DOCUMENT RESUME

ED 258 995 TM 850 363

AUTHOR McLean, Les
TITLE Drawing Implications for Instruction from Item, Topic, and Classroom-level Scores in Large-scale Science Assessment.
INSTITUTION Ontario Inst. for Studies in Education, Toronto.
SPONS AGENCY Ontario Dept. of Education, Toronto.
PUB DATE Apr 85
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985). Research was funded under contract by the Ministry of Education Ontario.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Bound Students; Instructional Improvement; *Item Analysis; *Item Banks; *Science Instruction; Science Tests; Scores; Secondary Education; Secondary School Science; *Student Motivation; Testing Programs; Test Results
IDENTIFIERS *Ontario Assessment Pool Instrument

ABSTRACT

Data gathered from large-scale assessments in the Ontario Assessment Instrument Pool (OAIP) are examined. Implications for science instruction are to be found at the item level; the items should not involve more than two or three steps if the responses are to be informative. Items are collected and linked to provincial curriculum guidelines. These are given a preliminary trial in a few schools. When a few thousand items have passed the screening trials and reasonable coverage of the curriculum has been attained, many of the items are used in a field trial. An item sampling design is used to administer many of the items to a representative sample of students from the target population. Items were discussed in a set if they were coded to a topic and involved common content or technique. The sets were analyzed and assigned to one or more of seven categories according to the implications drawn. The categories are described and discussed. (DWH)

Drawing Implications for Instruction from Item, Topic and Classroom-level Scores in Large-scale Science Assessment

Les McLean
Ontario Institute for Studies in Education
Toronto, Ontario, Canada

Paper presented as part of the symposium *Alternative Assessment Strategies for Improving Science Instruction*, AERA annual meeting, Chicago, April 2, 1985. The work on which this paper is based was funded under contract by the Ministry of Education Ontario.

Drawing Implications for Instruction from Item, Topic and Classroom-level Scores in Large-scale Science Assessment

Drawing implications for instruction from large-scale assessment may seem to some of you to be a contradiction in terms. Most province-wide (or statewide) assessments have other objectives--usually to monitor achievement levels at the school or district level. The implications for teaching are very general ones, if they are sought at all. After the experience of the first province-wide achievement survey in Ontario in 1981 (McLean, 1982) and Ontario's participation in the Second International Mathematics Study in 1982 (McLean, Raphael, & Wahlstrom, 1983; Raphael, Wahlstrom, & McLean, 1983), it was decided to seek information directly useful to teachers during the chemistry and physics field trial that followed in 1983. (The monitoring and item pool development objectives were retained.)

One inspiration for this ambition was the science monitoring work of the Assessment of Performance Unit (APU) in England. The APU had been doing assessments in England, Wales and Northern Ireland for a number of years, innovating with *practical* exercises in mathematics (e.g., APU, 1981), creative approaches to English assessment (e.g., Gorman, White, Orchard, & Tate, 1983). The science teams went farthest in their attempts to stay close to classroom practices and derive results directly useful to teachers. At the same time, and perhaps stemming from the same motivations, they opposed the use of scaling techniques based on the Rasch model.

At first they produced the usual thick reports that were very widely unread (e.g., Harlen, Black, & Johnson, 1981; McLean, 1982), but then shifted to small (15 by 21 cm), short (30 pages or so) paperbacks in an attractive series, *Science report for teachers*. Six reports have been published (by the Association for Science Education rather than the Department of Education and Science) covering science assessment at 11, 13 and 15 years of age.[1] The reports for teachers typically have sections entitled "Main findings" (4 pages), "Implications for practice" (4.5 pages), "Examples of test questions and children's responses" (14 pages) and "Summary of implications" (2 pages). There are pictures and graphs. A separate report on "The Assessment Framework" explains their view of science and of children and the way the assessment was structured.

A recent review found published science tests reflecting "lack of innovation and, adherence to narrowly defined concerns", but also cited examples to support the thesis that

---

[1] The Association for Science Education, College Lane, Hatfield, Hertfordshire, AL 10 9AA. In 1984, the price was L1.00 each, or L3.75 for any 5 booklets, including postage within the UK.

"innovations in testing in science education abound" (Wilson, 1981, p. 268). The above references would add support to the thesis, and refine it somewhat with the observation that it is not innovative *tests* we are seeing, but innovative exercises and uses of those exercises.

In answer to the question, "What can schools do to facilitate motivation and achievement in science?", Maehr (1983) offered four recommendations:

1. *External evaluation should be minimized.* Sustained motivation is not enhanced by external examinations and may be diminished. The social competition they foster is likely to have a negative effect on all but the highly competent elite.

