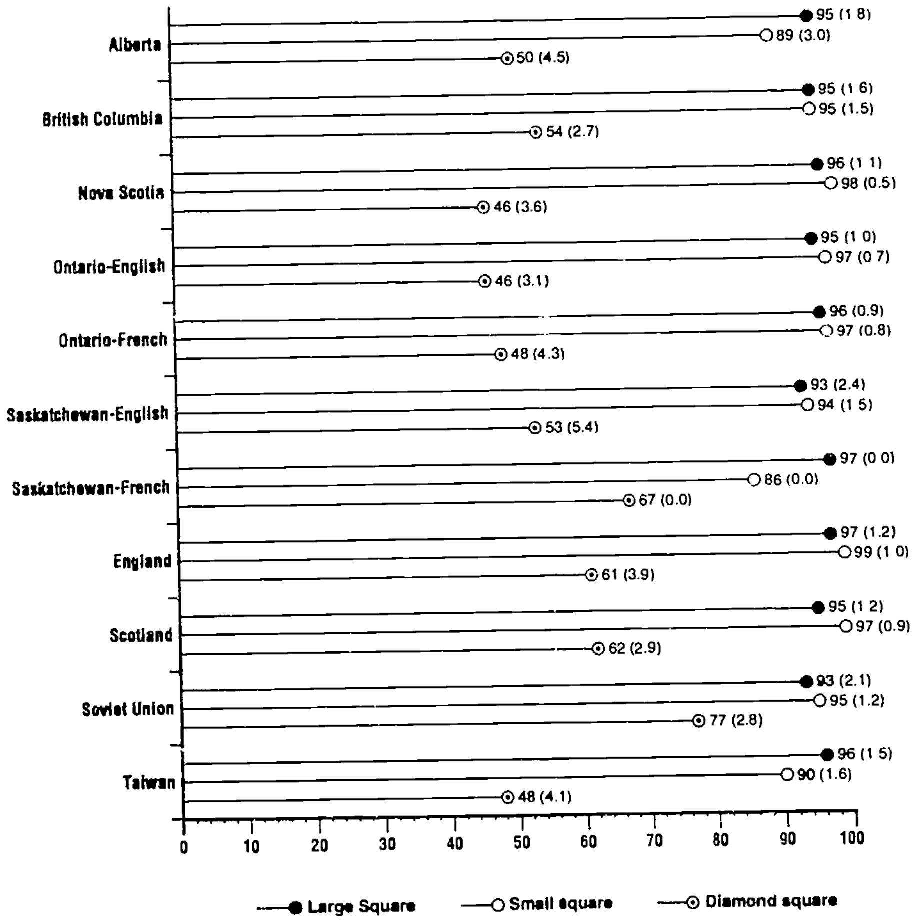
Credit given for: (1) drawing either version of the largest triangle, (2) drawing any five-sided shape, including side-by-side triangles, and (3) drawing a large square, any small square, and a diamond square.

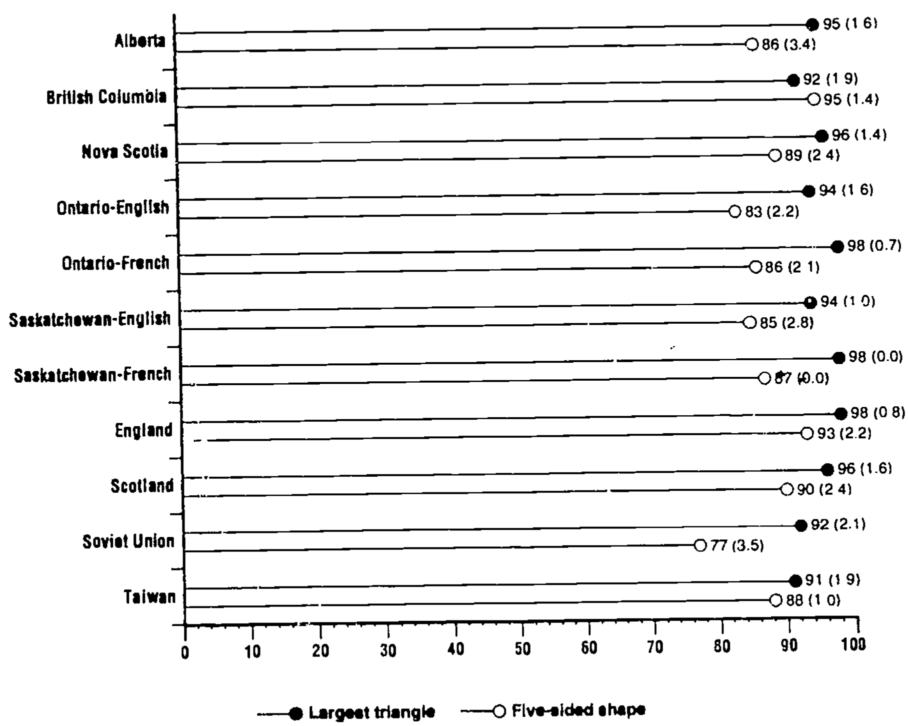
Comments

- Scores were very high for the largest triangle, the large square, and the small square and only slightly lower for the five-sided shape.

- In the identification of three different squares, there was a markedly poorer performance on the diamond square, presumably because of its atypical orientation.

Percentage of Correct Responses (with Standard Errors)





TRIANGLES

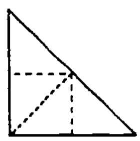
Task Descriptors

To combine four triangular shapes to form: (1) a square, (2) a large triangle, and (3) a six-sided shape.

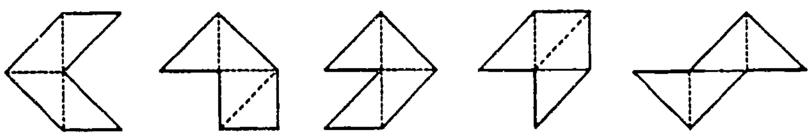
Equipment/Material

Four equilateral triangles. (Dotted lines have been added to show the position of the four triangles.)

Square



Large Triangle



Six-sided Shape

Student Instructions

Use the triangles to make the required shapes without leaving any spaces between the triangles. Trace around the outside of the shapes in your booklet.

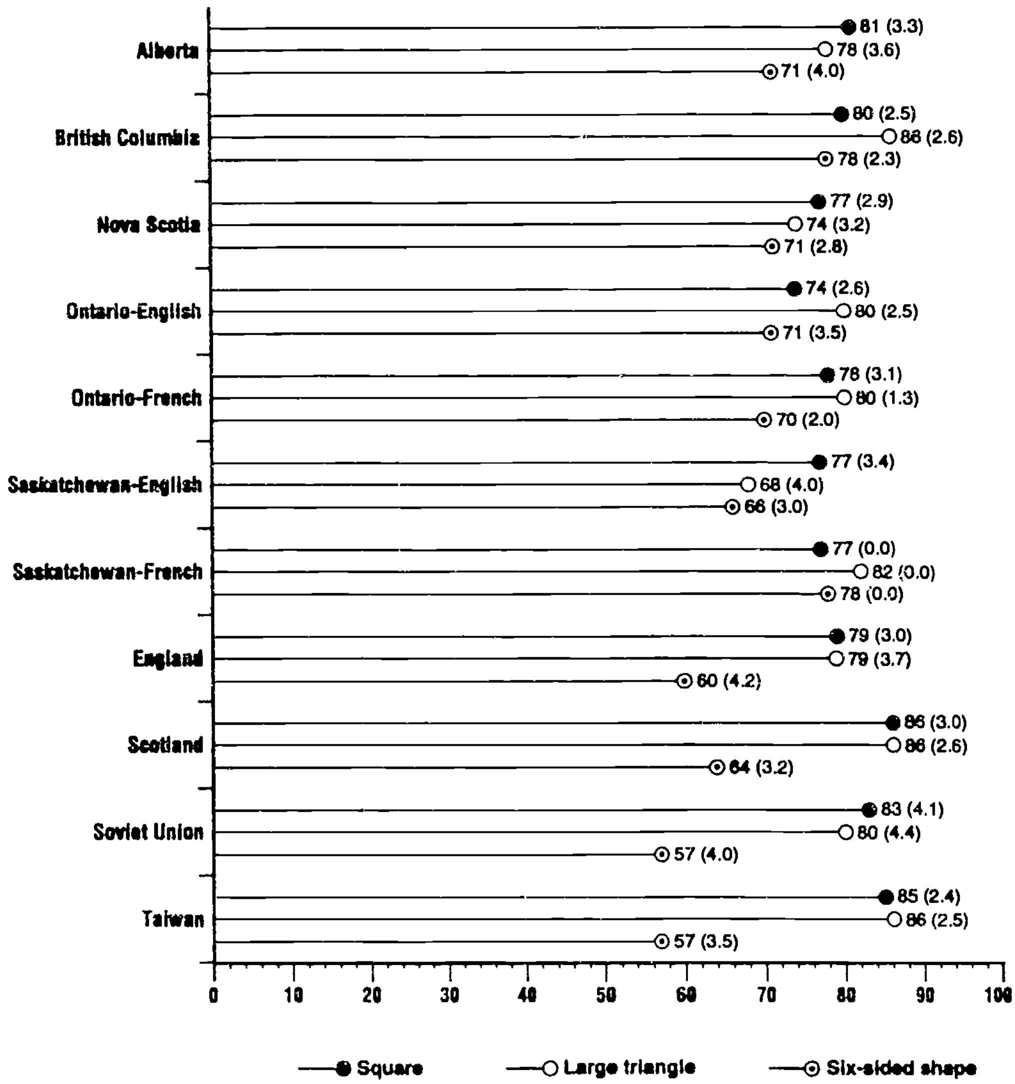
Scoring Scheme

Credit given for: (1) a square, (2) a large triangle, and (3) any six-sided shape.

