Development of a Physical Science Experimental Scenario Instrument for Measuring Teacher Trainees' Conceptions of Scientific Evidence

Tan Ming Tang
Jabatan Sains

ABSTRACT

This paper details the development of a physical science experimental scenario instrument for measuring teacher trainees' conceptions of scientific evidence. The five conceptions measured were those of repeated trials, evaluating the trustworthiness of data, treatment of anomalous data, internal and external validity of experimental design. Steps in determining the instrument's validity and reliability are presented. The potential advantages of assessing procedural understanding by written tests are also discussed.

INTRODUCTION

Scientists use various ways to collect and use data. The evidence collected can then be used to create new knowledge, expand existing knowledge or to solve problems. Although these various ways may differ, there are still some ideas or procedures common to many kinds of investigations (procedural understanding) which Millar, Lubben, Gott and Duggan (1994) categorized as being in the 'scientific evidence' (p. 245) domain. Roberts and Gott (2004) refer to "concepts of evidence" or "scientific evidence" as the understanding of a set of ideas that underpin the collection, verification, analysis and interpretation of data in order to handle scientific data effectively. These concepts of evidence involve cognitive abilities such as deciding on how many measurements to take, over what interval and range, how to interpret the pattern in the resulting data etc. and are in turn underpinned by scientific skills. Hence, collecting and using evidence in an investigative task is viewed as a tool kit to help in judging an experimental study for its design, the reliability of the measurements, the validity of the sample and the quality of the resulting data and its interpretation. Two of the most common and popular methods to assess procedural understanding in science are practical investigation and written scenario test.

RESEARCH CORRELATING STUDENTS' CONCEPTIONS OF SCIENTIFIC EVIDENCE IN PRACTICAL WITH SCENARIO TASK

Past research (Robinson, 1969; Ben-Zvi et. al., 1977) has found that there is a low correlation between practical and written tests of performance. In the APU survey involving 15 year old, Welford et al. (1985) found that most of them could describe the method to measure the mass of a single bung when presented with 1000 identical ones in a written probe but less than 10% could carry out the actual measurement properly in a practical situation. On this phenomenon of difference in performance on practical and written assessments, Welford et al. (1985) suggested that it is not necessarily that students do better on practical tasks but that they perform differently because the tasks themselves are different.

In the Assessment of Achievement Programme (AAP) in Scotland, pupils at the Primary 4 (8-9 years), Primary 7 (11-12 years), and Secondary 2 (13-14 years) were monitored on their procedural and conceptual science performance in the *Environmental Studies 5-14* subject. It was found that lower achieving children in all three age groups performed better on practical tasks than on paper and pencil tasks, even though the assessments were ostensibly measuring the same attribute (Stark, 1999).

Following on the heel of the AAP work, Gray and Sharp (2001) used both of these modes to assess the performance of a sample of about 128 Primary 6 (10-11 years old) pupils. Comparable tasks of practical and paper and pencil formats with as many variables controlled as possible were used. Once again, it was found that pupils, particularly lower achievers, performed better on more interactive practical than on comparable written tasks. According to Gray and Sharp (2001), the difference in performance in both the practical and written tasks could be due to something inherent in the tasks itself that causes individuals to perceive the tasks differently or that the very nature of the tasks themselves have been altered by their respective modes of assessment. If it is the latter, Gray and Sharp (2001) argued that these two modes of assessment could be actually measuring different attributes and thereby causing the validity of the entire assessment to be questionable.

In another study by Lawrenz et al. (2001), four assessment methods were compared and they found that each method measured entirely different competencies altogether and this is particularly true for hands-on assessment. From here, they concurred with the findings of Gray and Sharp (2001) that 'different assessment formats may be measuring different things'.

To assess four aspects of performance, namely (1) planning and designing, (2) a hands-on investigation, (3) analysis and interpretation and (4) application, Solano Flores et al. (1999) constructed a shell in which tightly

defined criteria in the same context were used. Despite a moderate level of instruction being given to the 109 fifth graders from two public schools in California, it was found that the correlations between the two hands-on elements was close to zero (r = -0.01). Furthermore, the scores for the hands-on investigation did not correlate well with the scores for the other three written assessments in both tasks (r varies from -0.06 to 0.17).