2. *Choice and freedom of movement must be fostered.* The most desirable long-term goals are fostered by freedom and independence in learning.

3. *Social competition must be minimized.* Those who excel in science seem to be able to operate effectively under most conditions, but the majority will not achieve the understanding of science society needs under conditions of strong social competition.

4. *Task-oriented classrooms are most productive.* The author seems to argue for a mastery learning approach, feeling (albeit tentatively) that it will be preferable to an "ego goal" approach.

The Science Council of Canada recently issued a series of reports on a four-year study of science teaching.[2] The eighth and last of their major recommendations was:

> Assessment techniques must be developed and implemented for *all* the objectives of science education to inform individual students about their progress and to monitor the effectiveness of provincial science education systems. (emphasis in original)

The authors stressed that school must not confine themselves to what John Goodlad (1983) called "the small piece of academic shoreline we measure with achievement tests". The Science Council report hardly mentioned tests at all, concentrating on teaching approaches and curriculum emphases.

The conclusion one comes to is that test scores will contribute little, if any, to the improvement of science teaching and learning. One can have a task orientation to evaluation as well as to teaching if one utilizes a variety of tasks. In this paper, we look at what can be learned from large-scale assessments, and we find that implications for instruction are to be found but that they are at the item level. Moreover, in assessment situations, the *tasks*, or

---

[2] See, e.g. Science Council, 1984, for a summary and reference to all the reports.

items, must not involve more than two or three steps if the responses are to be informative. Much can be inferred by an experienced science teacher from responses to a large variety of multiple-choice items, but the multiple-choice items must also not be too complex.

Province-wide Assessment in Ontario

Assessment has always been part of the development of the OAIP--Ontario Assessment Instrument Pool (Ministry of Education, 1980). First, a large number of items (instruments) are collected and linked to the provincial curriculum guidelines, and then these are given a preliminary trial in a few schools. When a few thousand instruments have passed the *screening trials* and reasonable coverage of the curriculum has been attained, many of the instruments are used in a *field trial*. Here, an item-sampling design is used to administer many of the instruments to a representative sample of students from the target population.

Field trials have a dual purpose--to submit the instruments to a final proving run under realistic conditions and also to estimate their difficulty. The field trial of the chemistry and physics *Assessment Instrument Pools* did not duplicate the practical approach of the APU teams, mainly because of the character of the pools.

The Chemistry OAIP has 2244 multiple-choice items and a much smaller number of other types, 120 of which are called *essay instruments*.[3] These require students to provide an extended response which may be entirely free or may be restricted by precise instructions. The responses may be oral, written, or graphic. Four essay-type instruments from the Grade 12 section and eight from the Grade 13 section were included in the 1983 field trial, along with 500 Grade 12 and 300 Grade 13 multiple-choice instruments.[4]

A large student sample was required to obtain reasonably precise difficulty estimates for so many instruments. A representative sample of schools was drawn (290 of 444 eligible schools), and where schools had more than two chemistry classes, the OISE team drew two at random; often there was only one. The details are described elsewhere (Talesnick & McLean, in press), but suffice to say that when the vast operation was over, records were in the files from 12,902 students in 620 chemistry classes.

---

[3]There are also 12 "Diagnostic", 66 "Storyline" and 28 "Laboratory" instruments in the 1983 version. Copies of the pool are available in two packages ($15Can. each from OISE Publications Sales or from the Ontario Government Bookstore).

[4]University-bound students in Ontario are required to complete six academic courses beyond Grade 12, usually requiring an extra year of study. These courses are taken in high schools in "Grade 13" at the moment, but a change is underway that will make it possible for many students to complete the requirements for university entrance by the end of Grade 12.

Questionnaires were completed by 276 teachers, describing how much time they spent on the guideline topics, how much homework they assigned, the emphasis they gave to objectives, how many courses they taught and much more. In many schools the same teacher taught two classes that were in the sample, so the responses on one questionnaire applied to two classes.

Because of the larger number of instruments designated for Grade 12, 60 per cent of the students were sampled from Grade 12 and the rest from Grade 13. In Grade 12, 54 per cent were male and in Grade 13 57 per cent. Ninety per cent planned to go on for some post-secondary study, 70 per cent to university--just what one would expect of students who elect science courses at the senior secondary level. About half of them said they planned on a career in science. It would be quite wrong, therefore, to generalize these results to high school students in general, but probably quite safe to regard the results as representative of able high school students who have an interest in science and mathematics.