Comments

- Scores were generally high for the square and triangle but lower for the six-sided shape, particularly in countries other than Canada.
- The omission rate on all of these tasks was relatively high, 8 percent across all participating countries and provinces.

Percentage of Correct Responses (with Standard Errors)



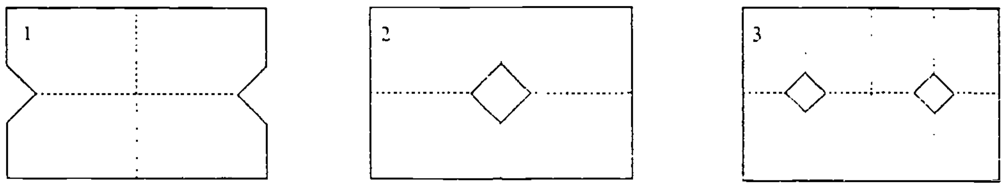
SCISSORS

Task Descriptors

To create three defined shapes from sheets of paper in one cut.

Equipment/Material

Drawings of three shapes (see below), a pair of scissors, and sheets of paper. (Dotted lines have been added to show the required folds.)



Student Instructions

Fold a sheet of paper and make one straight cut to produce each of the required shapes.

Scoring Scheme

Credit given for the correct cut out(s) and evidence of the correct folding (and no evidence of more than one cut).

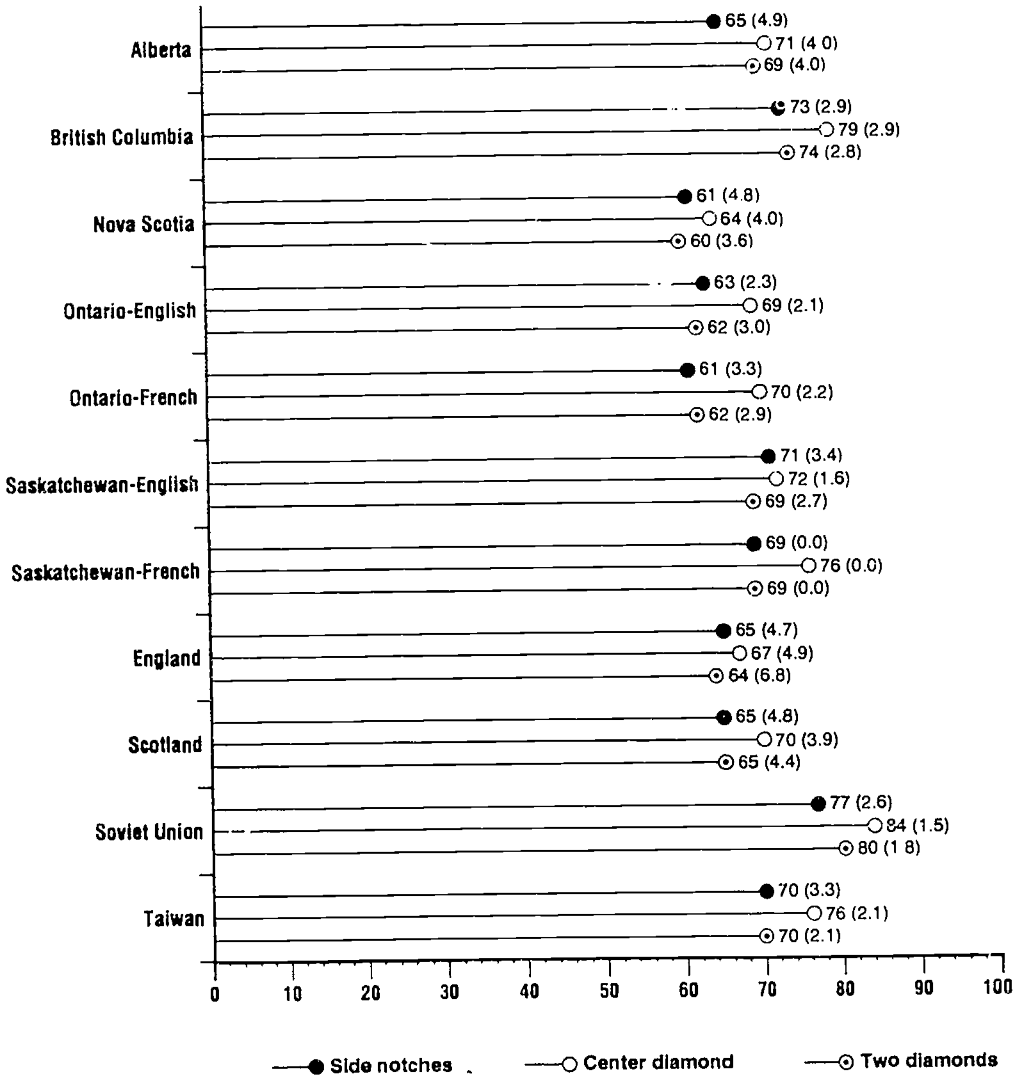
Comments

- Scores were generally quite high on these tasks, ranging from 60 to 84 percent correct, and differences among countries and provinces were relatively small.

- The consistency of the results on the three tasks could indicate that students who grasped the correct strategy in the first task tended to complete the remaining tasks successfully.

- Fifteen percent of students across all participating countries and provinces made more than one attempt on these tasks, most of them two attempts. Omission rates were about 6 percent across all students.

Percentage of Correct Responses (with Standard Errors)



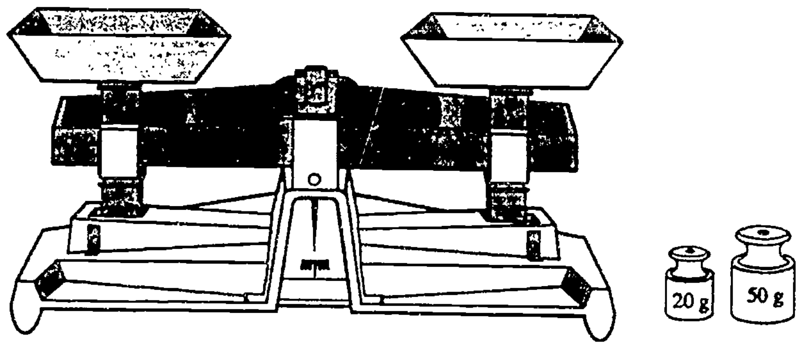
CLAY

Task Descriptor

To make a 15g lump of modelling clay from a larger lump of clay, using a two-pan balance and two masses.

Equipment/Material

A large lump of modelling clay, a two-pan balance, and two masses, 20g and 50g, respectively.



Student Instructions

Make a 15g lump of clay using the materials provided and explain how you did it. (Simple instructions on how to use the balance were provided.)

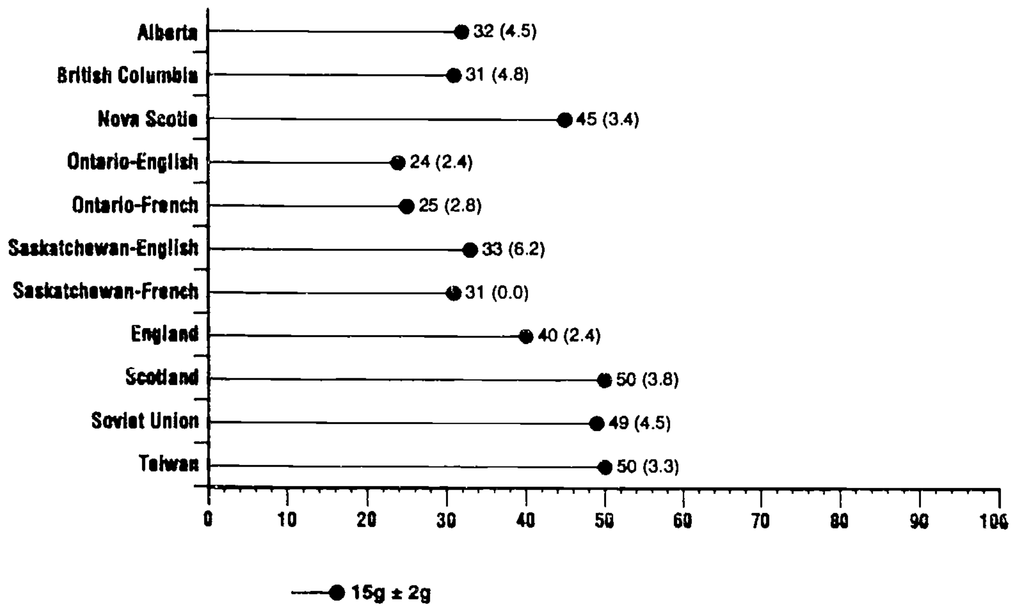
Scoring Scheme

Credit given for a clay lump of 15g ± 2g and for describing a feasible method of obtaining it.

