Errors: What are they and how significant are they?

John P. Keeves
Trevor G. Johnson
Tilahun M. Afrassa
Flinders University of South Australia  john.keeves@flinders.edu.au

THE NATURE OF ERRORS

Errors in educational research and measurement arise from four main sources:

(a) errors associated with the characteristic being measured – intrinsic errors,
(b) errors arising from the instrument being used – instrumental errors,
(c) errors involved in the act of measurement – observational errors.

In addition, since it is rarely possible to take measurements of a complete population there are
(d) errors arising from the process of sampling – sampling errors.

Sampling errors are of two types. First, there are errors arising from a sample not being fully
representative of the population from which it is drawn-sample bias. Secondly, there are errors
that arise from the variability among the cases included in the sample, which can be estimated from
information on the variability between cases and the number of cases-standard errors of sampling.

Estimates of the standard error of sampling permit inferences to be drawn about the range of a
characteristic in the population.

The word “error” has many meanings. The most common meaning is concerned with the idea of a
“mistake” which does not apply in this context. A further meaning is concerned with “the
difference between an observed or estimated numerical result and the true or exact one”. However,
in educational research the “true value” is both unknown and unknowable and this meaning does
not apply. In statistical work the term “error” simply means “the action of wandering”, since the
observed values are dispersed about a central value and are assumed to be as likely to be greater
than this central value as they are to be less than the central value. This “errant” or “wandering”
nature of observations applies to all four types of error considered above. However, it does not
apply to sample bias. In order to make some allowance for sample bias, prior knowledge is
needed. The making of statistical estimates that are based on prior knowledge lies in the domain of
Bayesian statistics. Bayesian procedures were employed in statistical estimation in the Australian
Studies of School Performance in 1976 and 1981 (Morgan, 1978), but have not been used since in
other Australian studies of student achievement.

The examination of error in educational research, when Bayesian procedures are not employed, is
built around the idea of the importance of findings; namely, pattern of results, size of effect
and statistical significance. This paper is concerned with the examination of errors in several
recent Australian research studies.
IMPORTANCE AND SIGNIFICANCE

Increased access to computers for the processing of data collected in large surveys of educational outcomes has recently raised issues concerned with assessing the importance or significance of findings in educational research. In Australia, although the Curriculum Survey (Radford, 1951) was conducted by ACER in 1947, it was not until the First IEA Mathematics Study (FIMS) was carried out in 1964, that the value of regular surveys of student achievement and attitudes was recognized. Today, not only do the IEA studies continue, but the PISA project conducted by OECD, the Basic Skills Testing Programs which are held annually in each state, and the Course Experience Questionnaire, administered to graduating university students, are in operation, while university entrance and completion tests are under development. In some studies efforts are made to draw large random samples, while in other studies attempts are made to capture all members of a target population. The speed and accuracy with which computers and electronic equipment can now scan documents and process data remove some constraints on the size of the student groups under survey.

The issues that have arisen are largely in the interpretation of the findings and their importance or significance. However, there are the related problems of the unknown bias that arises from a less than adequate level of response both in the proportion of respondents and in the meaningfulness of their responses to the survey instruments, due not only to omission, but also to lack of motivation to respond.

The strength and value of the surveys undertaken across educational systems in Australia today, whether they involve a sample or a census of a defined target population, lies in the opportunity to monitor change over time as well as to provide estimates for comparisons between like groups through replication. Since education is primarily concerned with learning and development, it is the monitoring of change over time that is emerging as the aspect of greater interest. The estimation of level of performance and the attainment of benchmarks or standards are also of some interest, but interpretation of findings suffers from the judgmental nature of the tasks of standard setting and the specification of desired levels of performance. The monitoring of change over time involves not only the system, but also the school and classroom, and above all else the learning achieved by each individual student and the development taking place in each individual child.

It is in the examination of learning and development, that the monitoring of change over time has the potential to make its greatest contribution to education. This paper has a main purpose of addressing certain issues that have arisen in educational research as a consequence of this increased emphasis on the use of large scale surveys of educational outcomes to monitor change over time. These problems involve, in the main, the procedures employed to assess the importance of the findings from surveys and censuses. Unfortunately, the standard texts on educational research methodology are written as course books for undergraduate and postgraduate students in education, psychology and sociology. These texts do not consider large scale surveys, but only small studies that are undertaken in a single school or institution. Consequently, some of the issues that arise are not raised in courses at universities.

Unfortunately too, there is marked controversy between statisticians with respect to the three major approaches to significance testing, namely:

1. the Fisherian test of the null hypothesis;
2. the Neyman-Pearson test for the choice between two alternative hypotheses leading to the making of decisions for policy and practice; and
3. Bayesian tests that take into account prior information before making a decision.

The approach that is increasingly being used in educational research for estimation and significance testing is based on Fisher’s ideas of likelihood. Under this approach three things must be available:
1. the hypothesis or model that specifies how the observed data were generated;
2. the nature of the distribution under which the data were generated; and
3. data that have been collected and that can be used in testing the hypothesis or model.

The question is whether the model or hypothesis is adequate to explain the generation of the observed data, or whether the model or hypothesis must be rejected. The probabilities that are employed in testing are not about the parameters of the model. They are concerned with the data and whether the data could have been generated by the model or the hypothesis. The model can never be said to be true, but it can be provisionally accepted as adequate. The true value of any parameter that is being estimated is unknowable. The question is whether the model fits the data, and if so, the task is to estimate the parameters of the model. The computer programs used today in educational research employ this likelihood approach whether they are concerned with multivariate or multilevel analysis, or with the scaling of data collected with achievement tests or attitude scales.