Classroom-level Scores

An *item sampling* approach means that different students answer different items, and when there are 500 items (as in Grade 12 chemistry), the number of different sets is large. As a result, the only summary measures available at the classroom level are averages over topics, where topics are defined by clusters of items. The classroom mean is therefore the number of correct answers to items in that topic divided by the number of items in that topic *that were presented in that class.* Classroom means are usually based on different numbers of items from one class to the other, and controls have to be introduced to keep the variability within bounds. Small classes cannot be used (not enough responses to each topic), and checks must be made, class by class and topic by topic, that enough responses were obtained to make the class average meaningful.

Trimming the Data

Items must be valid measures of the content, that is, there must be good reason why an item is in one topic and not another. A strict definition of *topic* would probably imply a unidimensional latent trait, but the uses to which topics are put and the inferences made make such strictness unnecessary. As we see below, there are sources of variation other than content that swamp such niggling concerns. In this paper, the controls imposed were (a) classes with less than 10 students were set aside (about 9 per cent of the classes), and (b) a class mean was retained only if there were 10 or more responses from that class to items in that topic.

When one trims the data in this fashion, some topics are lost as well as classes. This further weakens the analysis, since now the content coverage is smaller and the sample is no longer representative. The remaining classes and topics do have the strength that classroom achievement can be linked to other classroom variables, such as time on topic and teacher characteristics. Means were available for about 350 Grade 12 classes[5] on each of 9 major topics (out of 11) and around 250 Grade 13 classes on 9 major topics (out of 14, all different from Grade 12). Table 1 contains a list of the Grade 12 topics and some means (over all classes) associated with them. Table 2 has the same list, but for Grade 13.

Analysis and Implications

The median number of hours teachers devoted to each topic gives some indication of its importance, but the variability in number of hours is what strikes one. Box-and-whisker summaries are displayed in Figure 1 for three topics, and one sees that some teachers do not teach at all a topic to which other teachers devote more than 15 hours! One can grant that these teacher reports are likely to be imprecise and still be amazed at the range. In Tables 1 and 2, there is a consistency among student reports of amount of teaching (OTL) and teacher reports (number of hours) that makes one believe that the range is real.[6] Classroom teachers in Ontario academic senior secondary schools are experienced, well-trained professionals (McLean, in press), and they exercise fully the freedom of choice given to them in the guidelines. There were positive correlations between student reports that a topic was taught and achievement (Table 1 and Table 2, 'Ach/OTL'), but the correlations between teacher time reports and student achievement (not shown) were small and non-significant.

Whereas the students rated each item as taught or not taught, the teachers were responding to abstract category (topic) labels, so one cannot infer too much from the lack of correlation between their responses and achievement. The guidelines from which the labels were taken were several years old at the time of the field trial and about 10 per cent of the teachers wrote on their questionnaries that they did not use those labels in describing their course. According to the field trial Advisory Committee, however, the labels are widely known and used and lack of familiarity is an inadequate explanation for the large range of time reports. It is known that these times vary from teacher to teacher.

---

[5]The number of classes on which the means were based varied from topic to topic: 344 to 355 in Grade 12. The range was greater in Grade 13: 8 topics in the 248 to 260 class range, and one each at 212, 148 and 126.

[6]The students rated each item as "Taught this year", "Taught before" or "Not yet taught", and these ratings were averaged over items in a topic and over students in a class to arrive at average classroom OTL (the same way item responses were averaged to arrive at achievement).