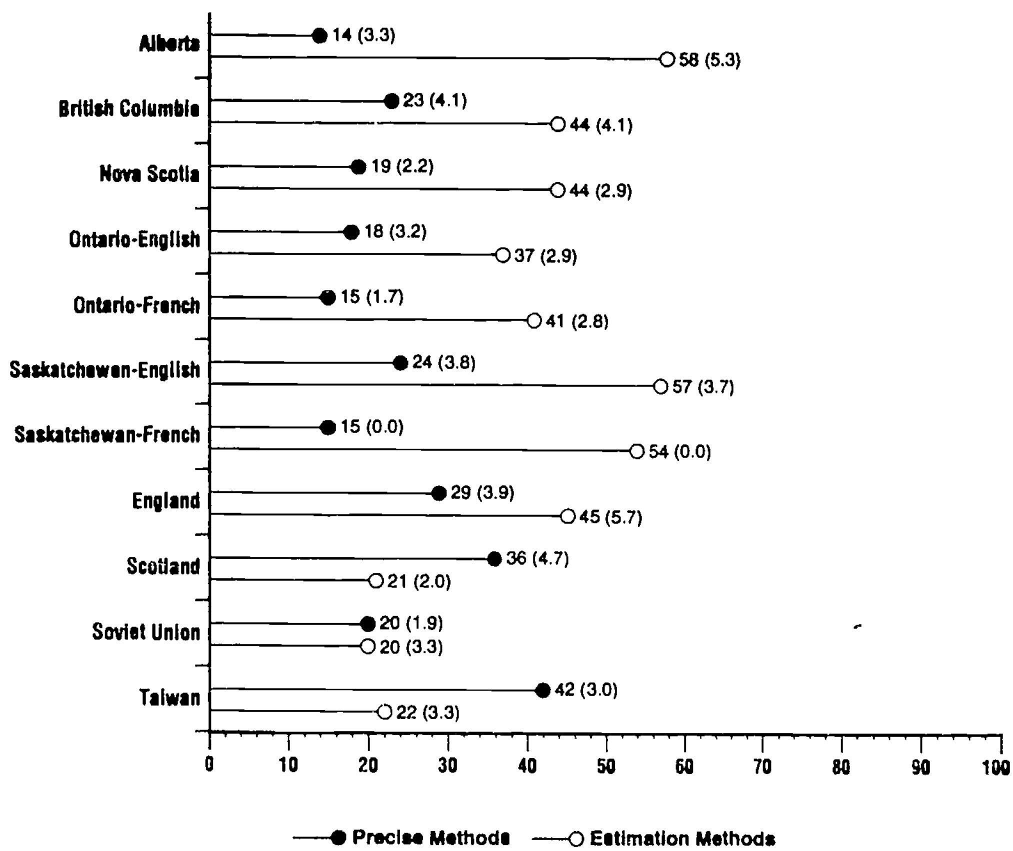
Comments

- All but 5 percent of the students across participating countries and provinces produced a lump of clay, but 11 percent of them did not provide a description of how they obtained it.
- Scores for obtaining clay lumps 15g ± 2g were generally lower in the Canadian provinces than elsewhere, the exception being Nova Scotia.
- There was a strong tendency towards the use of estimation rather than mathematically precise methods in England and the Canadian provinces.

Percentage of Correct Responses (with Standard Errors)



Percentage of Students Using Different Methods (with Standard Errors)



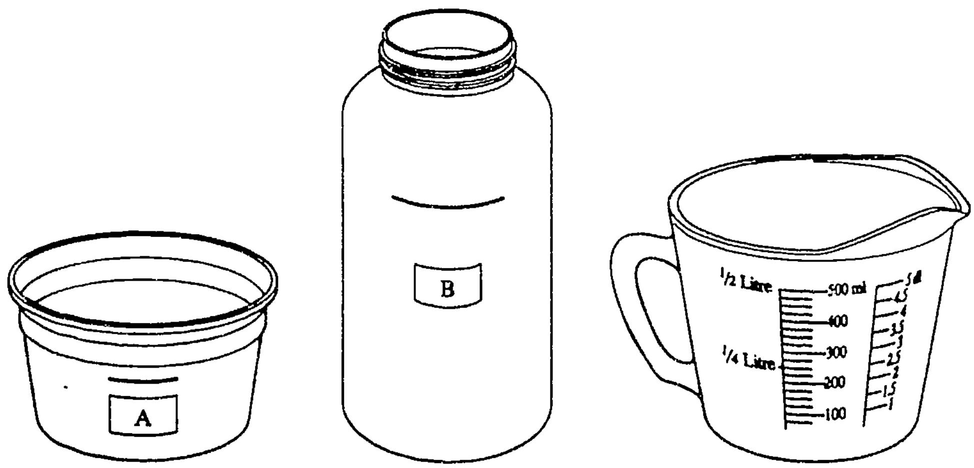
WATER

Task Descriptor

To measure the capacity of two plastic containers using a measuring cup.

Equipment/Material

A large container filled with water, two smaller containers labelled A and B and a measuring cup (500ml graduated in 25ml units). Containers A and B were marked with a black line at 375ml and 725ml, respectively.



Student Instructions

Fill containers A and B up to the black lines from the large container.
Measure the amount of water in each container in millilitres, using the measuring cup.

Scoring Scheme

Credit given for correct volume ± 25ml.

Problems

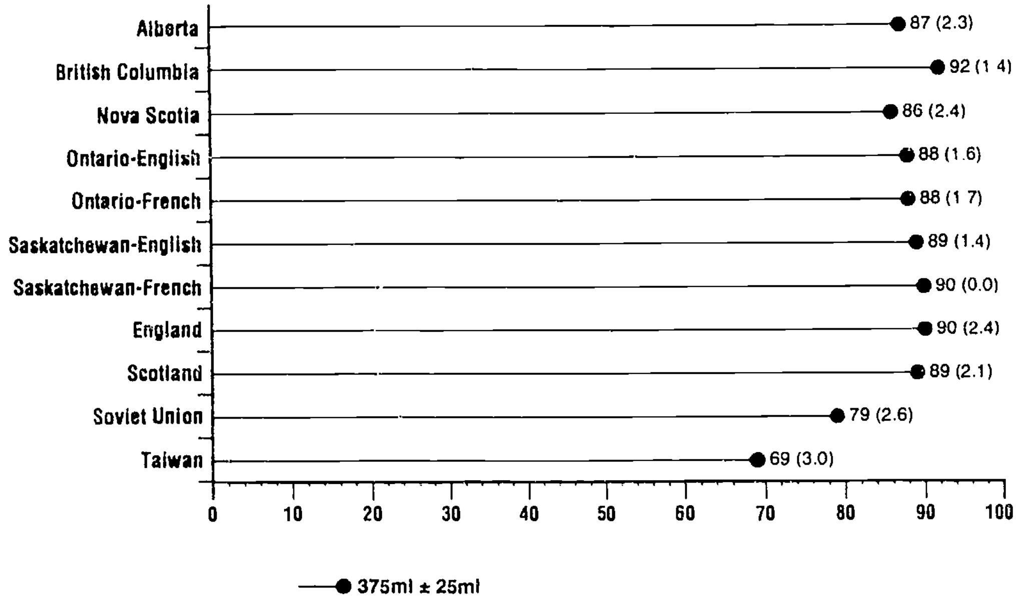
Checking the calibration of the black lines on the containers and measuring cups after the assessment revealed significant variations. For container A, these differences could be accommodated by giving credit for answers within ± 25ml. Because the problem was magnified in the measurement of container B, it proved impossible to make suitable adjustments and these results were discarded.

Comments

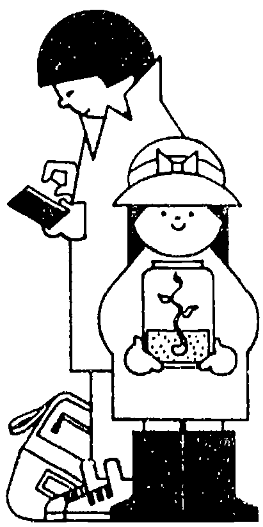
- Scores were generally high at about 90 percent, the exceptions being Taiwan and the Soviet Union, where scores were somewhat lower.

- Non-response rates were very low (1 percent across all countries and provinces) but spillage was high!

Percentage of Correct Responses (with Standard Errors)



Science Tasks



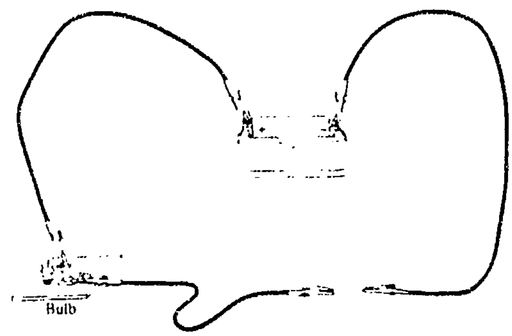
LIGHT-UP

Task Descriptor

To categorize objects according to their electrical conductivity by completing an electrical circuit; to explain why some objects enable a bulb to light; to predict whether an object in a sealed container would enable the bulb to light and to explain why (or why not).

Equipment/Material

An electrical circuit with a bulb and a gap with two contacts which could be bridged. Five objects as follows: wood strip, plastic strip, nail, foil strip and cardboard strip. Also an object (piece of copper wire) in a sealed, clear plastic box.



Student Instructions

Complete the circuit using the five objects in turn. List those objects that enable the bulb to light and explain why. Say whether you think object X would enable the bulb to light and explain why.

Scoring Scheme

Credit was given for identifying the nail and foil strip as conductors and for giving an explanation mentioning one of the following or its equivalent: objects conduct electricity, allow electricity/charge to pass, complete the circuit, are metal. Also credit was given for saying object X would enable the bulb to light and for giving an explanation as above.