From the review of these research findings correlating students' conceptions of scientific evidence in practical with written (scenario) task, it can be argued that a written test may be a useful complement to performance assessment in assessing procedural understanding in practical work. According to Roberts and Gott (2004), the potential advantages of the written test approach are:

a) a lighter touch assessment allowing teachers the freedom to teach open-ended investigative work outside the confines of the assessment system;
b) a focus is provided for discussion about experimental design, data analysis and the validity and reliability of evidence.

But can such a written test be developed? This paper details the steps taken to trial the written instrument's validity and reliability.

THE PURPOSE OF THE STUDY

The purpose of this study was to develop a valid and reliable physical science experimental scenario instrument to measure science teacher trainees' conceptions of scientific evidence.

METHOD

Sample

After obtaining permission from the Teacher Training Division of the Ministry of education, a pilot-test was carried out on a convenient representative sample of 29 final year science teacher trainees (15 males, 14 females) in an intact class from a teacher training college in the Kuching division. This pilot sample of science teacher trainees was rather similar to the actual sample of the study in that that they were trained in the same college and were in their final year of training. This pilot sample will not be involved in the actual study later on.

Formulation of the Physical Science Experimental Scenario Instrument

A typical unsound scenario described student investigations that did not produce the desired data. In this present study, the sample science teacher trainees reviewed two hypothetical scenarios with unsound experimental data sets and/or related conclusions and then respond to questions such as 'what would you do?' or 'what should you do?'. These two experimental scenarios, designed from the topic 'Force and Motion', were incorporated into this study to probe the science teacher trainees' conceptions of scientific evidence. The basis for structuring the scenarios around this topic was because it is the fundamental topic for the physical science subject. Moreover, a large number of the experiments in this topic utilizes Type 2 investigations (one independent and one dependent variable, both continuous) which was also the type of investigative scenario selected for use in this present study.

This paper and pencil instrument was developed by absorbing various aspects of the target conceptions in Lubben and Millar's (1996) PACKS project and Taylor's (2001) Classroom Passages Protocol. Each unsound hypothetical experimental scenario contained five different data sets to test the trainees' conceptions on five scientific evidence aspects. The five aspects investigated were that of repeated trials, evaluating the trustworthiness of data, treatment of anomalous data, the internal and external validity of experimental design (Table 1). For each correct response, a score was given. The breakdown of the scores into individual scientific evidence aspects is very useful in enabling the researcher to pin-point areas of relative strength or weakness.

To ensure construct and face validity of the unsound hypothetical physical science experimental scenarios, two experienced science lecturers (with more than 10 years teaching experience) were asked to examine its content and design to ensure that the task and its measuring instrument could adequately measure the underlying conceptions of scientific evidence. They suggested that the definition for 'believable' data should be told or stated in the related question asked so as to avoid any misunderstanding on the part of the respondents.

Table 1

Target Conceptions for Prompts 3 to 7 in the Unsound Hypothetical Physical Science Experimental Scenario

Prompt Number	Target Conception
3b	The need and rationale for repeats
4	Evaluating the trustworthiness of data as a measure of reliability
5	Recognition and treatment of anomalous data
6	Fair test
7	Manipulation of independent variables (range and interval)

Both the instrument and its measuring scale were initially formulated in English by the researcher before being translated into Bahasa Malaysia. In order to ensure that the Bahasa Malaysia content of each instrument has not deviated from its original English version, the three-step back-translation procedure (Brislin, 1986), was used to check on the accuracy of the translation.

Pilot Testing the Instrument

The teacher trainees were required to respond to two prompts designed to examine pertinent subject matter knowledge and to five other prompts of each hypothetical physical science experimental scenario (a sample item is shown in Appendix A), aimed at investigating specific conceptions of the scientific evidence associated with the measurement reliability and design validity categories. In prompts one and two, the trainees were asked to identify all the variables affecting a given dependent variable and then to describe the nature of each relationship. In prompts three to seven of the experimental scenarios, the trainees were asked to respond to unsound student-collected data sets and/or flawed conclusions.

This instrument was pilot-tested on a convenient representative group of 29 final year science teacher trainees and this was followed by the retest on the same sample a month later. The instrument was collected back after the first session and the trainees were not informed that there would be a retest. Before the administration of the instrument, the trainees were told that they were being involved in a survey but at the same time, they were advised to try their best. It was observed that on average, the trainees took about ninety minutes to complete each session.

