CONCEPTIONS OF SCIENTIFIC EVIDENCE IN INVESTIGATIVE TASKS AMONG TRAINEES IN A TEACHER TRAINING COLLEGE IN KUCHING

by

Dr Tan Ming Tang
Institut Perguruan Batu Lintang,

ABSTRACT

The main aims of this study were to describe science teacher trainees' understanding of scientific evidence regarding measurement reliability and design validity. Three physical science experiments and two scenario tasks were employed in this study. The sample consisted of 87 trainees from a teacher training institute in Kuching. A quantitative methodology involving observation, interview, practical report and scenario was used. The results reveal that at least 87% (76), 36% (31), 25% (22), 64% (56), and 5% (4) of the sample had the right conceptions on repeats, variance, treatment of anomalous result, fair test and external validity aspect respectively in the practical task. Similarly, 44% (38), 75% (65), 26% (23), 41% (36) and 6% (5) of the sample were found to possess the right conceptions on each of the above aspects respectively over both scenarios. As to the cross-protocol comparison of the sample's conceptions of scientific evidence across both practical and scenarios, weak to moderate correlation values (r = .04 to .38) were obtained. This seems to indicate that the scenario task may not be a suitable alternative to the practical task in measuring college students' conceptions of scientific evidence regarding measurement reliability and design validity.

Introduction

Being scientifically literate nowadays means more than just knowing factual science information. The emphasis now is more on the need of students to understand the procedures of scientific inquiry since it is the single most important ingredient necessary in students' investigative work to enable them to understand how scientific knowledge came to be established and be accepted generally as valid.

Roberts and Gott (2004) refer to this understanding of procedures of scientific inquiry as "concepts of evidence" or "scientific evidence". It is 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.

The Statement Of The Problem

In Malaysia, one of the main aims of science education is to develop the potentials of individuals in an overall and integrated manner so as to produce Malaysian citizens who are not only scientifically and technologically literate but also competent in scientific skills (Malaysia, Pusat Perkembangan Kurikulum, 2002). However, owing to a keen emphasis on examination-oriented teaching, 'the teaching and learning of science in some context, has becomes largely teachercentered, thereby ignoring the development and mastery of scientific and thinking skills among students as required by the curriculum' (Sharifah, 2001; p. 42).

Thus there is a need to look into the training of the Malaysian science teachers on whether enough emphasis is placed on teaching and facilitating the use of scientific skills in science laboratory investigations. At present, there is a lack of knowledge about their conceptions and applications of scientific evidence in science investigations. Hence there is a need to assess their 'thinking behind the doing' (Gott and Duggan, 1995; p. 26) of these scientific procedures in investigative tasks. This assessment is of utmost importance because for science teacher trainees to be effective users and future facilitators of investigative tasks in school, they need to possess appropriate conception of scientific evidence in order to be able to apply its vast repertoire of tools in the teaching and learning of science.

The methodologies of observation and interview have been widely used to assess students' performance and understanding of scientific procedures in practical task. As it is too time consuming to assess students' performance and understanding directly, alternative forms have to be found to be used as surrogates. Hence this study also aims to investigate whether a scenario task can be a reliable and valid alternative to practical task in assessing students' understanding of scientific procedures.

Objectives Of The Study

Specifically, this study aimed to:

(i) Describe the science teacher trainees' conceptions of five scientific evidence aspects associated with the measurement reliability and design validity categories that are employed in
  (a) carrying out the novel physical science experimental task and
  (b) reasoning about the unsound hypothetical experimental scenario task.

(ii) Investigate the relationship between science teacher trainees' conceptions of five scientific evidence aspects associated with the measurement reliability and design validity categories in the practical task, with corresponding conceptions for the same aspects in the scenario task.

Understanding The Key Ideas Of Scientific Evidence

Secondary school and college level education should provide science students with an understanding, at an appropriate level, of the scientific account of the natural world and of the processes of scientific inquiry (Black, 1993). Hence, practical laboratory work is widely used as a teaching strategy and is also seen as crucial in developing an understanding of the procedures of scientific inquiry.