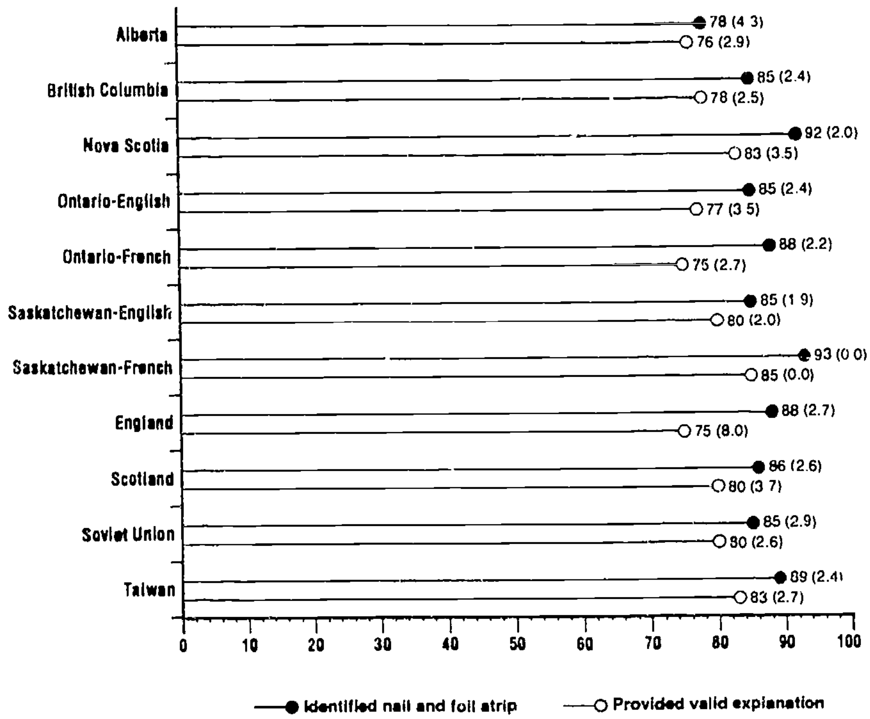
Problems

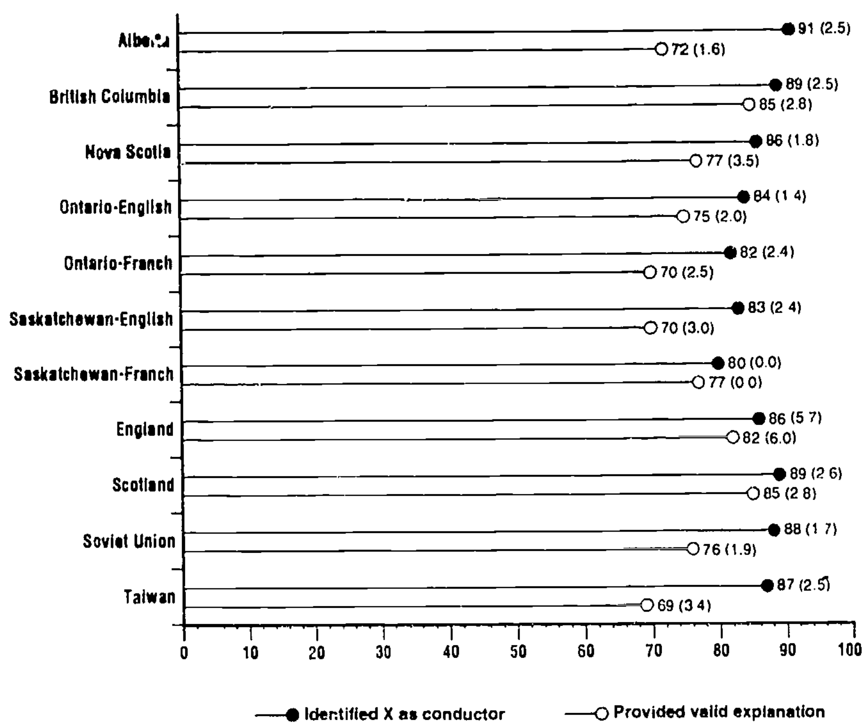
There was a problem in some Canadian provinces where the word "enable" in the instructions was read as "unable." Students who listed the nonconductors and provided an appropriate explanation were counted as giving the correct answers.

Comments

- Most students, 78 to 93 percent, categorized the objects correctly, but somewhat fewer were able to give a valid explanation for what they had done.
- In four of the countries and provinces, more students recognized the conductivity of object X than had categorized the original objects correctly and in two of these countries and provinces (and three in total), more students gave a valid explanation for their decision.

Percentage of Correct Responses (with Standard Errors)





CIRCUIT

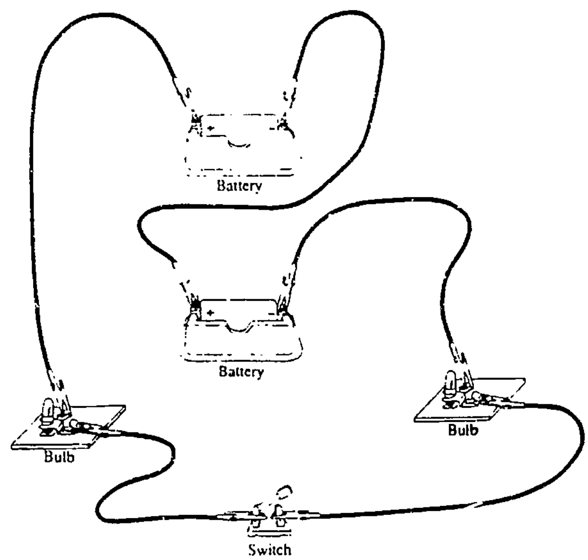
Task Descriptor

To construct an electrical circuit as represented in a drawing by selecting appropriate components and connecting them correctly.

Equipment/Material

Drawing of the circuit and a set of components as listed below. (Number of components required to construct the circuit are shown in parentheses.)

3 batteries (2)
2 battery holders (2)
3 bulbs (2)
2 bulb holders (2)
1 switch (1)
6 wires with clips (5)



Student Instructions

Use the objects on the card to make up the circuit shown in the drawing. You may not have to use all of the equipment. When your circuit matches the diagram, close the switch and see what happens. Raise your hand and ask the administrator to check your work.

Scoring Scheme

Credit was given for the correct positioning of batteries and bulbs, and for using five wires to form a closed loop, thus enabling the bulbs to light.

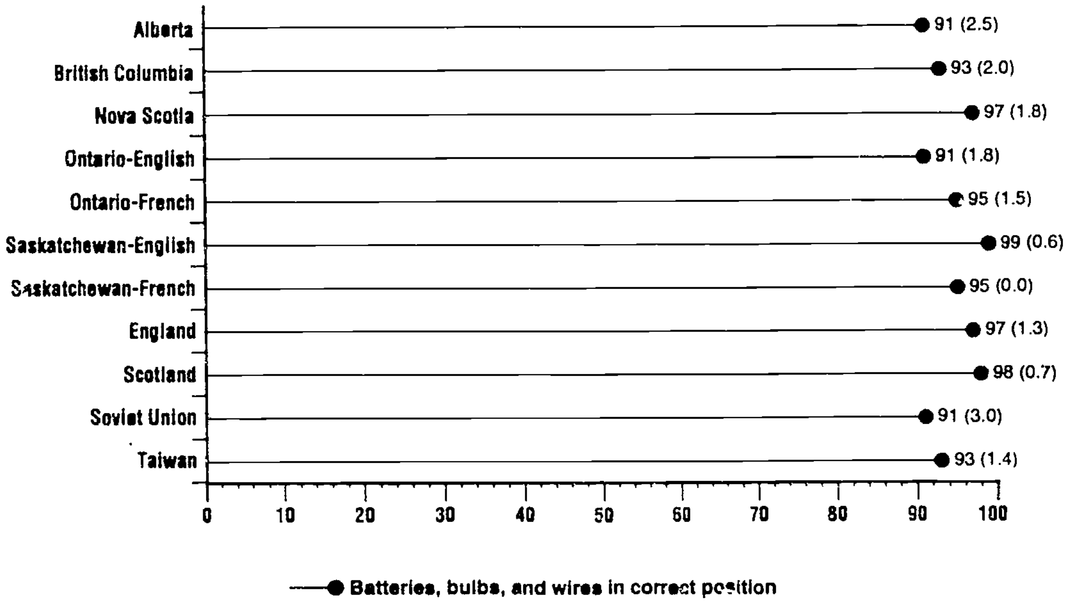
Problems

A loose connection in a bulb holder in one of the kits used in Ontario prevented the two bulbs from lighting, but students were credited for constructing the circuit correctly.

Comments

- Almost all students across participating countries and provinces completed this task successfully.

Percentage of Correct Responses (with Standard Errors)



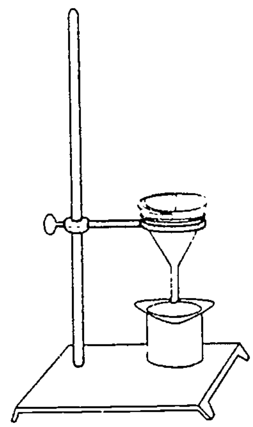
FILTER

Task Descriptor

To set up apparatus for filtering, as shown in a drawing, and to filter some muddy water.

Equipment/Material

A ring stand, a funnel, a beaker, and a folded filter paper. Also, a bottle of muddy water.



Student Instructions

Set up the apparatus as shown in the drawing above, put the folded filter paper into the funnel, and pour a small amount of muddy water into the funnel. Raise your hand when you have gotten some clear water and ask the administrator to check your work.

Scoring Scheme

Credit was given for the apparatus being assembled correctly, the filter paper being inserted correctly in the funnel, and for any clean water obtained.