**SCALING OF DATA**

The scaling of data in most Australian studies employs the Rasch or one-parameter logistic model. The model, based on the conjoint measurement of student performance and item difficulty, employs the logistic transformation and a probabilistic approach. In contrast to the two-parameter and three-parameter models, it provides measures of student performance on a scale that is independent of both the items employed and the persons who responded. Moreover, the scale has a natural metric provided as the unit of measurement - the logit. The only property of the scale that is arbitrary is the fixed point or origin. Thus the scale of measurement has the properties of an interval scale, but not of a ratio scale. The major advantages of the use of the Rasch model, provided that the model fits the data and the requirement of unidimensionality is satisfied, are:
1. the errors involved in measurement of items and persons can be estimated;
2. data sets can be readily equated;
3. differential item functioning for different subgroups of persons can be readily detected; and
4. the use of an interval scale provides measures of change on a consistent metric across the full range of the scale.

**SAMPLE SURVEYS AND CENSUSES**

It is essential to recognize that whether a sample survey or a census is conducted in educational research the units in operation are the student, the classroom, the school or institution and the system. As a consequence students are clustered within classrooms, classrooms are clustered within schools and schools are clustered within systems. Treatments are administered at all four levels of the system, the school, the classroom and the student. The analysis of data in educational research must recognize this nested or clustered nature of the data, whether the study is a sample survey or a census.

If a sample survey is undertaken there is a further complexity in so far as the school is commonly the primary sampling unit, and either intact classrooms or students are sampled from within the
school at a second stage. Random selection occurs at both stages, largely to ensure representativeness for subsequent generalization. Moreover, if random selection has occurred in sampling, it is also possible to make statements about the size of the errors of sampling. Nevertheless, it is rare in educational research to employ a simple random sample, as a consequence most samples employed are complex or clustered samples. Furthermore, it is rare for treatments to be applied solely at the level of the individual student and the complex structure of the data, for both surveys and censuses, must be taken into consideration in the analysis of the data. Since, estimates of error are used in significance testing, failure to consider the complex structure of both the sample and the data gives rise to gross mistakes in much testing for statistical significance. It is necessary to recognize that the estimates of sampling error are commonly faulty because students are nested within classroom groups, and classroom groups are nested within schools and schools are nested within systems. Consequently the simple random sample estimates of error provided by computer programs, with the exceptions of WesVarPC (1997) and multilevel analysis programs such as HLM (Bryk, Raudenbush and Congden, 1996) and MLwiN (Rasbush, Healy, Browne, & Cameron, 1998), are the sources of these gross mistakes.

In the main, significance testing is primarily conducted to identify findings that are considered of importance. In general, it is assumed that if an effect involves statistical significance at the five per cent level, it is of some importance, and if the finding involves statistical significance at the one per cent level it is of greater importance. However, the level of statistical significance is dependent on sample size. Nevertheless, sample size is not the only characteristic to be considered. It is also necessary to take into account the structure of the sample or the census data and sometimes to consider the proportion of the target population tested when the finite population correction is used.

The consequences of the structure of the data must be discussed in addressing issues of importance in the examination of the findings of sample surveys and censuses. It is here that text and reference works on educational research and measurement are seriously at fault.

**SOURCES OF DATA**

In the illustration of some of the consequences of failure to address in appropriate ways some of the issues that arise in the examination of findings in educational research, data have been drawn from several sources. First, there are analyses of data collected with the Course Experience Questionnaire from graduating students. Secondly, there are analyses of data collected in the Basic Skills Testing Program in South Australia in 1995 and 1999. Thirdly, there are data collected in the First, Second and Third IEA Mathematics Studies in Australia in 1964, 1978 and 1994 respectively. Work has been done on the analyses of these data sets by staff and postgraduate students at Flinders University. We are grateful to have access to these analyses to provide illustrations of effects that are of interest. These effects arise in attempts to assess the importance of findings. This discussion is presented not to draw attention to gross mistakes, but to stimulate discussion of the issues that have arisen in educational research and measurement in Australia.

**ISSUES OF SIGNIFICANCE**

Statisticians have promoted a concern for statistical significance as the over-riding indicator of the importance of findings. They have largely ignored consideration of the size of an effect, except in power calculations, and the estimation of appropriate sample size, and have underemphasized the usefulness of the pattern of results, which is involved both in replication and in the monitoring of change over time. All three aspects of the importance of findings, namely, statistical significance, size of effect, and pattern of results must be considered.
PATTERN OF RESULTS

Interest in the pattern of results employs two approaches. First, there is the replication of effects in comparison with other similar groups under survey. Thus there is interest in the Basic Skills Testing Program for comparisons within a groups of like schools. Alternatively, there is interest in comparisons within a school of the performance of similar classroom groups, since the classroom is the unit of instruction. There is also interest in comparisons between schools systems, as exist in each of the Australian states, or between countries as occurs in the IEA or PISA Projects. Secondly, there is interest in the changes that occur over time or the monitoring of performance at the student, school or system levels.