To describe this distinct set of conceptions relating to the procedures of scientific inquiry, Gott, Duggan and Roberts (2002) have come up with a list to define these concepts of evidence (Appendix A), which they believe, can be taught and which is a necessary but not sufficient condition for creative problem-solving. As an example, in order to ensure that the test or experiment conducted is fair, an understanding of the importance of isolating only the relevant variables while controlling others is necessary so that the resulting data collected is valid.

From the review of literature, studies had been carried out to probe different aspects of students' conceptions on measurement reliability such as the need for repeats (Schauble, 1996; Varelas, 1997), the treatment of anomalous data (Chinn and Brewer, 1993), and the reliability of data sets (Allie, Buffler, Kaunda, Campbell & Lubben, 1998; Lubben and Millar, 1996). Yet other studies focused on different aspects of students' conceptions on experimental validity such as fair test (Schauble, Klopfer & Raghavan, 1991) and data collection strategies (Strang, 1990).

Findings on science teachers' conceptions of appropriate sampling techniques and statistical significance in analyzing experimental scenarios are found in the studies of Jungwirth and his colleagues (Jungwirth, 1987, 1990; Jungwirth & Dreyfus, 1990, 1992). As for using different modes of assessment to compare students' understanding of procedural knowledge, recent studies (Welford et. al., 1985; Stark, 1999; Gray and Sharp, 2001; Lawrenz et. al., 2001) had found that there was low correlation between practical and written tests of performance.

Methodology

In this study, three physical science experiments were designed by the researcher based on the topic "Force and Motion" to gather data on the science teacher trainees' conceptions of scientific evidence. For conceptions of the measurement reliability category, the three scientific evidence aspects investigated were that of repeated trials, evaluating the trustworthiness of data and treatment of anomalous data. In the design validity category, the conceptions of internal and external validity aspects were probed. The instrument used was a focused 'Level of Conceptions Interview Protocol' (LoCIP), designed by the researcher based on a revised Lubben and Millar's (1996) "Levels of Students' Understanding of the Collection and Evaluation of Empirical Data" Model and the researcher's own "Levels of Students' Understanding of the Design Validity" Model.

Two weeks after the completion of the above practical task, the sample reviewed two hypothetical scenarios with unsound experimental data sets. They were designed from the topic 'Force and Motion' and were used to probe conceptions of similar aspects as above in both the measurement reliability and design validity categories. 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.

Respondents

The sample studied consisted of 87 science teacher trainees from the January 2002 intake in a teacher training college in the Kuching Division. These prospective science teachers had been routinely taught to use certain experimental procedures such as identifying key variables, controlling variables for fair test, doing repeats, devising data table, and drawing graph. They were also taught to write their practical reports by following a standard format: Aim of experiment, Equipment/Materials used, Methodology, Data table, Graph, Conclusion and Precautions taken (Lembaga Peperiksaan, 2002). There were 45 male and 42 female trainees in this selected sample and their mean age was 22.6 years (sd = 2.1). The majority (94.3%) of the respondents in this selected college were art-streamed students, having only taken the General Science paper in their upper secondary school years.

Findings And Discussion

Identification of Conceptions of Five Scientific Evidence Aspects in the Practical Task

The results (Table 1) reveal that the percentages of college students having the correct conceptions of five scientific evidence aspects in the practical task are as follows (in ascending order): rationale of repeats (~87% to 89%), fair test (~64% to 68%), how to evaluate the trustworthiness of data (~36% to 38%), treatment of anomalous data (~25% to 26%) and the external validity aspect (~5% to 9%).