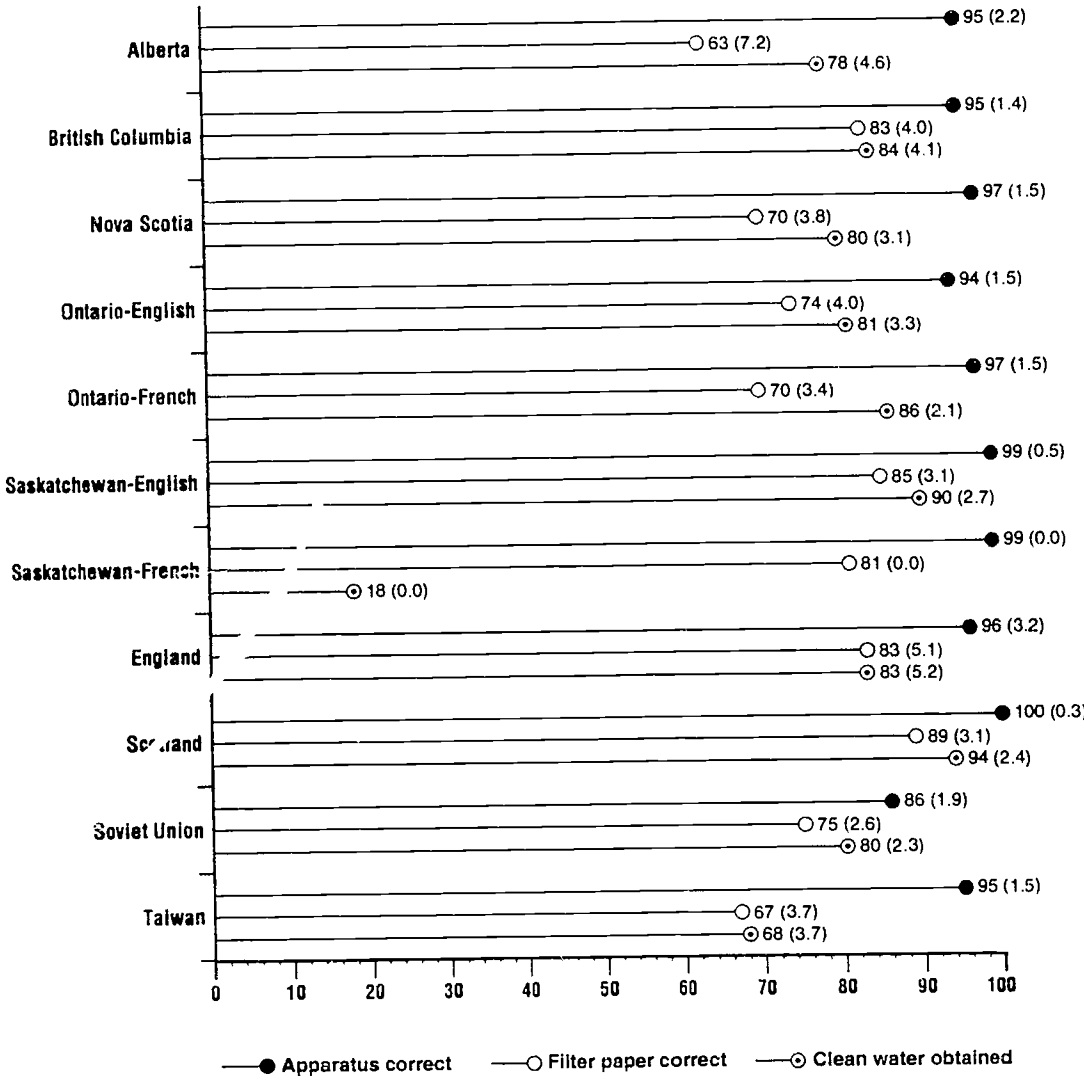
Problems

In the pilot-testing, filter papers were supplied unfolded and this caused widespread problems, but in the final assessment they were pre-folded.

Comments

- There was a high success rate, 86 to 100 percent correct, in assembling the apparatus; but more difficulty was experienced with correctly inserting the filter paper, where success ranged from 63 to 89 percent correct.
- Despite problems with the filter paper, many students were still able to obtain some clean water.

Percentage of Correct Responses (with Standard Errors)



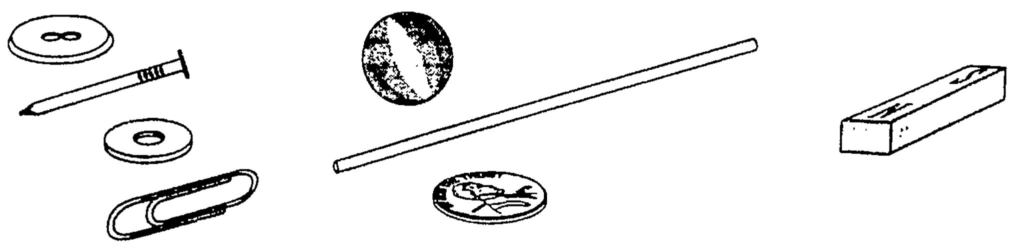
MAGNET

Task Descriptor

To use a magnet to identify magnetic and non-magnetic items and then to explain the difference between them.

Equipment/Material

A magnet and the following seven objects: plastic button, iron or steel washer, steel paper clip, iron nail, glass marble, plastic rod and copper coin.



Student Instructions

Test the objects with the magnet and divide them into two groups. List the objects in the two groups and explain what makes the objects in the two groups different.

Scoring Scheme

Credit was given for grouping the objects correctly. Four categories of explanations were recorded: namely, that one group was made of iron or steel, that one group was attracted by the magnet, that one group was made of iron and steel and was attracted by the magnet, and any other explanation.

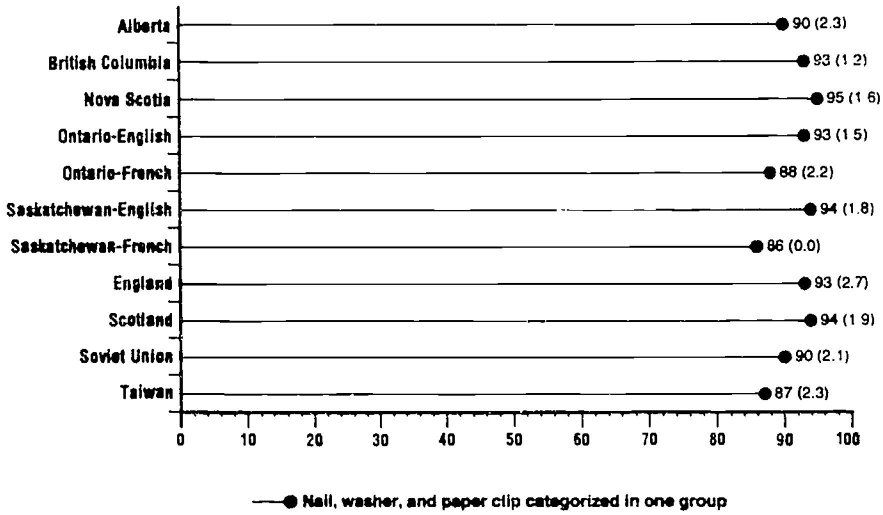
Comments

- Generally students performed the categorization task well, scores ranging from 86 to 95 percent correct; but 10 percent of the students across all countries and provinces gave irrelevant explanations.

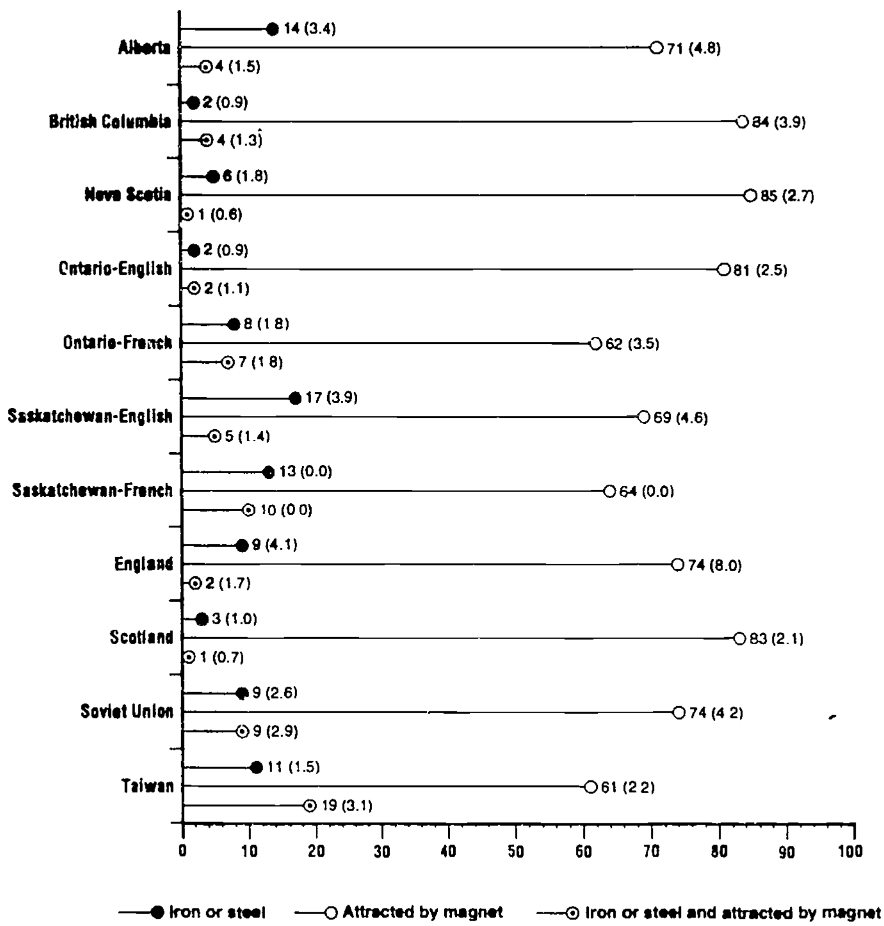
- Omission rates were generally low, but there was a 6 percent omission rate in England.

- The most frequent explanation for students' categorization was that one group of objects was attracted by the magnet; 79 percent of the students across participating countries and provinces gave this response. Fewer, between 4 and 30 percent, mentioned iron or steel, and this varied considerably among countries and provinces.

Percentage of Correct Responses (with Standard Errors)



Percentage of Students Giving Particular Explanations (with Standard Errors)



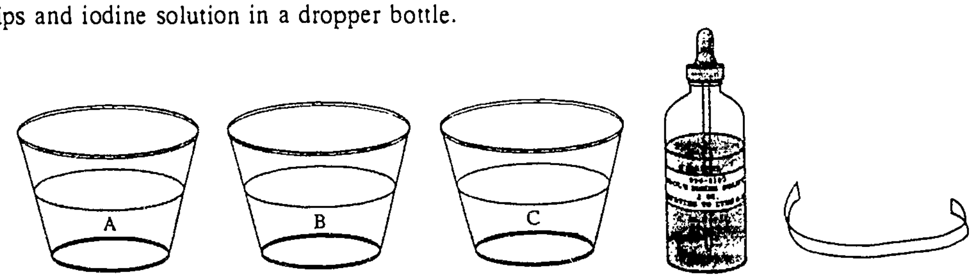
INDICATORS

Task Descriptor

To determine whether three solutions contain glucose, starch, or glucose and starch using indicators for glucose (test strip) and starch (iodine solution).