In its simplest form the pattern of results can be assessed by simple comparisons of greater than, equal to, or less than and the counting of pluses and minuses. However, such comparisons lend themselves to statistical analysis or meta analysis, that can now be readily carried out using multilevel analysis procedures. Likewise, the examination of change over time, provided three or more measures of an outcome are available for between time comparisons, can also be readily undertaken using multilevel analysis, if change is modelled as a simple linear or quadratic function of time. Alternatively, if a large enough number of time points is available, event history analysis can be employed. These procedures of analysis are, however, dependent on both issues concerned with the size of an effect that is considered to be of importance and with the statistical significance tests that are applied at many stages in the calculations. Thus the pattern of results of itself is not the sole test to be employed in the assessment of the importance of findings. It is here that the Bayesian approach to statistical theory is likely to come into greater prominence in the years ahead.

Some Illustrations of Pattern of Results

Examination of change over time

As a first example in the examination of change over time, data collected in the First, Second and Third IEA Mathematics Studies in 1964, 1978 and 1994 respectively are employed to illustrate trends over time (Afrassa and Keeves, 1997). The FIMS and SIMS data provided estimates of mathematics achievement for 13-year-old students for government schools for five of the eight school systems, while the FIMS and TIMS data provided estimates of mathematics achievement for Year 8 students in the same five government school systems. Performance in mathematics is recorded on a Rasch scale, using a metric of centilogits, and with the fixed point set at 500, which was the difficulty level of the 1964 FIMS mathematics tests (A, B and C). Figure 1 records the estimates of the group means and their standard errors in parentheses. The pattern of the findings is clear and consistent across occasions. However, the issue arises as to whether the declines in performance over time are of consequence or importance. Moreover, since five state school systems were involved it is of interest to consider whether the same pattern is observed for Australia overall as is detected for each of the five systems. Figure 2 presents the findings for each school system for the comparisons between FIMS (1964) and SIMS (1978). While Figure 3 presents the findings for the comparisons between FIMS (1964) and TIMS (1994). The larger standard errors for TIMS compared with FIMS and SIMS should be noted in spite of the greatly increased sample sizes. The issue that must be addressed is whether these effects are of importance both in magnitude and with respect to statistical significance. It is also of interest to consider whether the large standard errors for TIMS greatly reduce the probability of detecting a significant difference. It should be noted that in TIMS intact classes were sampled at the second stage after schools were sampled at the first stage while in FIMS and SIMS, students within schools were sampled at the second stage. This difference in sample design in TIMS gives rise to
rather larger design effects. It should be noted that one school system recorded an apparent rise in mathematics achievement over the 30 year period while one school system recorded no change, and the other three systems provided evidence of apparent declines.

**Figure 1. Comparison of Achievement in Mathematics between 1964, 1978 and 1994 in Australia**

Fixed point = Mean difficulty level of FIMS test; 100 units = 1 logit; 1 unit = 1 centilogit
D = difference in mathematics achievement between occasions
Values in parentheses are Rasch estimated mean scores and standard errors of the means respectively
$D_{1978-1964} = 19$ centilogits or half year of mathematics learning
$D_{1994-1964} = 31$ centilogits or 0.8 year of mathematics learning
One year of mathematics learning is estimated to be 37 centilogits.

**Figure 2. Comparison of Achievement in Mathematics between 1964 and 1978 in Australia for 13-year-old students**

Fixed point = Mean difficulty level of FIMS test; 100 units = 1 logit; 1 unit = 1 centilogit
Values in brackets are Rasch estimated mean scores and standard errors of the mean respectively
A = State A, B = State B, C = State C, D = State D, E = State E.
One year of mathematics learning is estimated to be 37 centilogits.
Errors: What are they and how significant are they?

Figure 3. Comparison of Achievement in Mathematics between 1964 and 1994 in Australia for Grade 8 students

Size of Effects

The apparent declines recorded in Figure 1 raise the question as to whether changes in mathematics achievement of 0.19 and 0.31 logits are of sufficient magnitude to be of practical significance.

There are three ways in which the size of an effect is estimated.

Standardized mean difference

Many studies have estimated correlations between a criterion measure and the performance of two groups and the correlation coefficient is used to summarize the findings of the study, since larger correlations indicate stronger relationships. Thus a correlation coefficient can be readily converted to a standardized mean difference in which the difference between the two group means is divided by the common standard deviation and the mean difference is expressed in standard deviation units. This standardized mean difference is generally known as an effect size. Cohen (1988) has provided three rules of thumb for assessing the magnitude of an effect size:

- Small: 0.20 standard deviations, or a correlation of 0.10, or the difference between 50 and 45 per cent;
- Medium: 0.50 standard deviations, or a correlation of 0.30, or the difference between 50 and 35 per cent;
- Large: 0.80 standard deviations, or a correlation of 0.50, or the difference between 50 and 25 per cent.

Table 1 records the estimated effect sizes for differences in mathematics achievement for FIMS, SIMS and TIMS that are shown in Figures 1, 2 and 3.
Table 1. Effect Size Difference in Mathematics Achievement for FIMS (1964), SIMS (1978) and TIMS (1994)

<table>
<thead>
<tr>
<th>State</th>
<th>$D_{1978-1964}$</th>
<th>$D_{1994-1964}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13-year-old-students</td>
<td>Grade 8 students</td>
</tr>
<tr>
<td>State A</td>
<td>-0.16</td>
<td>-0.02</td>
</tr>
<tr>
<td>State B</td>
<td>-0.11</td>
<td>-0.83 (L)</td>
</tr>
<tr>
<td>State C</td>
<td>0.05</td>
<td>-0.28 (S)</td>
</tr>
<tr>
<td>State D</td>
<td>-0.36 (S)</td>
<td>-0.59 (M)</td>
</tr>
<tr>
<td>State E</td>
<td>0.00</td>
<td>0.14</td>
</tr>
<tr>
<td>Australia</td>
<td>-0.19</td>
<td>-0.29 (S)</td>
</tr>
</tbody>
</table>