FINDINGS AND DISCUSSION

The unsound hypothetical experimental scenario instrument was checked for its reliability in measuring science teacher trainees' conceptions of five scientific evidence aspects. By using its measuring scale, the trainees' responses were rated by the researcher and two other experienced science lecturers. Approximately 34% (10 trainees) of the above pilot sample's scripts were chosen randomly for the inter-rater reliability check. A discussion was held with the two science lecturers involved to clear up or clarify any question they might have regarding the measuring scale guidelines before commencing the rating exercise.

For the total conception score in the measurement reliability and design validity categories in both the scenarios, the *k* values obtained for the inter-rater agreement between lecturers across ratings in the 10 aforementioned pilot sample responses, were at least .90, indicating a high level of agreement between raters. Also, data source triangulation was used to examine the respondents' conceptions of scientific evidence across two different physical science contexts of the scenario task. According to Creswell (1998), this form of triangulation involves the corroboration of evidence collected from the same data source (usually a person) but in different contexts or under different conditions. The Cohen's (1988) kappa values obtained for each pairs of the five aspects in the two scientific evidence categories was found to be reasonably high, ranging from .77 to . 87 (Table 2).

Table 2

Cohen's (1988) Kappa Values for the Five Scientific Evidence Aspects in the Unsound Hypothetical Experimental Scenario Instrument of the Pilot Study

Scientific Evidence Aspects	Rationale of Repeats	Evaluating the Trustworthiness of Data	Treatment of Anomalous Data	Fair Test	External Validity Aspect
k	.87	.77	.77	.86	.82

After a month, a retest of the above instrument was carried out on the same pilot sample students, followed by another session of ratings by the researcher. Table 3 shows the test-retest correlations for each of the five aspects of scientific evidence in the unsound hypothetical experimental scenarios. The high Pearson-r correlations obtained show that the unsound hypothetical experimental scenario is a reliable instrument to measure science teacher trainees' conceptions of scientific evidence in Malaysia.

Table 3

Test-Retest Correlations for the Five Scientific Evidence Aspects in the Unsound Hypothetical Experimental Scenario Task of the Pilot Study

Scientific evidence Aspects	Rationale of Repeats	Evaluating the Trustworthiness of Data	Treatment of Anomalous Data	Fair Test	External Validity Aspect
Pearson-r	.79	.83	.72	.72	.70

As mentioned earlier, the five items in the unsound hypothetical experimental scenario task were adapted from Lubben and Millar's (1996) PACKS project and Taylor's (2001) Classroom Passages Protocol. To test whether all these items are appropriate to probe local students' conceptions of scientific evidence, an item analysis was also conducted. By using the results of the pilot study, the items in the unsound hypothetical experimental scenario instrument were analyzed by calculating the facility index (FI) and item discrimination index (DI). The facility index of an item denotes the percentage of subjects who have answered the item correctly whereas the item discrimination index compares the percentage of top students with the percentage of poorer students who have answered the same item correctly. For the former, a good spread in results can be obtained if the mean of the FI of items is around 50% or 60% and if the FI of the items vary from about 20% to 80%. For the latter, the discrimination index of all items should be positive, that is, more of the better students should answer the item correctly (Bloom et. al, 1971).

The score awarded for a particular aspect in the two given scenarios was based on their average score. The FI and DI of the items in the scenario task are shown in Table 4. The FI of the five items ranges from 20.7% to 79.3% with a mean value of 46.9%. Since the item facility values obtained are reasonable and there are no negatively discriminating items, it was decided to retain all items for use in this present study.

Table 4

Facility Index and Discrimination Index of the Items in the Unsound Hypothetical Experimental Scenario Instrument

Item	Scientific evidence	Facility Index (%)	Discrimination Index (%)
3	Repeats	48.3	80.0
4	Evaluating the trustworthiness of data	79.3	40.0
5	Anomalous Data	44.8	90.0
6	Fair Test	41.4	50.0
7	External Validity	20.7	60.0

CONCLUSION

This paper describes the development of a written test instrument that facilitates the assessment of students' conceptions of scientific evidence in practical work. It consists of two hypothetical scenarios with unsound experimental data sets and/or related conclusions that aim to provide a cost-effective and systematic approach to complement performance assessment in assessing procedural understanding. The Cohen's Kappa values obtained for inter-rater reliability and data source triangulation during the coding of categorical conceptions were found to be reasonably high. As for the test-retest stability of the instrument, the high Pearson-r correlations obtained show that the unsound hypothetical experimental scenario is a reliable instrument to measure science teacher trainees' conceptions of scientific evidence in Malaysia.