Equipment/Material

Three dishes labelled A, B and C containing the standardized, unknown solutions. Glucose test strips and iodine solution in a dropper bottle.



Student Instructions

The glucose test strip will turn from yellow to green on contact with a solution containing glucose and the iodine solution will turn blue-black when starch is present. The dishes A, B and C contain three different solutions which you are to test for glucose and starch using the indicators. Take the dish filled with solution A and dip the glucose test strip into it. Let the test strip dry. Add a drop of iodine solution to dish A. Observe all the results, report what solution A contains and repeat for solutions B and C.

Scoring Scheme

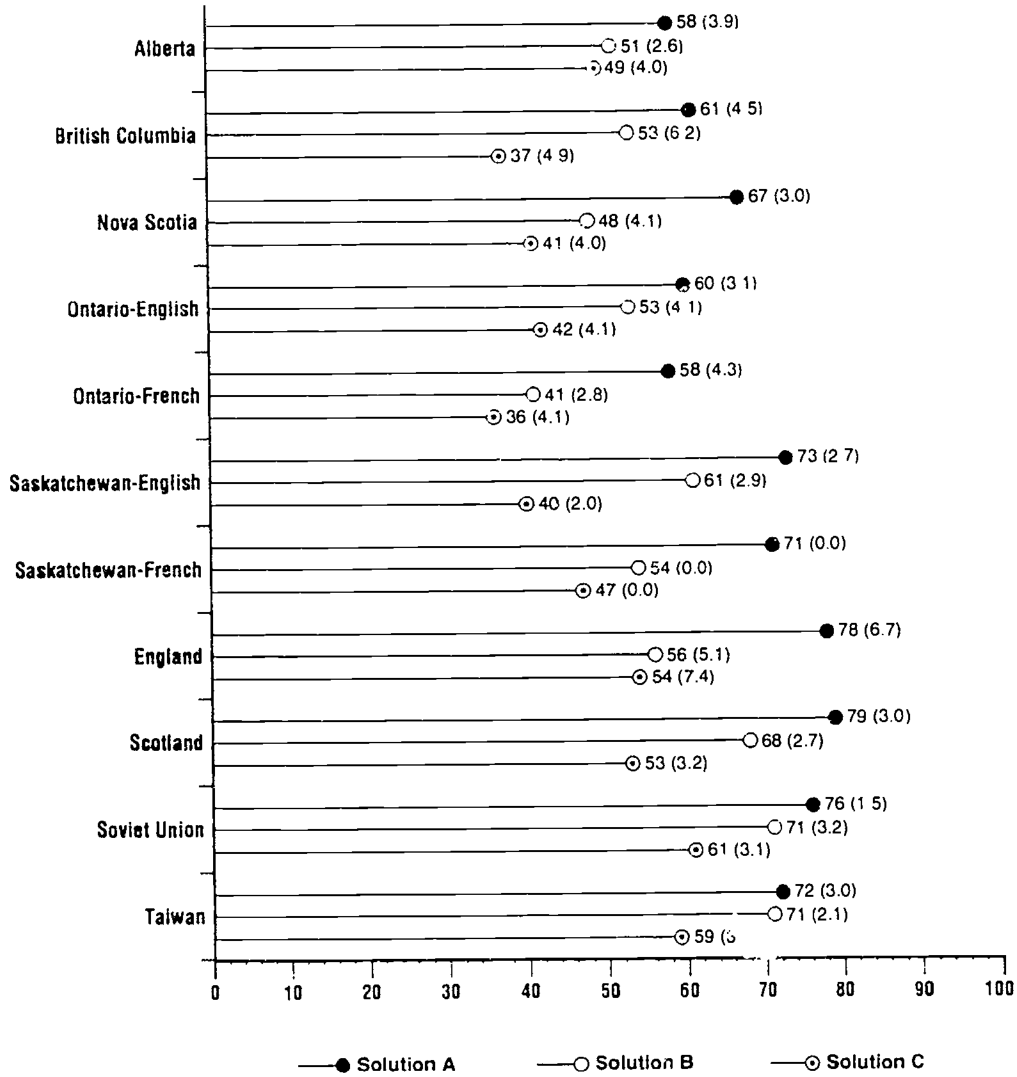
Credit was given for identifying glucose only in solution A, starch only in solution B, and glucose and starch in solution C.

Comments

- The differences in performance among countries and provinces were substantial in all three tasks. For each task, the difference in the scores of the highest and lowest performing populations was at least 20 points.

- Success rates in identifying the solution containing only glucose were highest, averaging 68 percent correct across participating countries and provinces. Those for the starch-only solution and the mixture of both averaged 53 and 47 percents correct, respectively.

Percentage of Correct Responses (with Standard Errors)



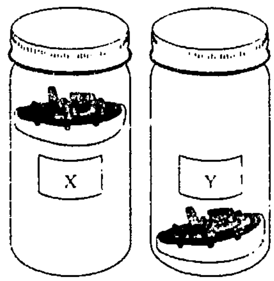
FLOAT

Task Descriptor

To select correct observations about flotation from two sets of objects.

Equipment/Material

Two small glass jars labelled X and Y containing clear liquids and identical plastic toys, one floating (in jar X) and one submerged (in jar Y).



Student Instructions

Look carefully at the two jars -- you may pick them up. Five other students looked at these jars and made the following statements. Which statements are observations, that is, they describe what the student actually saw?

A. I see a toy floating in jar X.
B. I see a toy floating in jar Y.
C. I see a toy in jar X that is made of a different plastic than the toy in jar Y.
D. I see jars containing colourless liquids and coloured toys.
E. I see a toy in jar Y that is heavier than the toy in jar X.

Scoring Scheme

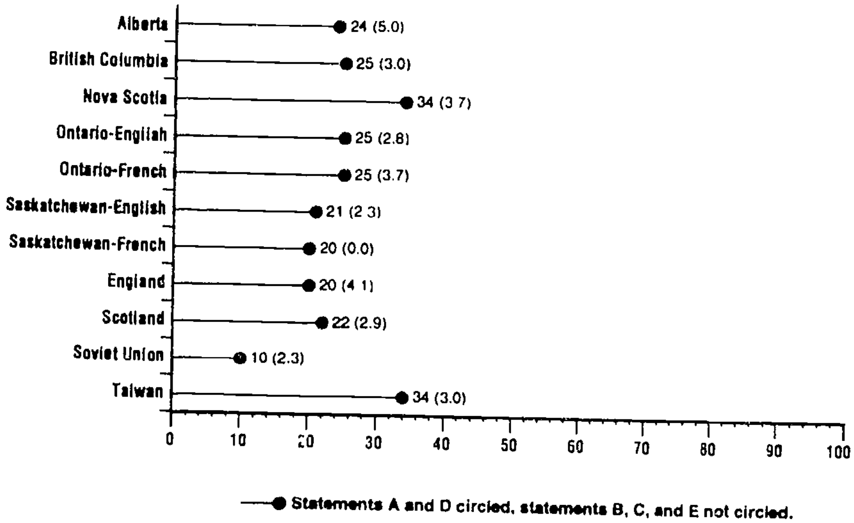
Credit was given for circling correct statements A and D and not circling incorrect statements B, C and E.

Comments

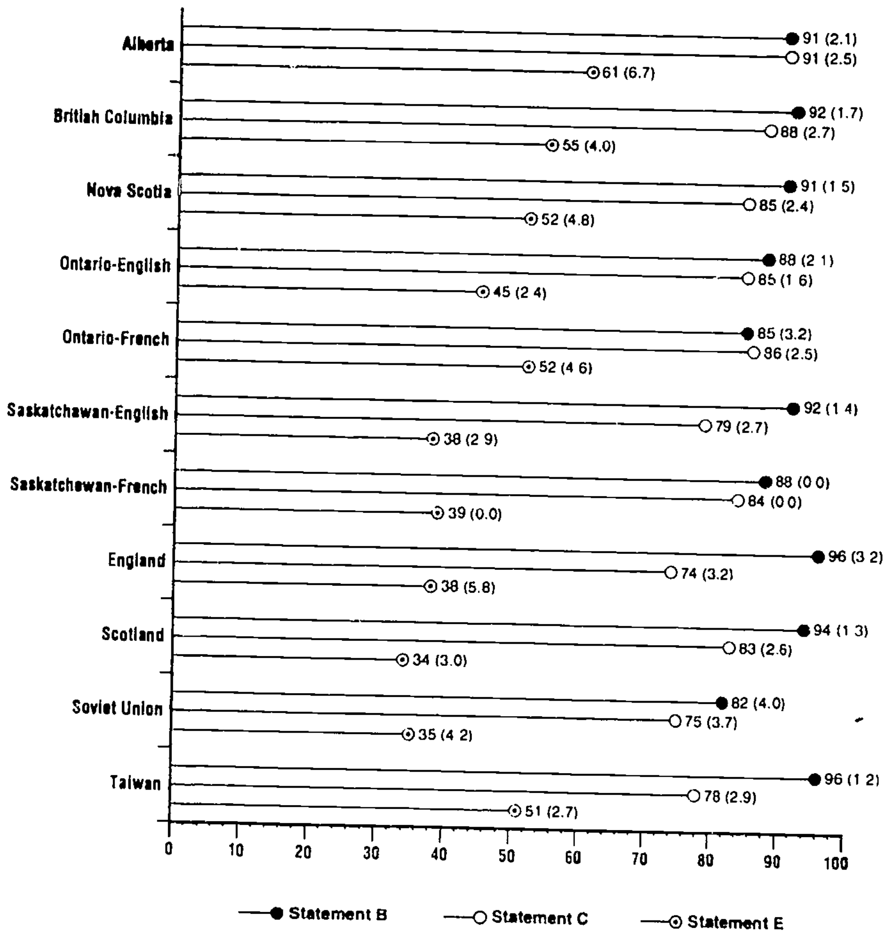
- The percentages of students who circled both correct statements *and* none of the incorrect ones were low, ranging from 10 to 34 percent.