Recorded in parenthesis are the assessed magnitudes of effect sizes:
- S – Small; M – medium; L – Large
- All scores are recorded on the TIMS standardized scales for mathematics

**Standardized scale difference**

The tradition at the Educational Testing Service in the United States has been to construct a standardized scale of achievement in which either 50 units (for National Assessment of Educational Progress) or 100 units (for SAT scores) correspond, at least initially when the scale was first calibrated, to a pooled student standard deviation for a somewhat arbitrary sample of students. Table 2 records the mean scores in mathematics achievement for the eight Australian school systems for Year 8 level, Level 1, Level 2 and the difference between Level 1 and Level 2 which indicates the gain in scores between adjacent grades obtained for the TIMS study in Australia using standardized scale differences on a 100 unit scale (Lokan, Ford and Greenwood, 1996), that is similar to the scales constructed by the educational testing service.

Table 2. Estimated gains in mathematics achievement across adjacent grades in TIMS study using standardized scale values

<table>
<thead>
<tr>
<th>Mean mathematics achievement score measured in centilogits</th>
<th>Age In years</th>
<th>Year 8</th>
<th>Adjacent year Levels</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Level 1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Level 2&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>New South Wales</td>
<td>14.1</td>
<td>522 ± 9</td>
<td>495 ± 9</td>
<td>522 ± 9</td>
</tr>
<tr>
<td>Victoria</td>
<td>14.0</td>
<td>511 ± 7</td>
<td>475 ± 8</td>
<td>511 ± 7</td>
</tr>
<tr>
<td>Queensland</td>
<td>13.5</td>
<td>512 ± 8</td>
<td>512 ± 8</td>
<td>512 ± 8</td>
</tr>
<tr>
<td>Western Australia</td>
<td>13.5</td>
<td>532 ± 8</td>
<td>532 ± 8</td>
<td>561 ± 11</td>
</tr>
<tr>
<td>South Australia</td>
<td>13.8</td>
<td>516 ± 6</td>
<td>516 ± 6</td>
<td>557 ± 6</td>
</tr>
<tr>
<td>Tasmania</td>
<td>14.0</td>
<td>501 ± 12</td>
<td>473 ± 12</td>
<td>501 ± 12</td>
</tr>
<tr>
<td>Australian Capital Territory</td>
<td>14.1</td>
<td>548 ± 12</td>
<td>540 ± 14</td>
<td>548 ± 12</td>
</tr>
<tr>
<td>Northern Territory</td>
<td>13.8</td>
<td>478 ± 20</td>
<td>478 ± 20</td>
<td>494 ± 14</td>
</tr>
<tr>
<td>Australia</td>
<td>-</td>
<td>-</td>
<td>498 ± 3.8</td>
<td>530 ± 4.0</td>
</tr>
</tbody>
</table>

All scores are recorded on the TIMS standardized scales for mathematics

- <sup>a</sup> Year 7 in NSW, Vic, Tas and ACT, and Year 8 in Qld, WA, SA and NT
- <sup>b</sup> Year 8 in NSW, Vic, Tas and ACT and Year 9 in Qld, WA, SA, and NT

Scale units in centilogits; Standard errors of mean values are recorded alongside estimated mean values

The gain in mathematics achievement for a year of schooling across Australia is estimated to be 0.32 units on a rather arbitrary international scale (Lokan, et al. 1996). It should be noted that it is only meaningful to compare performances between states at the Year 8 level, and even here the differences in ages between the Year 8 groups in the different states should be noted. However, it would also be possible to report an age-adjusted score, since the growth during a year of schooling across Australia has been estimated. It is clearly not meaningful to make comparisons between states within Level 1 or Level 2, because of the very different periods of schooling involved.
Logit scale differences

The natural unit of the logistic scale employed in Rasch measurement procedures is the logit, and it is now common in the construction of Rasch scaled scores to employ a scale mean of 500 and to use the centilogit as the scale metric (Keeves and Schleicher, 1992). Work on the 1995 Basic Skills Test data in South Australia to examine change in scaled scores between Year 3 and 5 in both Literacy and Numeracy yielded a surprisingly simple relationship that the growth over a two year period was almost exactly 1.00 logits or 100 centilogits. It is estimated that the growth across Year 3 and Year 5 in both Literacy and Numeracy is 50 centilogits. Keeves and Schleicher (1992) have estimated that the growth in Australia from Age 10 to Age 14 in Science is 135 centilogits and the growth per year in this age span can be estimated as 34 centilogits. Schleicher (1994) has estimated in several different ways that the approximate growth in reading achievement from Age 9 to Age 13 is 21 centilogits. Moreover, Afrassa (1998) has estimated that the growth in mathematics achievement between one grade level and the next in Australia in the TIMS study is 37 centilogits, which is consistent with the estimate recorded in Table 1 on the TIMS scale, since the standard deviation employed was estimated to be approximately 120 logits. Table 3 summarizes in scale units the learning in one year obtained from different studies.