For each science experiment, it was found that from about 87% (76) to 89% (77) of the respondents who possessed the right conception on the rationale of repeats, only about 25% (22) to 26% (23) of them understood the correct way to handle the repeats measured in each of the three experiments conducted. When asked on why they repeated their measurements and how they subsequently handled the repeats, their typical responses are as follows: "By taking one reading only, it may be less accurate" (*Dengan mengambil satu bacaan sahaja, ia mungkin kurang tepat*) and "By finding its average" (*Dengan mencari puratanya*).

To test the college students' understanding on how to evaluate the trustworthiness of their measurements in both the measured and given data, about 36% (31) to 38% (33) of the respondents managed to provide a viable explanation on how they evaluated the trustworthiness of their measurements in each of the three experiments. By looking at the spread of the measurements, a typical response was “All the measurements are almost the same” (*Semua ukuran adalah hampir sama*). They also justified correctly the existence of an anomalous data point in the three experiments by pointing out significant differences in the measurements. A typical answer in response to the question, “From the data, is there any reading which is less believable and if there is, identify it and explain why you think so?” is illustrated as follows:

Table 1. Percentages and Frequency Counts of Sample’s Conceptions of Five Scientific Evidence Aspects in Each Physical Science Experiment.

Conceptions of Scientific Evidence	Physical Science Experiment (N = 87)		
	First	Second	Third
Repeats			
Wrong conceptions	11.5 (10)	11.5 (10)	12.6 (11)
Correct conceptions on rationale of repeats only	63.2 (55)	62.1 (54)	63.2 (55)
Correct conceptions on both rationale of repeats & way to handle repeats	25.3 (22)	26.4 (23)	24.1 (21)
Evaluating trustworthiness of data			
Wrong conceptions	62.1 (54)	64.4 (56)	63.2 (55)
Correct conceptions	37.9 (33)	35.6 (31)	36.8 (32)
Handling anomalous data			
Wrong conceptions	74.7 (65)	73.6 (64)	74.7 (65)
Correct conceptions	25.3 (22)	26.4 (23)	25.3 (22)
Fair test			
Wrong conceptions	33.3 (29)	35.6 (31)	32.2 (28)
Correct conceptions	66.7 (58)	64.4 (56)	67.8 (59)
External validity aspect			
Wrong conceptions	91.0 (79)	93.1 (81)	95.4 (83)
Correct conceptions	9.0 (8)	6.9 (6)	4.6 (4)

"Yes, at the height of 80 cm where the time at t_4, that is 13.88 because this reading t_4 differs quite a lot from the rest of the readings at this height" *(Ya, iaitu pada ketinggian 80 cm dimana bacaan masa pada t_4 iaitu 13.88 kerana catatan bacaan t_4 ini agak jauh berbeza dengan bacaan yang lain pada ketinggian ini)*. (Respondent no.: 55, first experiment)

As to the conception on handling the anomalous data in each of the three experiments (an anomalous data was included in the results of each supplementary question), about 25% (22) to 26% (23) of the respondents identified and handled the anomalous data correctly when asked on whether a measurement which differs appreciably from most of the others can be included in calculating an average. This group of students realized that if the anomaly is included, it will affect the value of the mean calculated. When questioned on whether all the data in the supplementary question can be accepted or not, a typical response from this group of respondents is as follows:

> "Cannot. This is because the calculation of unacceptable data will influence the value of the mean at that measured height (*Tidak boleh. Ini kerana pengiraan data yang tidak boleh diterima akan mempengaruhi purata pada ketinggian yang diukur*). (Respondent no.: 36, first experiment)

In the design validity aspect, about 91% (79) to 95% (83) of the sample failed to have a real understanding of the importance of selecting a suitable range and interval in collecting their data in each of the three experiments conducted. When asked to provide a rationale for their preference of choosing an adequate range of independent variable values and at equal interval in their measurements, about 44% (38) to 47% (41) of the respondents described issues related to both graphical construction and interpretation in each of the three experiments conducted. The following excerpts contain portions of the respondents' rationales:

> To make easier the transfer of data or information onto the graph (*Untuk memudahkan pemindahan data atau maklumat ke dalam graf*). (Respondent no.: 55, first experiment)
>
> To make the construction of graph easier (*Untuk memudahkan membina graf*). (Respondent no.: 53, second experiment)
>
> Because with suitable interval, it will make reading and the drawing of graph easier (*Kerana selang kelas yang sesuai akan memudahkan bacaan dan melukis graf*). (Respondent no.: 57, third experiment)

Hence, nearly half of the sample chose a suitable range and appropriate interval in their measurements in order to make the construction and/or interpretation of the resulting graph drawn easier. Thus, these trainees failed to see the connection between the usage of a suitable range and interval in collecting data and the external validity of experimental design overall.

As to the concept of fair test, about 64% (56) to 68% (59) of the respondents in each experiment understood the necessity of controlling all the relevant variables in order that only the independent variable is allowed to affect the dependent variable. When questioned on whether they can consider their experiments to be a fair test, a typical excerpt of their responses is as illustrated below:

> "Yes, because there is no obstacle for us to do the experiment. Time is only influenced by the height" (*Ya, sebab tiada ada halangan untuk kita melakukan ujikaji. Masa dipengaruhi ketinggian sahaja*). (Respondent no.: 77, second experiment)

Of these respondents, about 59% (51) had the right conception on fair test in all three experiments and another 8% (7) in any two experiments. This shows that these respondents were quite consistent in their conceptions of the fair test.

Identification of Conceptions of Five Scientific Evidence Aspects in the Scenario Task

Through the analysis of the respondents' written artifacts, it was found that out of about 52% (45) of respondents who had the right conceptions on repeats in each of the scenarios, only about 44% (38) of the sample were able to provide a viable explanation on the purpose of repeats in both scenarios.

As to evaluating the reliability of data sets, about 75% (65) of the respondents focused correctly on the spread of the results (variance) over both scenarios. They realized that results that have a smaller range are more trustworthy. Thus, they used consistency as a criterion for judging the reliability of experimental data. This result concurs with the findings of Lubben and Millar (1996) that young adults frequently used consistency in judging the reliability of data sets.

As to handling an anomalous result, only about 26% (23) of the respondents concluded correctly that the anomalous result must be excluded in working out an average across both scenarios. This finding is almost similar to that obtained by Allie et al. (1998) who administered nine pencil and paper probes on 121 undergraduate physics students. About a third (~33%) succeeded in identifying and excluding the anomaly in working out an average.

As to the design validity category, only 41% (36) of the respondents had a good knowledge of what constitute a 'fair test' in both scenarios. This result is consistent with the findings of Renner and Lawson (1973) who had found that many adults too do not possess appropriate conceptions of controlled experimentation.

As for conception on the external validity issue of experimental design, only 6% (5) were able to explain the importance of having a suitable range and interval over both scenarios. In general, the rationale of the majority of the respondents on this external validity aspect was based mainly on a wrong premise, that is, the reason for having a suitable range and interval was to make the construction of graphical representations of the collected data easier. Thus, they failed to see the importance of using appropriate data range and interval in the experimental design and how these aspects are related to the extent in which the experimental results can be generalized.

A Cross-Protocol Comparison of Trainees' Conceptions of Scientific Evidence

By comparing trainees' conceptions of rationale of repeats, evaluating the trustworthiness of data and the external validity aspect in the practical task with their corresponding conceptions in the scenario task, the Pearson's correlation values ($r = .13$ to $.25$) obtained show that each aspect was only weakly correlated overall. Even weaker correlation value ($r\ \ .04$) was obtained for the fair test aspect across both protocols.

The contributing factor for the weak relationships in these scientific evidence aspects between the two protocols may be due to the differences in the format of the questions used in the protocols although the target conceptions measured were the same. In this present study, the formats of the scenario task's questions are similar to those utilized by many researchers (e.g. Lubben & Millar, 1996; Varelas, 1997; Taylor, 2001) undertaking such study but they differ markedly from those of the practical task. In the case of the fair test, only one extraneous factor was required to be controlled (i.e., mass) in the scenario task to ensure fair test whereas in the practical task, the factors to be controlled were many and varied. As to the data set format for repeats in the practical task, the data set collected by each respondent had a few measurements for each value of the independent variable whereas for the scenario task, only one value for each independent variable was given.