- Most students recognized statements A and D as observations. Almost all students recognized that statement B, the opposite of statement A, is not an observation. Most students recognized statement C, that the two toys were made of different plastic, is an incorrect statement, probably because the toys looked so similar. However, statement E, that the mass of the toys were different, proved attractive to many students and they circled it, even though they had no way of knowing the mass of the two toys.

Percentage of Correct Responses (with Standard Errors)



Percentage of Students Recognizing Incorrect Statements (with Standard Errors)



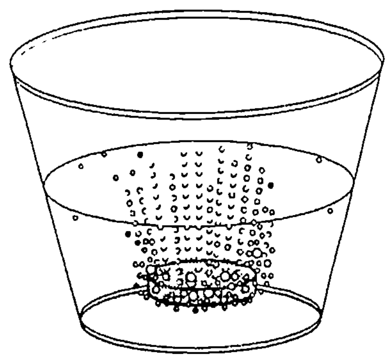
TABLET IN WATER

Task Descriptor

To observe and record all the changes which take place when a tablet dissolves in water.

Equipment/Material

Water supply, plastic cup, and fruit-flavoured, coloured fizzy tablets.



Student Instructions

Observe what happens when the tablet is in the water. Write as many different things as you notice.

Scoring Scheme

Credit given for all appropriate visual, auditory, and olfactory changes recorded.

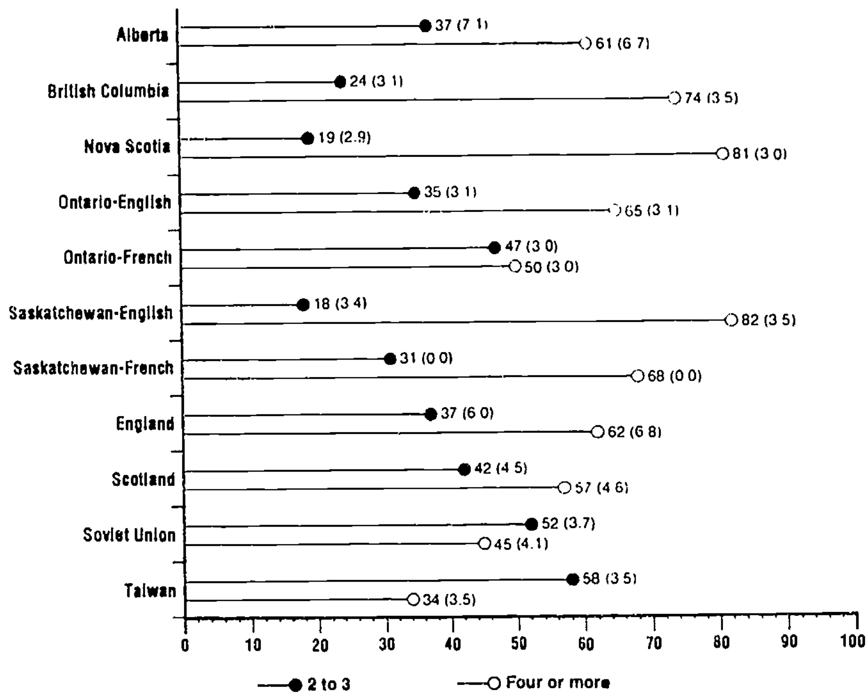
Comments

- The changes that were recorded by most students were in the size of the tablet, the colour of the water, and the bubbling of gas. These are all visual changes and it may be that the use of the word "observe" in the students' instructions biassed their responses towards such changes. However, there were substantial differences in the reporting of different visual changes and among different countries and provinces.

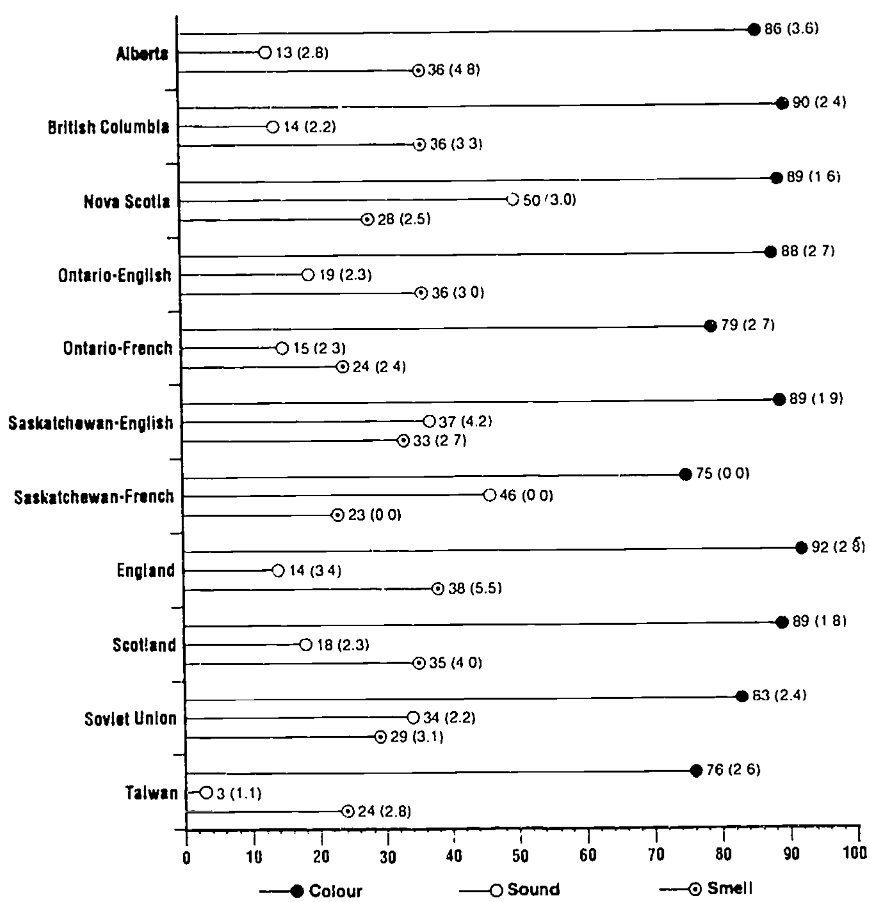
- A notable feature was a wide range in the reporting of the fizzing sound as the tablet dissolved, from 3 percent in Taiwan to 50 percent in Nova Scotia.

- At least one-half of the students in participating countries and provinces mentioned four or more observations, except in the Soviet Union and Taiwan, where the percentages were 45 percent and 34 percent, respectively.

Percentage of Students Mentioning Correct Observations (with Standard Errors)



Percentage of Students Mentioning Specific Observations (with Standard Errors)



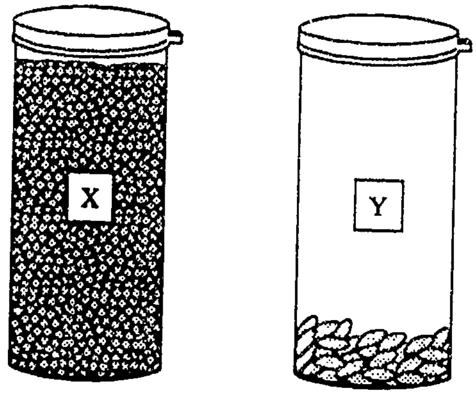
SEEDS

Task Descriptor

To categorize two different types of seeds according to their size, shape and colour.

Equipment/Material

Three groups of seeds labelled 1, 2, and 3 and two containers labelled X and Y with the "unknown" seeds.



Student Instructions

Your task is to decide in which group seeds X and Y belong and to state your reasons. Look carefully at the three groups of seeds 1, 2, and 3 -- you may pick up the containers.

Scoring Scheme

Credit was given for relating seeds X to group 3 and mentioning colour and size. Also for relating seeds Y to group 2 and mentioning shape.

Problems

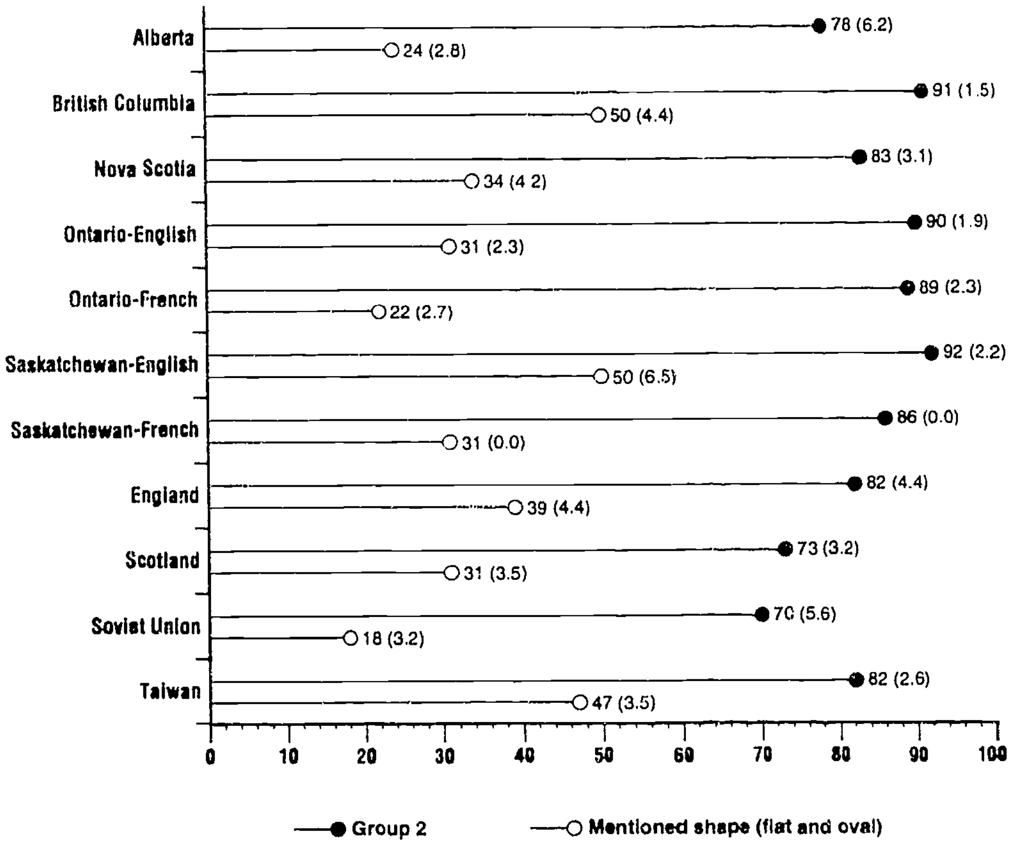
Seeds could not be included in the kits and sometimes it proved impossible to obtain comparable seeds in all of the countries involved. Because the colour of the sesame seeds varied (sometimes white, sometimes yellow), comparable scores could not be obtained for the first part of the task, categorizing seed X.