Table 3. Growth for learning in one year expressed in centilogits

<table>
<thead>
<tr>
<th>Field</th>
<th>Year</th>
<th>Grade</th>
<th>Country</th>
<th>Growth/year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Literacy</td>
<td>1995</td>
<td>Year 3 to 5</td>
<td>South Australia</td>
<td>0.50</td>
</tr>
<tr>
<td>Numeracy</td>
<td>1995</td>
<td>Year 3 to 5</td>
<td>South Australia</td>
<td>0.50</td>
</tr>
<tr>
<td>Science</td>
<td>1984</td>
<td>Ages 10 to 14</td>
<td>Australia</td>
<td>0.34</td>
</tr>
<tr>
<td>Mathematics</td>
<td>1994</td>
<td>Grades 7 to 8</td>
<td>Australia</td>
<td>0.37</td>
</tr>
<tr>
<td>Reading Literacy</td>
<td>1990</td>
<td>Ages 9 to 13</td>
<td>All countries</td>
<td>0.21</td>
</tr>
</tbody>
</table>

It is evident that using the Rasch scale it is possible to interpret the differences recorded in Figure 1 for the decline (19 units) in mathematics achievement across Australian between FIMS (1964) and SIMS (1978) in terms of half a year (19/37 units) of student learning, and the decline (31 units) in mathematics achievement across Australia between FIMS (1964) and TIMS (1994) as approximately four-fifths (31/37) of a year of learning of mathematics. It should be noted that in one school system between 1964 and 1994 the gain (13 units) is estimated to be approximately one third (13/37 units) of a year of mathematics learning. The interpretation of change in achievement in terms of years and/or months of school learning would seem to add meaning to the estimates made when the Rasch scale is employed. Nevertheless, it is also of some concern to estimate in terms of years of school learning the magnitudes of the standard errors that are recorded in Table 2 for system mean values for the TIMS study. Moreover, the estimated magnitudes of growth recorded in Table 2 for the eight school systems are so different as to suggest serious bias in some of the mean values recorded in this table.

STATISTICAL SIGNIFICANCE

The initial problem to be considered in the examination of data collected in the Basic Skills Testing Program and through the Course Experience Questionnaire, where an attempt is made to provide a study in which all members of the target population are involved, is whether a sample survey or a census has been undertaken. Inevitably, there are losses at both the institution and student within institution levels. An initial question must be asked is whether the losses introduce bias, and how the extent of bias could be assessed. Further questions must be addressed.
1. If a sample survey is considered to be a meaningful description, with the initial target population being the year group population to which generalization is made, should a finite population correction be made to the estimates of error?

2. If a sample survey is considered to be a meaningful description, is the target population the successive year groups from which one year group forms the sample, and is a finite population correction inappropriate?

3. If a census is considered to be a meaningful description can a standard error of the mean value be meaningfully calculated for the census data?

4. In the calculation of the standard error of a mean value, how is the complex structure of the data collected taken into consideration, since it involves institutions, classrooms, and students at successive levels of analysis? This latter question must apply irrespective of whether a sample survey or a census is considered to be a meaningful description, and if an estimate of the standard error of a mean value is to be calculated.

5. How should the standard errors for comparisons between schools within one system and within one year group be calculated?

6. How should the standard errors for comparisons for one school across several years be calculated?

It is clear that many computer packages completely ignore the problems raised, as do most standard text and reference books. Yet the problems remain, because it would seem desirable to make some estimate of the error associated with a mean value, even if statistical significance testing were abandoned.

Successive attempts to calculate standard errors of the mean values for complex samples have involved:

(a) the use of four subsamples (see Husen, 1967; Keeves, 1966);

(b) the use of jackknifing with ten subsamples (see Peaker, 1975; Ross, 1978);

(c) the use of jackknifing deleting one primary sampling unit at a time (see Rust, 1985; Ross, 1991)

Westat in the United States has released a computer package WesVar PC (Brick, et al., 1997) that builds on the jackknifing procedure advanced by Rust (1985) in order to compare mean values, and percentages, as well as to calculate sampling errors in regression analysis for complex sample designs. The WesVar PC program considers each institution as a primary sampling unit with students selected from within each primary sampling unit at a second stage. It employs a procedure involving the dropping of one primary sampling unit at a time and then estimates the parameter for the truncated sample. This step is repeated (n-1) times, where n is the number of primary sampling units, and the (n-1) estimates of the parameter are used to calculate both the jackknife mean and a jackknife standard deviation of that mean. The deviation of the jackknife mean from the full sample estimate of the mean is considered to be an index of bias, and the jackknife standard deviation is considered to be an estimate of the standard error of the mean (see Ross and Rust, 1997).

This approach regards the sample as a subpopulation and jackknifing involves successive samples drawn from that subpopulation to provide an estimate of the error involved in calculating the mean of the subpopulation. The treatment of the estimation of error in this way would seem to avoid
many of the problems that are involved in the use of both formula estimates of standard errors and
the logical problems associated with the relationships between the sample and the target
population. However, all that is produced is an estimate of a parameter and an estimate of the
standard error of that parameter from which a confidence interval can be specified. The WesVar
PC program provides for the statistical comparison of subgroups with appropriate significance
tests. Comparisons between separate subgroups must be carried out through further analyses.

Table 4 presents a comparison between WesVar PC estimates of error and the SPSS estimates of
error for the six Course Experience Questionnaire scales. The table records estimates of the design
effect (deff) which is an indicator of the effect of the complex sample design on the estimation of
error for the mean value of each scale.