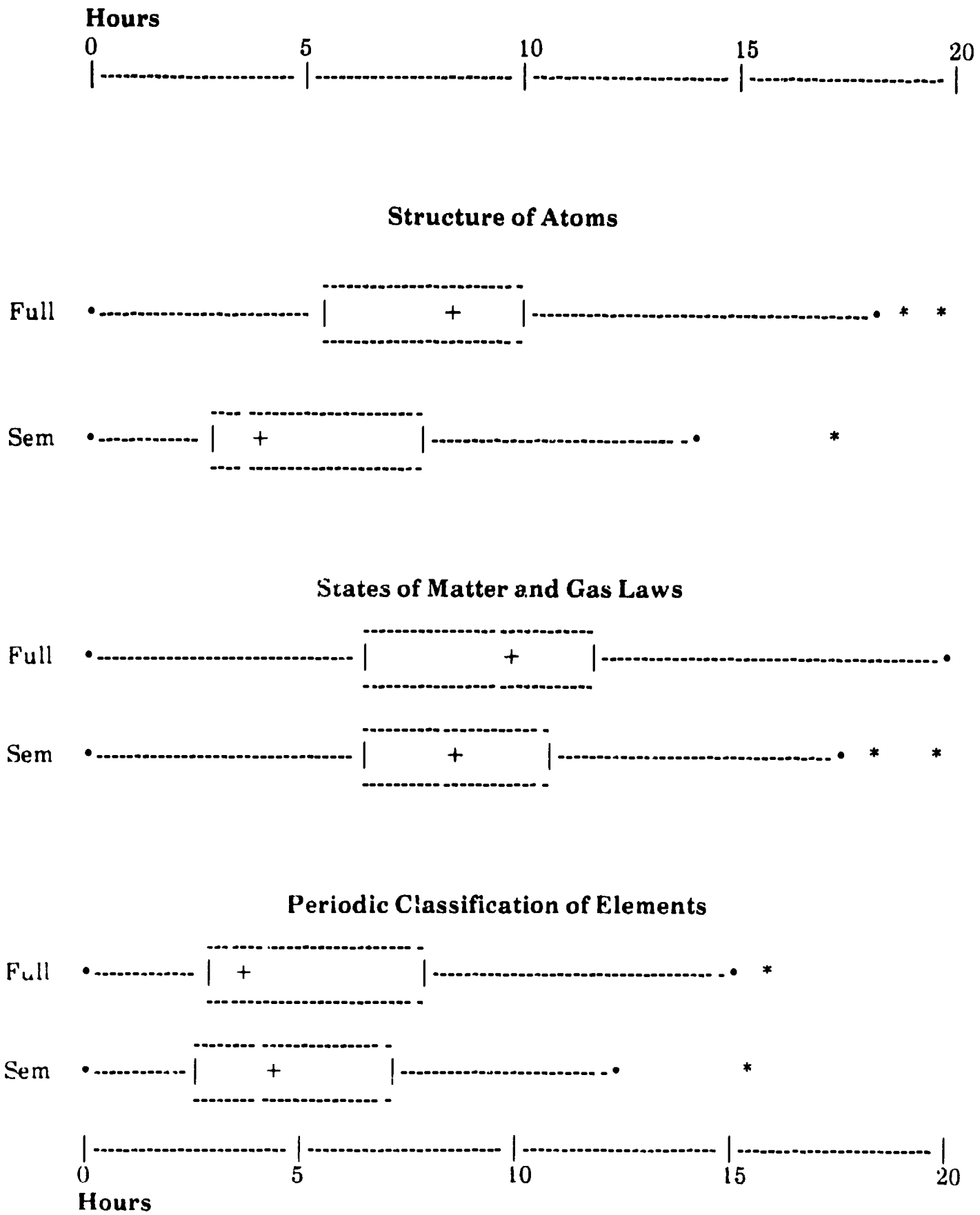
Table 1: Grade 12 Chemistry Topics and Associated Statistics for Correlation between Achievement and OTL, Opportunity to Learn (OTL), Median Number of Hours Devoted to the Topic, Average and Standard Deviation of Class Achievement Means.
OTL Scale: 1 = not taught, 2 = taught before, 3 = taught this year

Topic	Ach/ OTL	OTL	No. Hours	Mean Ach.	s.d. Ach.
Structure of Atoms [Unit 2]	.37	2.6	6	0.53	0.13
Structure of Aggregate Atoms [Unit 3]	.33	2.5	5	.45	.11
States of Matter and Gas Laws [Unit 4]	.47	2.3	9	48	.11
Oxygen and Hydrogen [Unit 5]	.29	2.3	3	.51	.11
The Mole, Atomic Weight, Molecular Weights [Unit 6]	.60	2.6	10	.55	.12
Formulas, Nomenclature and Equations [Unit 7]	.38	2.6	10	.52	.11
Water and Solutions [Unit 8]	.31	2.2	4	.45	.11
Ions in Aqueous Solution [Unit 9]	.37	2.2	3	.41	.11
Elements of Group 2 [Unit 10]					
Elements of Group 7 [Unit 11]					
Periodic Classification of Elements [Unit 12]	.50	2.4	4	.46	.09

Table 2: Grade 13 Chemistry Topics and Associated Statistics for Correlation between Achievement and OTL, Opportunity to Learn (OTL), Median Number of Hours Devoted to the Topic, Average and Standard Deviation of Classroom Achievement Means.
OTL Scale: 1 = not taught, 2 = taught before, 3 = taught this year