Comments

- In general, high proportions of the students were able to assign seeds Y to the correct group, ranging from 70 percent in the Soviet Union to 92 percent in English-speaking Saskatchewan.

- Many fewer students provided the correct reason for their categorization. The percentages doing so ranged from 18 to 50 percent.

Percentage of Correct Responses for Container Y (with Standard Errors)



Appendix: Problem Solving In Mathematics

In order to provide a more detailed picture of the rich information produced from the performance assessment, this appendix describes the outcomes produced by Scottish students on one of the problem solving tasks in the mathematics circuit. Results are based on unweighted analyses of student responses.

The station in the mathematics circuit named *Clay* required students to produce a 15g lump of plasticine (modelling clay). To achieve this they were provided with a large lump of plasticine, a two-pan balance with a centering needle (but no weight scale), and two masses, 50g and 20g. Almost all Scottish students produced a lump of plasticine intended to be 15g and in most cases an explanation of how it was produced. Fifty percent of students produced a plasticine lump within ± 2g of the required mass, the agreed tolerance for a correct response. Many of the explanations provided were interesting, but some could not have produced the 15g!

What strategies were adopted by students?

Two types of strategy were used to produce lumps of plasticine of mass 15g -- estimation and precision approaches.

Estimation Approaches. Some 25 percent of the Scottish students used an estimation approach involving only the 20g mass and the balance to produce a plasticine lump less than 20g. Some of these students said they removed 5g once the plasticine lump balanced the 20g mass, others said they removed one-quarter of it or left three-quarters of it. Clearly, the adoption of an estimation procedure in these cases was not due to any lack of ability to make the necessary calculations for a precision approach.

> "I made my 15g lump of plasticine by weighing out 20g and taking 5g away."

However, one student did not get even this correct.

> "Found out 20g and then took a 3rd of the plasticine away."

Others may not have been able to calculate correctly:

> "I adjusted the plasticine until their was 15g."

> "I kept taking bits from the poke and tryed each time of the scales."

Other students gave various explanations which provided only spurious accuracy. For instance, they mentioned deviations in the position of the scale pans and in the centering needle.

> "I placed a 20g mass in one pan and measured the plasticine until it was 20g and then bits off until it was 3/5 of the way to the centre line."

Others gave more exotic but even less meaningful explanations:

> "First I weighed a lump of plasticine against a 50g mass to find out if it was more or less than 5g. It was less, so I weighed it against a 20g mass and just kept pulling bits off and putting bits on the plasticine until the 20g was just a bit heavier than the lump of plasticine."
>
> First I found a lump that weighed 20g, gradually taking small pieces of. When I thought I had 15g I put the 50g in one and 20g and plasticine into another, putting extra pieces into it soon reached 50g with another 20g added."

One or two students seemed to carry estimation to extreme:

> "You get some plasticine and roll it in your hands and make it round." (This student produced a 20g lump of plasticine).

One student started off promisingly in what seemed a precision strategy but in the end relied on estimation.

> "I made 50g of plasticine and then took 20g away from it. 15g is half of 30g so I found out what 20g and 10g looked like and estimated how much plasticine to add to my 10g ball."

Precision Approaches. Use of the 20g mass led 22 percent of the students into precision approaches. Having obtained the 20g of plasticine, they then systematically reduced this to 15g by:

- halving the 20g lump, then halving one of the 10g lumps and adding a 5g lump to the first 10g lump, or

- halving the 20g lump, halving both the 10g lumps and then combining 3 of the 5g lumps.

The first, more efficient, approach was used by 14 percent of students, the second by 8 percent.

The most popular precision method involved the use of the 50g and 20g masses in opposite pans on the scales. With this approach, used by 26 percent of students, it was possible to make a 30g lump of plasticine in the first weighing operation, and then halve it. Of the students using this approach, only 41 percent specifically mentioned using the scales for the halving process, although more may have done so in practice.

Two other students seemed to start with this approach but failed to halve the 30g lump, and so failed the task.

A variation on this approach, used by six students, was to use the 50g mass to produce a 50g lump of plasticine. The 20g mass was then weighed against plasticine removed from this lump, until 30g of the lump was left. This was then halved.

Use of a precision approach was usually associated with a good written description. However, there were exceptions as one student's explanation shows:

> "First a put the 50g weight in one pan and the twenty in another then a I balanced it which maid thirty then I halft it."

One enterprising student seemed to set out using this approach but with success within his/her grasp used it only as an entry into the simpler precision approach involving only the 20g mass:

> "50g in left balance. 20g in right balance. added into right until it was 50g also. took out 20g from right. 30g left. put 20g in left balance. took out from right until equal, stuff taken out was 10g. broke 20g one up till both were 10g. divided the 10g into two and weighed the bits till they were equal. added 10g bit to 5g bit = 15g."

Convoluted. but the student achieved success -- and a very long sentence!

Another 11 students used precision approaches of varying complexity. For instance, at one end of the scale:

> "I used the 20g first and made two 20g balls. I then made a 10g ball using the 20g weight. I then measured all the balls made with the 50g weight then took the 10g and made another 5g ball. and stuck the 10g and the 5g ball together."

At the other end of the scale, one student used the 20g mass to produce two 20g plasticine lumps. He/she halved one of these, added it to the other and halved the resulting 30g lump.

Two students used both the 50g and 20g masses to produce a 70g lump of plasticine, halved it to give two 35g lumps, then used the 20g mass to remove 20g from it, leaving the required 15g.

What does this mean?

Almost all of the students tackled this problem and produced a lump of plasticine, over three-quarters of the lumps being between 10g and 20g. The task instructions did not ask for the lump of plasticine to be precisely 15g, and it is interesting to speculate whether this would have changed the proportions of students using estimation and precision approaches. In some countries and provinces, one or the other approach was predominant -- for instance in Taiwan and Scotland about twice as many students used precision approaches as estimation ones. In contrast, between two and four times as many students in Alberta and Saskatchewan used estimation approaches as compared to precision ones.

The explanation for these differences presumably lies, at least in part, in the curriculum and teaching methods in the countries and provinces involved. However, our overall background variables can do nothing to illuminate such relationships, as neither problem solving in small groups nor students saying that "knowing how to solve problems is as important as getting the correct answer" appear to be correlated with the approaches used in this task.

In Scotland, the most popular approach used by 26 percent of students was also the most economic in terms of weighing operations and arguably the most complex conceptually. This involved halving a 30g lump obtained by using both the 50g and 20g masses on opposite pans of the scales. The simpler halving of a 20g lump obtained by straight weighing using the 20g mass and then halving again, before recombining, attracted 22 percent of Scottish students. Estimation using the 20g mass attracted 20 percent of the Scottish students. One can only speculate whether the increasing emphasis put on problem solving in the Scottish school curriculum in recent years influenced these proportions. A similar exercise at age 9 might have illuminated the influence of intellectual development on students' problem solving capabilities. Next time perhaps!

Acknowledgements

The success of the IAEP assessment was dependent on the talents and dedication of many staff within the participating countries and the international coordinating center. This report is the result of the hard work, persistence, and creativity of these professionals. A number who have made outstanding contributions are acknowledged below.

The IAEP project was directed by Archie Lapointe and Nancy Mead of Educational Testing Service. Working with them was Janice Askew of Educational Testing Service. Sampling was conducted by Adam Chu and Leslie Wallace of Westat, Inc. Assessment development was coordinated by Brian Semple of the Scottish Education Department and Derek Foxman of the National Foundation of Educational Research, with the assistance of Ray Stark of Jordan Hill College. Data processing was conducted by the Canadian Data Analysis Group: Jean Guy Blais, University of Montreal; Maurice Dalois, GRICS; Michel Frechette, GRICS; Leo Laroche, Quebec Ministry of Education; Solange Paquet, Educan. Report preparation was coordinated by David Hobson and major contributions in its production were made by JoAnn Calabrese of Educational Testing Service.

Finally, thanks must be given to the students, teachers, and administrators whose cooperation and interest made the assessment possible.

Alberta

Dennis Belyk, Alberta Education
Chantey MacKay, Alberta Education

British Columbia

Bill Postl, Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights
Daina Watson, Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights

Nova Scotia

Bette Kelly, Department of Education
Turney Manzer, Department of Education

Ontario

Michael Kozlow, Ministry of Education
André Vézina, Ministry of Education

Saskatchewan

Rick Jones, Saskatchewan Education
Beverley Mazer, Saskatchewan Education

England

Clare Burstall, National Foundation for Educational Research
Derek Foxman, National Foundation for Educational Research

Scotland

Cathy MacGregor, Moray House College of Education
Brian Semple, Scottish Education Department

Soviet Union

Galina Kovalyova, Russian Academy of Education

Taiwan

Jong Hsiang Yang, National Taiwan Normal University