<table>
<thead>
<tr>
<th>CEQ Scale/Index</th>
<th>SPSS Mean</th>
<th>SE_{SRS}</th>
<th>WesVar PC Mean</th>
<th>SE_{C}</th>
<th>N</th>
<th>Deff</th>
<th>Deft</th>
<th>SE_{C}/SE_{SRS}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Teaching</td>
<td>11.12</td>
<td>0.165</td>
<td>11.12</td>
<td>0.988</td>
<td>6,009</td>
<td>33.72</td>
<td>5.98</td>
<td>5.98</td>
</tr>
<tr>
<td>Clear Goals &amp; Standards</td>
<td>19.51</td>
<td>0.157</td>
<td>19.51</td>
<td>0.652</td>
<td>6,009</td>
<td>17.18</td>
<td>4.14</td>
<td>4.14</td>
</tr>
<tr>
<td>Appropriate Workload</td>
<td>3.49</td>
<td>0.155</td>
<td>3.49</td>
<td>0.796</td>
<td>6,009</td>
<td>26.29</td>
<td>5.13</td>
<td>5.13</td>
</tr>
<tr>
<td>Appropriate Assessment</td>
<td>28.75</td>
<td>0.177</td>
<td>28.75</td>
<td>0.888</td>
<td>6,009</td>
<td>25.16</td>
<td>5.02</td>
<td>5.02</td>
</tr>
<tr>
<td>General Skills</td>
<td>33.72</td>
<td>0.137</td>
<td>33.72</td>
<td>0.559</td>
<td>6,009</td>
<td>16.59</td>
<td>4.07</td>
<td>4.07</td>
</tr>
<tr>
<td>Overall Satisfaction Index</td>
<td>36.65</td>
<td>0.198</td>
<td>36.65</td>
<td>0.995</td>
<td>6,009</td>
<td>25.38</td>
<td>5.04</td>
<td>5.04</td>
</tr>
</tbody>
</table>

\[ \text{deff} = \frac{\text{Variance of complex sample estimate}}{\text{Variance of simple random sample of same size}} = \frac{V_c}{V_{SRS}} \]

The value of \( \sqrt{\text{deff}} \) or \( \text{deft} \) is the factor by which sampling errors calculated using simple random
sample formulae must be multiplied in order to obtain estimates that reflect the clustering effects
of students within institutions. The far right hand column in Table 4 shows that \( \text{deft} \) as the square
root of \( \text{deff} \) is the same as the standard error for the complex sample divided by the standard error
of the simple random sample of the same size, given in columns 4 and 2 of Table 4 respectively,
when the finite population correction is not used. When the finite population correction is used,
deff is somewhat smaller than \( \sqrt{V_c/V_{SRS}} \).

**Multilevel analysis programs**

Several multilevel analyses programs are now available that take into account the nesting of
students within institutions and, as a consequence, generate more appropriate estimates of
standard errors than do traditional computing programs. These multilevel analysis programs include:

(a) HLM 4.01 (Bryk, Raudenbush and Congdon, 1996)

(b) MlwiN (Rasbash, Healy, Browne, and Cameron, 1998).

It should be noted that HLM is relatively easy to use and the program reads data from a number
of sources such as SPSS and SAS. WesVar PC is specially designed for the analysis of data in
survey research and not for multilevel analysis.
The data for the six scales of the Course Experience Questionnaire (CEQ) were analysed using both MLwiN and HLM4.01 to demonstrate that the estimates were comparable. In addition, it was possible to compare the estimates of the mean values and their standard errors for these two multilevel programs with those obtained by WesVar PC. Table 5 compares the estimates for MLwiN1.02 and HLM4.01 for the fully unconditional model that is equivalent to a one-way ANOVA in the analysis of the CEQ data.

Table 5. Comparisons for Means and Standard Errors of CEQ scores between MLwiN and HLM

<table>
<thead>
<tr>
<th>Scale/Index</th>
<th>MLwiN1.02 Estimates</th>
<th>HLM4.01 Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>Good Teaching</td>
<td>12.77</td>
<td>1.02</td>
</tr>
<tr>
<td>Clear Goals &amp; Standards</td>
<td>20.09</td>
<td>0.61</td>
</tr>
<tr>
<td>Appropriate Workload</td>
<td>4.24</td>
<td>0.76</td>
</tr>
<tr>
<td>Appropriate Assessment</td>
<td>29.49</td>
<td>0.95</td>
</tr>
<tr>
<td>General Skills</td>
<td>34.27</td>
<td>0.51</td>
</tr>
<tr>
<td>Overall Satisfaction Index</td>
<td>38.01</td>
<td>0.89</td>
</tr>
</tbody>
</table>

ρ(rho) is the intraclass correlation coefficient.

The grand means for the scales differ noticeably for MLwiN and HLM from those obtained from SPSS and WesVar PC. One reason for this is that the single level and multilevel estimates are for slightly different data sets. Cases containing missing data at Level 2 (the institutional level) can not be included in the multilevel analyses. It should also be noted that the multilevel estimates of the standard errors of the mean values recorded in Table 5 are of similar magnitude to the WesVar PC estimates recorded in Table 4. Consequently, these findings further reinforce the view that traditional methods recorded from SPSS in Table 4 seriously underestimate the standard errors of clustered or nested data. It follows that the confidence intervals calculated by traditional procedures are likely to be too narrow and erroneous conclusions are likely to be recorded for the statistical significance of observed differences. In addition, it should be noted that these differences arise even where the intraclass correlation coefficient which is also an index of the clustered nature of the sample is seemingly small.