Topic	Ach/ OTL	OTL	No. Hours	Mean Ach.	s.d. Ach.
Energy Effects in Chemical Reactions [Unit 2]	.18	2.6	10	.52	.12
Rates of Chemical Reactions [Unit 3]	.20	2.7	10	.58	.11
Equilibrium in Chemical Reactions [Unit 4]	.43	2.8	11	.55	.13
Solubility Equilibria [Unit 5]	.08	2.5	7	.61	.14
Aqueous Acids and Bases [Unit 6]	.44	2.5	11	.55	.12
Oxidation-Reduction Reactions [Unit 7]	.60	2.3	10	.50	.14
Experimental Base for Atomic Theory [Unit 8]			1		
Electron Arrangement and the Periodic Table [Unit 9]	.49	2.2	2	.48	.12
Molecules in the Gas Phase [Unit 10]	.36	2.1	1	.45	.13
Bonding in Solids and Liquids [Unit 11]	.29	2.1	2	.51	.14
Chemistry of Carbon Compounds [Unit 12]	.55	2.0	5	.50	.18
3rd Row of the Periodic Table [Unit 13]					
4th Row Transition Elements [Unit 14]					
Some 6th and 7th Row Elements [Unit 15]	.40	1.9	0	.46	.17

Figure 1: Box-and-whisker displays for number of hours chemistry teachers reported spending on grade 12 topics in full-year and in semester classes



There were similarly large ranges in achievement among the classes, as can be inferred from the means and standard deviations in Tables 1 and 2.[7] The implications are that one is unlikely to derive implications for instruction from classroom-level summaries in large-scale assessment. There are too many unmeasured variables and there is too much variation to get at classroom processes from this vantage point.

Lessons from the Essay Questions

This section can be short because there were many fewer instruments and because many of the lessons were negative ones. The first lesson was that senior secondary school students do not exert themselves on *tests* unless the results are important to them. Figure 2 contains an example of one of the essay questions, along with the marking scheme and some observations by the teacher who marked the responses. Only about one-third of the students attempted to answer in any serious way, and those who did usually got bogged down early in the solution and were unable to show much of what they knew.

In other words, the second lesson was that complex, multi-step questions are poor instruments for use in large-scale assessment because you learn very little from them. It would help if the students were highly motivated, but that is difficult to arrange in the assessment framework of school and item sampling. If open-ended questions are used, they should be of the short-answer variety. There is such a thing as being too open-ended.

The second point would be equally valid for an end-of-course or secondary school graduation examination where there was no opportunity to discuss or give feedback on the answers. Several multi-step questions can be asked if there is time, where the solution to various combinations of the steps are given and the students asked to supply the missing link. In large-scale assessment, the *same* problem can be given in several different forms to different samples of students to determine which steps are causing problems in the student population as a whole.

Item-level Summaries

Professor Irwin Talesnick, of Queen's University, Kingston, Ontario, has to be recognized as a gifted and energetic analyst for his insightful summary of results on 800 items. It helps to understand why he might try (and why he might succeed) if you know that

---

[7] These summary statistics were chosen for their brevity rather than their quality. The standard deviation is especially poor, as the distributions are skewed and leptokurtic. One can see, however, that the classroom means were quite widely spread.

Figure 2: Example of essay-type question given to grade 12 chemistry students. The marking scheme and step descriptors are in the Chemistry **OAIP.** The discussion was prepared by a teacher who marked the responses collected during the field trial

Coke is used in the steel manufacturing process. Coke oven gas is a by-product of the process. A certain steel plant (Acme Steel) produces 150,000 L (at 101 kPa of pressure) of coke oven gas per 24 h period. Government legislation allows only 5% of that amount to be burned off into the atmosphere due to the gas's pollution effects. The remaining gas has to be stored in a tank. Once every ten days the tank is emptied to provide for the plant's internal fuel needs. The tank is designed to withstand a maximum pressure of 1 MPa. The temperature range for the locality where the plant is located is -20° to -30°. The temperature of the gas at any one time can be assumed to be the same as the outside temperature. What is the smallest acceptable volume of the tank?

MARKING SCHEME

Total Number of Marks 15
Number of Major Steps 8

STEP DESCRIPTOR	Number of marks per step
1. Volume of gas stored per day	2
2. Volume of gas stored in 10 days	1
3. Temperature decisions	2
4. T correction	3
5. P correction	2
6. V of remaining gas at 101 kPa	1
7. P correction to 1 MPa	2
8. Volume of holding tank	2

This question is wordy and complex, and it appeared to intimidate and confuse many students. Opportunity-to-learn data shows that over half the students felt they had not been taught this material, suggesting that they did not recognize that this is a gas laws problem. As a result, most did not even attempt the first two steps, which did not require more than simple arithmetic.