However, moderate Pearson correlation value (r = .38) was obtained for the relationship between trainees' conceptions on the treatment of anomalous data across both tasks overall. The slightly higher correlation values obtained here could be due to the fact that the same question format for the anomalous data aspect was being used in both practical and scenario tasks.

Overall, the weak to moderate correlation values obtained in this present study are in line with the findings of other researchers (Solano Flores et al., 1999; Gray and Sharp, 2001) who found that hands-on investigation utilized and tapped a kind of knowledge not addressed by the other forms of tasks. Gray and Sharp (2001) argued that different modes of assessment may be actually measuring different attributes since the nature of the tasks themselves could been altered by their respective modes of assessment. Hence different forms of task may induce in the minds of the trainees slightly different perceptions of what they are suppose to grasp and understand.

Overall, the results reveal that, on the average, more science teacher trainees had appropriate conceptions for both the rationale of repeats and fair test in the practical task than in the written scenario task. But the reverse occurs for conceptions of 'evaluating the trustworthiness of data', 'treatment of anomalous data, and the external validity aspect. The routinization of the scientific evidence aspects of repeats, identifying key variables and controlling relevant variables in the practical task may have enabled more science teacher trainees to grasp the necessary understandings of the related evidence aspects.

Implications and Recommendations of Findings

Conceptions of Scientific Evidence in the Practical Task

The findings seem to suggest that the routinization of scientific evidence aspects did enable more trainees to grasp the appropriate conceptions of the related evidence aspects. In other word, the constant applications of these scientific evidence aspects did increase the strength of the weightings in the memory units, thus aiding more college students in generating appropriate generalizations or inferences about the scientific evidence applied. This finding thus substantiates Rumelhart and McClelland's (1986) connectionistic information processing model which states that learning can occur with gradual changes in connection (memory) strength by experience and an environment within which the system must operate. However, the results also reveal that there are still some college students who have yet to grasp the necessary conceptions of the routinized scientific evidence aspects from their practical experiences in the science laboratory. Thus, more time is needed for these college students to grasp the necessary understandings of their applications of the related scientific evidence aspects or they might not even grasp it at all.

The three critical areas were the sample's failure to grasp the correct conceptions on how to evaluate the trustworthiness of the data, how to handle anomalous data and on the external validity issue in all three experiments. For the first aspect, teacher trainers might want to instruct them on the use of variance and standard deviation to evaluate the trustworthiness of the results. As to the second aspect, the students' attention needs to be drawn to see the difference in the mean values calculated if the anomalous data is both included and excluded in their calculations. In this way, it will help the respondents to determine which mean value is the better representative of the measure of central tendency of the collected data. For the last aspect, teacher trainers can help the trainees by encouraging them to reflect on the consequences of failing to use an appropriate range and interval on the overall experimental generalization drawn. To do this, trainees are encouraged to construct graphs from the provided data sets. Beside encouraging the practice of self-regulation of learning among the trainees, teacher trainers might also want to instruct them on the correct way of constructing line graphs so that the full pattern in a relationship can be explored and displayed.

Conceptions of Scientific Evidence in the Scenario Task

The findings on science teacher trainees' conceptions of scientific evidence show that only about 6% (5) of the respondents were found to be able to explain the importance of having a suitable range and interval for the external validity aspect over both scenarios. By examining their responses, it was found that the majority of them had the misconception that the reason for having a suitable range and interval was to make the construction of graphical representations of the collected data easier. Hence, the above findings will provide diagnostic information to aid teacher trainers in focusing their teaching on specific areas of procedural knowledge in which trainees had misconceptions or difficulties in understanding.