The intraclass correlation coefficient is defined as:

\[ \rho = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_w^2} \]

Where

- \( \sigma_g^2 \) = variance associated with groups,
- \( \sigma_w^2 \) = error variance within groups,
- \( \sigma_g^2 + \sigma_w^2 = \text{total variance.} \)

In making comparisons between institutions these substantial differences between the standard errors calculated for a sample of complex design must be taken into consideration. Over 40 years ago Kish (1957) discussed the consequences of applying the usual standard error formulae found in text books to data obtained from complex samples and concluded that:

In the social sciences the use of SRS (simple random sample) formulas on data from complex samples is now the most frequent source of gross mistakes in the construction of confidence statements and tests of hypotheses. (Kish, 1957, p. 157)

Nevertheless, 40 years later these gross mistakes are still wide spread. The same problem arises within schools where classrooms are the units of instruction and all members of the school are
Errors: What are they and how significant are they?

tested. The students can not be considered to be independent of one another and there is 
commonly marked clustering of students within class groups. The calculation of the standard 
errors of estimate for the mean value of a school must be obtained either by WesVar PC or by 
multilevel analysis with students nested within classrooms. Failure to recognize this problem in 
data analysis must lead to seriously erroneous conclusions about performance at the school level. 
Where students are randomly assigned to classes at the beginning of a school year the clustering 
effects might be thought to be small. However, teachers do have an effect during a school year and 
thus treatment conditions vary and must be taken into consideration in the estimation of errors. 
The effects of streaming or setting into class groups significantly accentuate this problem.

This discussion of issues of importance and statistical significance provides the necessary basis 
for consideration of errors of measurement and their effects on the estimation of student 
performance in such studies as the Basic Skills Tests and the TIMS and PISA testing programs.

ERRORS OF MEASUREMENT

For 20 years the ACER has gradually advanced the use of Rasch measurement in school systems 
across Australia, in spite of marked opposition led by academics from the University of London 
and some statisticians in the United States. The major advantages of the use of Rasch 
measurement procedures are, first, that provided a test can be considered to be unidimensional, 
and the items contained within a test satisfy the requirement of unidimensionality, then estimates 
of performance on that scale are independent of the items employed. Furthermore, the estimates 
of performance are also independent of the persons employed in the calibration of the scale. As a 
consequence of these properties it is readily possible for tests employed at different grade levels 
and on different occasions to be equated on a single common scale.

Secondly, it is possible to make estimates of the errors of measurement for both different items 
and different persons and different points on the scale of measurement.

ERRORS IN RASCH SCALED SCORES

Table 6 records the errors of measurement for persons on the 1999 Basic Skills Tests of Literacy 
and Numeracy at selected score levels on the scales of measurement developed which are 
expressed in logits.

The score values and their standard errors show different standard errors at different levels of the 
scale. However, the standard errors can be interpreted in terms of a year of learning that is 
estimated to be 0.50 logits. Thus the smallest standard error of a Rasch scaled score in the Basic 
Skills Tests recorded for Year 5 Literacy is estimated to be as large as half a year of learning, while 
the largest standard errors are equivalent to two years of learning. These estimates were made 
using the QUEST computer program (Adams and Khoo, 1993).

The standard deviation of the Rasch scaled scores is found to be approximately one logit. This 
indicates that all standard errors have at least a small effect size lying in the range from 0.20 to 
0.50, and the standard errors in the outer categories have a medium effect size being greater than 
0.50 and less than 0.80, while the extreme scores have standard errors of effect size in the large 
category being greater than 0.80. The magnitudes of these standard errors are surely unacceptably 
large, when assessed in this way.

It must be asked whether the use of CONQUEST (Adams, Wilson and Wu, 1998) would yield 
oticeably smaller standard errors, perhaps down to the size of 0.10 logits. It must also be asked
whether the procedures employed for calculating standard errors are really appropriate. Further
questions readily come to mind.

1. Do the procedures for estimating standard errors take into consideration the fact that
individual items are nested within blocks with a common stem?

2. What are the consequences of relatively large standard errors in the estimated scores on the
Basic Skills Tests for the classification of the performance of students in skill bands?

3. Is performance on the subscales of Number, Measurement and Space of sufficient accuracy to
be meaningful, without the use of Bayesian procedures?

4. How does the estimated accuracy of the raw scores compare with the estimated accuracy of
the Rasch scaled scores?

5. Should the Basic Skills Tests be focussed at the ability level of the student with the same
number of items, but differing for each student so that the items are spread over a narrow
band. With more focussed items, a more accurate estimation of student performance could be
made. Computer adaptive testing might provide this gain in accuracy?

**Table 6. Score values and their standard errors in Basic Skills Tests in 1999 in South
Australia.**

<table>
<thead>
<tr>
<th>Year 3</th>
<th>Literacy</th>
<th>Numeracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw Score</td>
<td>Scaled score</td>
</tr>
<tr>
<td>Top of Scale</td>
<td>62</td>
<td>4.77</td>
</tr>
<tr>
<td>Band 5</td>
<td>57</td>
<td>2.73</td>
</tr>
<tr>
<td>Band 4</td>
<td>48</td>
<td>1.42</td>
</tr>
<tr>
<td>Band 3</td>
<td>39</td>
<td>0.57</td>
</tr>
<tr>
<td>Band 2</td>
<td>29</td>
<td>-0.27</td>
</tr>
<tr>
<td>Band 1</td>
<td>13</td>
<td>-1.78</td>
</tr>
<tr>
<td>Band 0</td>
<td>2</td>
<td>-4.14</td>
</tr>
<tr>
<td>Bottom of scale</td>
<td>1</td>
<td>-4.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year 5</th>
<th>Literacy</th>
<th>Numeracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw Score</td>
<td>Scaled score</td>
</tr>
<tr>
<td>Top of Scale</td>
<td>83</td>
<td>6.06</td>
</tr>
<tr>
<td>Band 6</td>
<td>77</td>
<td>3.74</td>
</tr>
<tr>
<td>Band 5</td>
<td>65</td>
<td>2.22</td>
</tr>
<tr>
<td>Band 4</td>
<td>55</td>
<td>1.40</td>
</tr>
<tr>
<td>Band 3</td>
<td>42</td>
<td>0.50</td>
</tr>
<tr>
<td>Band 2</td>
<td>30</td>
<td>-0.34</td>
</tr>
<tr>
<td>Band 1</td>
<td>13</td>
<td>-1.84</td>
</tr>
<tr>
<td>Band 0</td>
<td>2</td>
<td>-4.18</td>
</tr>
<tr>
<td>Bottom of scale</td>
<td>1</td>
<td>-4.92</td>
</tr>
</tbody>
</table>