Steps 3 and 4, requiring a volume change with temperature, were often ignored, possibly because of misleading wording. The prefix in MPa was often not understood, giving poor results of step 5. Very few students reached step 5, and none went further, possibly because students took the word "emptied" to mean exactly that. If the question had been worded "Every ten days the gas is discharged at atmospheric pressure", more students might have recognized that some gas was left behind. Units were frequently omitted, and there was no evidence that the students were aware of the importance of significant figures.

he headed the committee that developed the chemistry OAIP and chaired the field trial advisory committee. He defined sets of items by subtopic within topic and prepared the detailed analysis in the report on the field trial (Talesnick & McLean, in press). The 800 items and per cent choosing each alternative are displayed in the report, along with the 12 essay instruments. Obviously, only a few snippets can be reported here, but even a few should suffice to show how useful such summaries can be for instruction. The implication is that *virtually all of the pedagogically relevant information in test responses is to be found at the item level.*

Summarizing the Summaries

Only the 500 Grade 12 items will be considered in this paper. T ey would seem to be quite enough, and Ontario Grade 12 is comparable to the final year of secondary school in the USA and in all other provinces of Canada except Quebec. Professor Talesnick found 224 *sets* to discuss, many of them single items but a few containing 12 to 15. Items were discussed in a set if they were already coded to a topic and involved common content or technique. For the purposes of this paper, the sets were analyzed and assigned by the present author to one or more of seven categories according to the implications drawn. About 10 per cent were not assigned, since no inference was d n, and about the same number were the source of more than one implication, since the tota ne to 226. The categories, numbers of sets assigned and proportions are reported in Table 3. No one has attempted to replicate the categorization, so it must be regarded as a possibly idiosyncratic assignment.

Specific point about content. The largest number of sets gave rise to a specific pedagogical point about chemistry content. Here are a few examples:

> Responses to number 198 suggest that electronegativity and electropositivity were not understood as *tendencies.* The same error was made on instruments similar to this one. The gain and loss of negative charge was confused with the gain and loss of particles. Students appeared to think that loss (in this case, of electrons) was related to increased electronegativity. [Unit 2]

> More than 30 per cent of students responding to each of the instruments in this group (12, 39, 42, 60) thought that the diamond lattice consists of ionic bonding. This is a double misconception. Students assumed that ionic compounds are hard and that all crystalline substances must be ionic. Such misconceptions can often be overcome by performing a number of simple experiments to test the properties of ionic as well as other types of crystalline substances--and calling attention repeatedly to the nature of the substance under test. [Unit 3]

Evaluation of achievement. The next larger number of implications concerned the

Table 3: Classification of Grade 12 Achievement Item Sets by Implication Drawn from Them

Category	No. sets	per cent
Specific pedagogical point--content	59	27
Evaluation technique	46	20
Specific pedagogical point--technique	31	14
General pedagogical point	27	12
Definitions, symbols, language	22	10
Problem with prerequisites	9	4
Good performance	31	14

process of evaluation. This is not surprising, since one purpose of the field trial was to learn about the instruments, but the two categories under pedagogical points together accounted for 41 per cent of the inferences and evaluation technique only 20 per cent. There were many good suggestions, as a few examples will illustrate:

> The distribution of responses to number 57 suggested that students guessed at random. This instrument is too demanding for the multiple-choice format, since the student must analyze each molecule before attempting an answer. This instrument would be better suited to a full, written response. [Unit 3]

> The selection of two true statements is always dificult, as number 88 showed.[Unit 6][8] Two of the possible responses included statement IV and these two possibilities attracted 58 per cent of the respondents. If students had the opportunity to justify their answers they could be awarded partial credit for their work on an instrument of this type.

> The group of 15 instruments (15 to 190) was generally well answered. Students were able to apply formulas such as MF_3 to the corresponding oxide, but they were not able to relate the fluoride to a phosphate. The phosphates were clearly not as familiar to the students as the oxides. Students were not as able to select one *incorrect* formula from the group of four formulas listed in instrument 16 as they were able to select the one *correct* formula in instrument 25. This is encouraging,

[8]Stems containing distractors such as "A and B" have been studied under the heading "complex multiple-choice (CMC)". One study found they failed to measure achievement and might inhibit improvement in education (Kolstad, Briggs, Bryant, & Kolstad, 1983).

because selecting a correct formula is better pedagogy than selecting an incorrect formula. Emphasizing negative instances reinforces incorrect ideas. [Unit 7]

The instruments in the set 76 to 156 illustrated again that the numbers that appear in the stem are chosen by many students as the correct answer. They appeared to have no appreciation that the molar volume of gases is 22.4 L at STP. The alternative selected strongly reflected numbers specifically included in the stem of the instrument. Students should be required to justify their answers, or else an entirely different instrument type should be used for this topic. [Unit 5]