As to conceptions on fair test, repeats and treatment of anomalous data, only about 41% (36), 44% (38) and 26% (23) knew what constitute a fair test, the purpose of repeats and the correct way to handle the anomalous result over both scenarios respectively. Consequently, there is a need for more discussion, problem solving in the laboratory, data analysis task, paper and pencil scenario etc. to be incorporated into the integrated approach to be employed by the teacher trainer to help trainees to grasp the necessary understanding of the underlying conceptions that underpin evidence. Explicit explanatory teaching coupled with more opportunities to practice self-regulation of learning may also be utilized, seeing that procedural understanding is a knowledge domain of science.

A Cross-Protocol Comparison of the Trainees' Conceptions of Scientific Evidence

The weak to moderate correlation values obtained for the cross-protocol comparison between the practical and scenario tasks seem to indicate that the kind of knowledge utilized and tapped by hands-on activities in the practical task may not be the same as that of the scenario task. Thus, the scenario task may not be a suitable alternative to the practical task in measuring college students' conceptions of scientific evidence associated with the measurement reliability and design validity categories.

Hopefully, this finding will serve as an eye-opener to the Malaysian Ministry of Education, which has been emphasizing more on the use of experimental scenarios in Paper 3 of the Form Five's science practical examination than on actual practical assessment. Hence, it is recommended that at this present time, the assessment of procedural understanding should not be based only on a single assessment format. Instead, to be fair to the students, a multiple assessment format should be used for the time being until a single valid assessment format could be found.

Conclusion

The findings of this study show that the routinization of certain scientific evidence aspects enhanced trainees' conceptions of the related scientific evidence aspects. Thus it is recommended that the routinization of laboratory experimental procedures be extended to include the other scientific evidence aspects (e.g. using appropriate accuracy in measurements, identification of an anomalous data etc.) as well. By integrating this routinization process with explicit instruction of the related procedural knowledge through the integrated approach in both practical and scenario tasks at the same time, it is hoped that science teacher trainees' conceptions of scientific evidence in investigative tasks will be enhanced.

REFERENCES

Allie, B., Buffler, A., Kaunda, L., Campbell, B. and Lubben, F. (1998). First-year physics students' perceptions of the quality of experimental measurements. *International Journal of Science Education, 20*, 447-459.

Black, P. (1993). The purposes of science education. In R. Hull (ed.), *ASE Secondary Science Teachers' Handbook* (pp. 6-22). London: Simon and Schuster.

Chinn, C.A. and Brewer, W.F. (1993). The role of anomalous data in knowledge acquisition: Atheoretical framework and implications for science instruction. *Review of Educational Research, 63,* 1-49.

Gott, R. and Duggan, S. (1995). *Investigative work in the science curriculum.* Buckingham: Open University Press.

Gott, R., Duggan, S. and Roberts, R. (2002). *Research into Understanding Scientific Evidence.* Retrieved June 10, 2003 from *http://www.Understanding%20 Scientific%20Evidence.htm*

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.

Jungwirth, E. (1987). Avoidance of logical fallacies – a neglected aspect of science-education and science-teacher education. *Research in Science and Technological Education, 5,* 43-58.

Jungwirth, E. (1990). Science teachers' spontaneous, latent or non-attendance to the validity of conclusions in reported situations. *Research in Science and Technological Education, 8,* 103-115.

Jungwirth, E. and Dreyfus, A. (1990). Identification and acceptance of a posteriori causal assertions invalidated by faulty enquiry methodology: An international study of curricular expectations and reality. In D. Herget (Ed.), *More history and philosophy of science in science teaching* (pp. 202-211). Tallahassee, FL: Florida State University.

Jungwirth, E. and Dreyfus, A. (1992). After this, therefore because of this: One way of jumping to conclusions. *Journal of Biological Education, 26,* 139-142.