- a - Perfect scores and zero scores omitted
- b - Scores recorded for midpoint of band
- c - Scaled scores and standard errors expressed in logits

It should be acknowledged that these measurement procedures using Rasch scaling were originally
developed for large scale survey testing programs, where the errors of measurement were small
because of the large sample sizes, and the sampling errors were of greater concern. However, today
these measurement procedures are being used for the estimation of individual student performance,
where the errors of measurement are large and the sampling errors are irrelevant.
**EQUATING ERRORS**

The tests administered each year are converted to measures on a standard scale that was constructed several years earlier. Each year the test for that year is equated with tests employed during earlier years and with the scale developed from those tests. Consideration needs to be given to the estimation of the equating errors introduced. The magnitude of the equating errors will depend on the equating procedure employed, whether it involves concurrent equating, or anchor item equating, or common item difference equating, and the sizes of the errors associated with the calibration of the separate tests being equated. It would seem that the magnitude of the errors of equating should involve both the numbers of common items and the numbers of common persons involved in the test equating operation, as well as the complexity of structure of both the data concerned with the items and persons.

**CALIBRATION ERRORS**

The test of fit to the data recorded for a particular item in order to assess whether the item satisfies the requirement of unidimensionality is a relative test and not an absolute one. Moreover, a balance must be achieved between the bandwidth of a range of items, each assessing slightly different aspects of the characteristics under examination and the fidelity of the items with respect to the unidimensional scale (see Cronbach, 1960). As a consequence, it is sometimes necessary to eliminate from a test those items that have apparently high fidelity but assess over too narrow a bandwidth, and thus supply redundant information, as well as items that have too low fidelity and are spread over too wide a bandwidth. Different computer programs for Rasch scaling use different indices for the testing of item fit, that give very different results for items of marginal fit. Moreover, the effects of sample structure on indices of item fit need to be further investigated, since only one study has been carried out in Australia (Farish, 1984).

Research and debate would appear to be urgently needed into such issues in calibration as:

1. treatment of omitted responses;
2. treatment of not-reached responses;
3. effects of guessing where only two or three alternatives are provided;
4. treatment of misfitting persons in calibration;
5. magnitude of fit statistics for identifying misfitting persons;
6. magnitude of fit statistics for identifying misfitting items;
7. interpretation of partial credit parameters;
8. inclusion or exclusion of items which are biased for particular subgroups, and
9. minimum number of items in a subscale for effective calibration.

These sources of error may be regarded, in the main, as instrumental errors. However, the issue of bandwidth also involves intrinsic errors, in so far as the particular characteristic under survey has many different manifestations. While the scale formed by Rasch measurement procedures is considered to be independent of the items employed, it would seem necessary that a sufficient range of manifestations of a characteristic should be assessed, and without the use of items that merely supply redundant information.

The use of multiple choice test items and constructed response items, with carefully specified guidelines for scoring, serve to reduce the effects of observational errors. However, the move towards embedded assessment, in which teachers embed their assessment procedures within their daily teaching in a classroom, introduces substantial observational error into the measurement process. Research is urgently needed into the nature and magnitude of these observational errors. Nevertheless, it is frequently argued that multiple choice and constructed response test items
reduce the bandwidth of the aspects of performance or the characteristic being assessed and subsequently the range of skills on which instruction is provided. The balance between bandwidth and fidelity must be given greater consideration as the use of Rasch scaling procedures is more widely employed in the assessment of student performance.

CONCLUSION

While this paper has addressed issues concerned with errors of measurement and sampling in testing programs, it is necessary to recognize the marked contribution that advances in educational measurement have made and are making to education, particularly in Australian schools. The emphasis has turned during recent decades from innovation and the liberation of teachers from the dominating effects of public examinations to the advancement of student learning and development in which growth in student performance is recorded on profiles each with a scale of performance. These scales of performance not only relate to the areas of the curriculum and the strands within those areas, but also the instruction and testing provided across the grades of schooling and the learning outcomes attained by individual students. Underlying these scales of performance is the principle of strong and meaningful measurement. As a consequence it is becoming possible for teachers and school principals to report to students, their parents, the school community and the community at large the extent of learning that is taking place in schools. Of particular importance, however, is the information given to individual students in order to show them in clearly identifiable terms the growth that they have achieved over a particular period of time as a result of their efforts in the classroom and at school. It is also important to recognize that the scales of learning extend across all levels of schooling and beyond. Learning does not end at the completion of schooling or tertiary education, but extends throughout life as opportunities for lifelong learning and development are followed.

REFERENCES