REFERENCES

Ben-Zvi, R., Hofstein, A., Samuel, D. & Kempa, R. F. (1977). Modes of instruction in high school chemistry. *Journal of Research in Science Teaching, 14,* 433-439.

Bloom, B. S., Hastings, J. T., and Madous, G.F. (1971). *Handbook of formative and summative evaluation of student learning.* New York: McGraw Hill.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner and J. W. Berry (eds.), *Fields methods in a cross-cultural psychology (pp. 137-164).* Newbury Park, CA: Sage Publishers.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Creswell, J. (1998). *Qualitative Inquiry and Research Design; Choosing Among Five Traditions*. London, New Delhi, Thousand Oaks, Sage Publications, 372 p. + notes, index.

Gray, D. and Sharp, B. (2001). Mode of assessment and its effect on children's performance in science. *Evaluation and Research in Education, 15*(2), 55-68.

Lawrenz, F., Huffman, D. & Welch, W. (2001). The science achievement of various subgroups on alternative assessment formats. *Science Education, 85*(3), 279-290.

Lubben, F. and Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education, 18*, 955-968.

Millar, R., Lubben, E., Gott, R. and Duggan, S. (1994). Investigating in the school science laboratory: conceptual and procedural knowledge and their influence on performance. *Research Papers in Education, 9* (2), 207-248.

Roberts, R., and Gott, R. (2004). A written test for procedural understanding: a way forward for assessment in the UK science curriculum? *Research in Science and Technological Education, 22*(1), 5-21.

Robinson, T. J. (1969). Evaluating laboratory work in high school biology. *The American Biology Teacher, 34*, 226-229.

Solano-Flores, G., Jovanovic, J., Shavelson, R. J. and Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education, 21*, 293-315.

Stark, R. (1999). Measuring standards in Scottish schools: the assessment of achievement programme. *Assessment in Education, 6*(1), 27-41.

Taylor, J. A. (2001). *Secondary School Physics Teachers' Conceptions of Scientific Evidence: A Case Study*. Unpublished doctoral dissertation, The Pennsylvania State University, USA.

Welford, G., Harlen, W. and Schofield, B. (1985). *Assessment of Performance Unit. Science Report for Teachers: 6. Practical Testing at Ages 11, 13 and 15.* London: Department of Education and Science.

APPENDIX A

Albert wanted to find the relationship between the slope angle and the time taken for the ball to roll down the incline plane. By varying the slope angles, the time taken for the ball to roll down from point X to point Y of the incline plane is measured by a stop-watch (Figure 1). The results of the experiment are shown in Table 1.

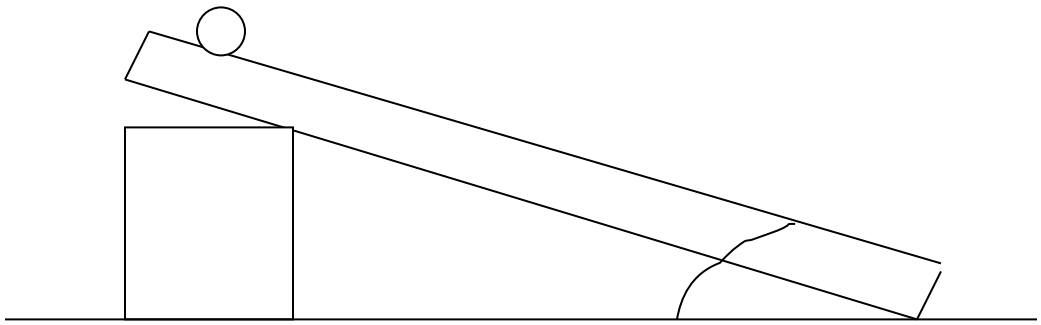

Figure 1

Φ /^{o}	Time for the ball to roll down from point X to Y of incline plane
40	9.30
50	9.26
60	9.20

Table 1

Based on the results of the experiment in Table 1, Albert concluded that the steeper the incline plane, the faster the ball will roll down from point X to point Y of incline plane.

a) Do you agree with his conclusion? Explain why you agree or do not agree

_____

_____

b) If you did not agree, how would you help Albert improve his experiment?

_____

_____