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

Lembaga Peperiksaan (2002). *Format Pentaksiran Mata Pelajaran Fizik Mulai SPM 2003: Instrumen Contoh Bagi Kertas 3 – 4531/3.* Kuala Lumpur: Kerajaan Malaysia.

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

Renner, J. and Lawson, A. (1973). Promoting intellectual development through science teaching. *Physics Teacher, 11*, 273-275.

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.

Rumelhart, D.E, & McClelland, J.L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32*, 102-119.

Schauble, L., Klopfer, L.E. and Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28*, 859-882

Sharifah Maimunah Syed Zin. (2001). *Malaysia*. Retrieved August 12, 2003 from h*ttp://www.ibe.unesco.org/National/China/NewChinaPdf/11Malaysia.pdf*

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.

Strang, J. (1990). *Measurement in School Science.* Assessment Matters No. 2. London: SEAC/EMU.

Taylor, J.A. (2001). *Secondary School Physics Teachers' Conceptions of Scientific Evidence: A Case Study*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, St. Louis, MO, Marh 2001.

Varelas, M. (1997). Third and fourth graders' conceptions of repeated trials and best representatives in science experiments. *Journal of Research in Science Teaching, 9,* 853-872.

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

Examples of Concepts of Evidence and their Definitions

(Adapted from Gott, Duggan and Roberts, 2002, pp. 1-12)

Reliability And Validity	*Concepts of Evidence*	*Definition*
Associated with design	**Variable Identification**	The design of an investigation requires variables to be identified and measured. The independent variable is the variable for which values are changed or selected by the investigator whereas the dependent variable is the variable the value of which is measured for each and every change in the independent variable.
	Fair Test	A fair test is one in which only the independent variable has been allowed to affect the dependent variable. Laboratory-based investigations ………. involve the investigator changing the independent variable and keeping all the controlled variables constant.
Associated with Measurement	**Relative Scale**	… the choice of sensible values for quantities is necessary if measurements of the dependent variable are to be meaningful e.g. in differentiating the dissolving times of different chemicals, a large quantity of chemical in a small quantity of water causing saturation will invalidate the results.
	Range and Interval	…….the range over which the values of the independent variable is chosen is important in ensuring that any pattern is detected. …….the choice of interval between values determines whether or not the pattern in the data can be identified.
	Choice of Instrument	Measurements are never entirely accurate for a variety of reasons…..of prime importance is choosing the (right) instrument to give the accuracy and precision required.
	Non-repeatability	….repeated readings of the same quantity with the same instrument never give exactly the same answer. (Due to the inherent variability in any physical measurement, repeats are necessary to give more reliable data).

(continued) *Examples of Concepts of Evidence and their Definitions (Adapted from Gott, Duggan and Roberts, 2002, pp. 1-12).*

Reliability And Validity	Concepts of Evidence	*Definition*
Associated with Measurement (con't)	**Accuracy or trueness**	….trueness is a measure of the extent to which repeated readings of the same quantity give a mean that is the same as the 'true' mean. According to Gott and Duggan (1995), an appropriate degree of accuracy is required to provide reliable data which will allow meaningful interpretation.
Associated With Data Handling	**Tables**	….a table is a means of reporting and displaying data. Simple patterns such as directly proportional or inversely proportional relationship can be shown effectively in a table but it has limited information about the design of an investigation e.g. control variables.
	Anomalous data	…patterns in tables or graphs can show up anomalous data points which require further consideration before excluding them from further consideration (the 'bad' measurement due to human error perhaps)
	Patterns	Patterns can be seen in tables or graphs or can be reported by using the results of appropriate statistical analysis and they represent the behavior of variables.
Associated with the evaluation of the complete task	**Reliability**	….the reliability of the design includes a consideration of all the ideas associated with the measurement of each and every datum. It relates to the question 'Will the measurements result in sufficiently reliable data to answer the question?'
	Validity	….the validity of the design includes a consideration of the reliability and the validity of each and every datum. Beside the above question, another overarching question is 'Will the design result in sufficiently valid data to answer the question?'