Each of the instruments 75, 84, 85 and 92 requires that the students perform three operations: interpret the equation, calculate molar ratios and calculate the number of moles of product formed or reactant consumed. The students continued to confuse mass and volume, confuse mass and number of moles and invert molar ratios. All of these instruments were answered poorly. Students will continue to answer instruments of this type poorly, and unless they are required to justify their answers the responses will provide little useful information. As questions, they would be poorly suited to summative or formative evaluation. [Unit 7]

Specific points about technique. There is clearly a fine line between a point of technique and a point of content, but it seemed useful to mention some of these examples separately.

Generally good performance was observed on this type of instrument. The statistics indicate that students know the definitions but have some difficulty in using the atomic and mass numbers to calculate the numbers of each of the particles in an atom. [6, 24, 102, 204, 222 and 228; Unit 2]

The volume of a gas varies directly with temperature *on the Kelvin scale.* Temperatures were given in instrument 129 in °C. About 40 per cent of the respondents did not change the temperatures to the Kelvin scale, and this was so whether they said they had been taught it or not. [Unit 4]

General. A number of the inferences extended across most, if not all, of the curriculum and deserve special mention. Included in this category were comments that the material appeared not to have been taught.

As revealed in instruments 33 and 102, the vast majority of the students could not distinguish between mass of solvent and mass of solution. The difference is fundamental and must be reviewed carefully when the solubility topic is studied in Grade 12. The two different solubility expressions are used interchangeably and considerable experience is necessary in order to shift from one system to the other without too much difficulty. [Unit 8]

It is difficult to explain why almost half the students did not answer instrument 16 correctly. The compound contains the element oxygen with which all students should be familiar. It would appear that students should spend much more time writing formulas and equations. [Unit 9]

> The instruments in the set 3 to 129 involving the terms deliquescence and efflorescence were generally not answered as well as those involving terms such as dehydration, hydrate, etc. Students generally indicated that they had not studied this topic. The actual terms are not as significant as the experience with the materials. Students should be aware that many compounds do lose water spontaneously and others take on water spontaneously. [Unit 8]

Language, definitions, symbols. Occasionally, the language used or a special symbol was overly prominent. On other occasions, such as in the immediately preceding example, it was hard to separate the language from the content. If the meaning of a word or a symbol was the central focus of a comment, it was listed in this category.

> There was marked confusion between the term "valence electrons" and the word "valence" in instrument 84. It appeared that many students remembered that the "valence of sulphur is -2", but they did not recognize that the number of valence electrons is not always numerically equal to the common valence of an element. [Unit 2]

> The majority of students appeared to misinterpret the question. They tended to select the response which included two common numbers without regard to the arithmetic operations indicated in the response, and failed to distinguish between the mass of a *mole* of atoms (Avogadro's number of atoms) and the mass of a *mole of moles* of atoms. [105, Unit 6]

Problems with the basics. It was notable how few comments dealt with deficiencies in basic skills, even the skills *basic* to senior secondary science. The system has either weeded out those with poor preparation or else the instruments were not sensitive to such problems.

> These instruments (42, 96) revealed an arithmetic misunderstanding--the concept of weighted average. In calculating relative atomic mass, 27 per cent of the respondents calculated the simple average of the two numbers. Drill on this topic can be related to non-chemical as well as chemical situations. [Unit 2]

> All three of instruments 36, 123 and 135 were poorly answered but reported widely as taught. Students appeared not to have an understanding of ratio and proportion or the manipulation of symbols--a finding that appears in mathematics classes as well. About a third of the sample would have added 273 to change a temperature from the Kelvin to the Celsius scale. Half of the sample confused Boyle's Law with Charles' Law and selected $V_1T_1 = V_2T_2$ as the correct expression of Charles' Law. Even if there is confusion between the laws associated with Boyle and Charles students should recognize direct variation in the form $V_2T_1 = V_1T_2$ or its equivalent, but alas, this was not the case. [Unit 4]

Good performance. In about 14 per cent of the cases, the analyst was moved to note that the students had learned something reasonably well. Instruments that involved recall of factual material, simple calculations or straightforward application of information in the

stem of the question were well answered. The comment was frequently made that performance was much better among those students who said they had been taught the material. It would be easy to dismiss this as an artifact--students say they have been taught only when they know how to answer--but that would be too facile. In this and previous studies (McLean, 1982; McLean, Raphael, & Wahlstrom, 1983) students often reported they had been taught material when their performance was poor. Where corroborative evidence was available, the student reports showed considerable validity.

> This trio of instruments (19, 22, 26), which does not require knowledge of the Periodic Table, reveals that the students have been well trained in the electronic configurations of the inert gases and the significance of the stable octet. [Unit 12]

> Instruments 4 and 23 were answered correctly by more than 80 per cent of the students. The questions are straightforward and factual; the students had only to refer to the Periodic Table to verify their selections. [Unit 12]

Lessons from the Item-level Summaries

Most of the lessons are explicitly stated in the examples above and summarized in the categories chosen to report them. A few additional observations can be made:

- Useful information often came when responses to several instruments were compared. Such comparisons are available in abundance in large-scale assessment but are rarely made. This observation leads into the next one.

- Making inferences useful for instruction requires both substantive knowledge and a thorough familiarity with the item pool. So far, sophisticated mathematical or statistical manipulations cannot provide the insights that come to a knowledgeable expert who studies very basic summaries of the data. After seeing what a real expert can do with simple item summaries, it is tempting to suggest that the sophisticated statistical analyses be kept well away from the content-oriented analyses, lest the substance be lost in the smoke and mirrors of, for example, item response theory.

- Achievement instruments in general, and multiple-choice instruments in particular, are difficult to construct well. There are many ways to do so, however, and most are not technical in nature. Several examples were given above. Careful attention to language and avoidance of incorrect content are general principles that bear repetition. Multi-step questions must be used with great care, usually not at all in multiple-choice format. There is often no substitute for full, written answers, but asking for justification of an answer to a multiple-choice question is a long-known technique that is too seldom used.

- Where applicable, students should be asked if they have been taught the material needed to answer each item (OTL). Such a question is not applicable in most language classes, either mother tongue or other language, but it is appropriate in mathematics and in the physical and life sciences. One of the lessons emerging

from the Second International Mathematics Study is that there is curriculum diversity everywhere, even in countries where it was thought there was strong and effective central control. If we set out to draw implications for instruction, then we have to be sensitive to variation in instruction, and experience shows that the student OTL question can be of considerable value in large-scale assessment.

References

APU. (1981). *Mathematical development--Primary survey report 2.* London: HMSO.

Goodlad, J. I. (1983). A study of schooling: Some findings and hypotheses. *Phi Delta Kappan, 64*(7), p. 468.

Gorman, T. P., White, J., Orchard, L., & Tate, A. (1983). *Language performance in schools--Secondary survey report no. 2.* London: HMSO.

Harlen, W., Black, P., & Johnson, S. (1981). *Science in schools--Age 11: Report no. 1.* London: HMSO.

Kolstad, R. K., Briggs, L. D., Bryant, W. B., & Kolstad, R. A. (1983). Complex multiple-choice items fail to measure achievement. *Journal of Research and Development in Education, 17*(1), 7-11.

Maehr, M. L. (1983). On doing well in science: Why Johnny no longer excels; Why Sarah never did. Chapter 8 in S. G. Paris, G. M. Olson, & H. W. Stevenson (Eds.), *Learning and motivation in the classroom.* Hillsdale, NJ: Lawrence Erlbaum Associates.

McLean, L. D. (in press). *Teaching and learning chemistry in Ontario grade 12 and grade 13 classrooms--Teachers, students, content, methods, attitudes and achievement.* Toronto: Ministry of Education.

McLean, L. D. (1982). *Report of the 1981 field trials in mathematics and English . . . Intermediate division.* Toronto: Ministry of Education.

McLean, L. D., Raphael, D., & Wahlstrom, M. W. (1983, October). The Second International Study of Mathematics: An overview of the grade 8 study. *Orbit 67.*

Ministry of Education. (1980). *Ontario Assessment Instrument Pool: A general introduction.* Toronto: Ministry of Education.

Raphael, D., Wahlstrom, M .W., & McLean, L. D. (1983, December). The Second International Study of Mathematics: An overview of the grade 12/13 study. *Orbit 68.*

Science Council of Canada. (1984). *Report 36. Science for every student--educating Canadians for tomorrow's world.* Ottawa: Minister of Supply and Services. Available from the Canadian Government Publishing Centre, Supply and Services Canada, Hull, Quebec, Canada, K1A 0S9.

Talesnick, I., & McLean, L. (in press). *Student achievement in Ontario grade 12 and grade 13 chemistry classes--Report of the 1983 field trial of the chemistry OAIP.* Toronto: Ministry of Education.

Wilson, J. T. (1981). Toward a disciplined study of testing in science education. *Science Education, 65*(3), 259-270.