Making the Most Effective Use of Statistics and Test Results

Some common terms used in research design, statistical analysis, and testing

Suggestions for interpreting, explaining, and reporting test results

Suggestions for questionnaire design and analysis

Marjorie Clegg

Quality Assurance Division
Program, Staff Development, and Quality Assurance Department

revised July 2000
Acknowledgements

The authors wish to thank the following people for their advice and assistance in preparing this manual:

Linda Clarke, Research Officer
Yasmin Sankar Khan, Research Analyst
Laura McAlister, Principal of Special Education
Rose Marie Batley, Superintendent
Michael T. Carlon, Superintendent (retired, 1998)
Dan Mason, Superintendent
Judy Turriff, Superintendent
David S. Ireland, Consultant

© 1997 The Carleton Board of Education
© 1998 The Ottawa-Carleton District School Board
© 1999, 2000 The Ottawa-Carleton District School Board

Copyright of this work is owned by The Ottawa-Carleton District School Board. No reproduction of this work, in whole or in part, may be made without the express authorization of The Ottawa-Carleton District School Board.
# Table of Contents

Introduction .................................................................................................................. 1  

Research Design ........................................................................................................... 2  
  Questions .................................................................................................................. 2  
  Terms ....................................................................................................................... 2  

Statistical Analysis ....................................................................................................... 3  
  Central Tendency ...................................................................................................... 3  
  Dispersion .................................................................................................................. 3  
  Measures of Shape .................................................................................................... 5  
  Ways of Displaying Data ........................................................................................... 8  
  Samples and Populations ........................................................................................... 11  
  Significance (Hypothesis testing) ............................................................................ 13  
  How important are the results? ................................................................................ 18  

Testing .......................................................................................................................... 20  
  Testing Terms ......................................................................................................... 20  
  Choosing Tests/Assessments .................................................................................... 27  
  Explaining Results and Planning Improvement ...................................................... 29  
  School Action Plans .................................................................................................. 31  
  Reporting Results ..................................................................................................... 35
Questionnaires

Before You Begin .................................................................................................................. 38
Designing the Questions ........................................................................................................ 39
Distribution and Return ......................................................................................................... 40
Analysing the Results ........................................................................................................... 41
Reporting the Results ............................................................................................................ 42

Sources and Resources .......................................................................................................... 43
INTRODUCTION

The use and misuse of research and test data to influence public opinion and make decisions have become commonplace in education.

The purpose of this manual is to provide educators with information about statistics and the use of data, particularly the use of group assessment data, in the hope that it will help them evaluate their own and those reported by others.

The manual was prepared by the staff of the Quality Assurance Division for use by teachers and administrators.

Suggestions and comments are welcome.
RESEARCH DESIGN

Questions

What do you want to know?

What evidence do you need?

What will you do as a result?

Terms

experimental group
An experimental group is a group of people given a particular, preferably well-defined, treatment or intervention to see if it works.

control group
A control group is a group of people similar to an experimental group on relevant dimensions (e.g. gender, age, experience, etc.) but receiving a different treatment (or no treatment at all) from the experimental group to determine if the experimental treatment has had any effect.
Example: a teacher wanted to find out whether a commercial keyboarding course worked well. Two groups of students participated in the study: those taking the course in the first school term (experimental group) were tested before and after the course; the control group consisted of students who were taking the course during the second school term. The students in the control group were tested at the same times as the experimental group (to control for maturation effects), as well as after their turn at the course.
Central Tendency

mean  The mean is an arithmetic average, the sum of the values of all the observations divided by the number of observations.

median  The median is the value above and below which half the data fall.

This measure is less influenced (than the mean) by extreme values.

mode  The mode is the most frequently occurring value(s).

In the following example, scores on a spelling test ranged from 60% to 100%; the mean was 80.8%, but the value above and below which half of the cases fall (median) was 80% and the most frequent value (mode) was 75%.

Scores: 60, 65, 70, 75, 75, 75, 80, 85, 85, 90, 95, 95, 100

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>80.8</td>
</tr>
<tr>
<td>Median</td>
<td>80</td>
</tr>
<tr>
<td>Mode</td>
<td>75</td>
</tr>
</tbody>
</table>

In a normal distribution, the mean, median, and mode would all be the same value (see page 4).

Dispersion

range  The range is the difference between the maximum and minimum observed values.
variability

Variability is a general term relating to the amount of scatter or dispersion there is among a set of scores.

variance

Variance is an estimate of the variability or dispersion of a population based on a sample.

The variance is computed by summing the squared differences from the mean for all observations and then dividing by one less than the number of observations. If all observations are identical – i.e., if there is no variation—the variance is 0. The more spread-out they are, the greater the variance.

standard deviation

The standard deviation is a measure of how much numbers vary around the average, how widely values are dispersed from the average value (mean).

The standard deviation is the square root of the variance.

In the example below, the standard deviation (variability) for the Intervention (Experimental) group was twice that of the Control group.

<table>
<thead>
<tr>
<th>Test Measure</th>
<th>Group</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Task</td>
<td>Control</td>
<td>13.52</td>
<td>.53</td>
</tr>
<tr>
<td></td>
<td>Intervention</td>
<td>13.97</td>
<td>1.08</td>
</tr>
</tbody>
</table>

standard error

The standard error is an estimate of what the standard deviation of a statistic would be if successive values were found for that statistic through repeated testings (usually on different, but similar, samples drawn from the same population).
| standard error of measurement | The standard error of measurement is an estimate of the *standard deviation* that would be found in the distribution of scores for a specified person if s/he were to be tested again and again on the same or a similar test (assuming no learning). |
normal distribution  A normal distribution is the distribution of a set of random events.

For many variables, most observations are concentrated near the middle of the distribution. As distance from the central concentration increases, the frequency decreases. Such distributions are often described as “bell-shaped.”

The normal distribution is symmetrical, with the three measures of central tendency (mean, mode, median) coinciding. 95% of all observations fall within two standard deviations of the mean, and 68% within one standard deviation.

A broad range of observed phenomena in nature and in society are approximately normally distributed because they tend to occur randomly. The normal distribution is the most important theoretical distribution in statistics.

Note to teachers: There is no reason to suppose that your students will conform to a normal distribution unless, which is unlikely, they are a random selection from the population.
Skewness is the degree of asymmetry of a distribution around its mean.

*Positive* skewness indicates a distribution with an asymmetric tail extending towards more positive (higher) values: more cases occur below (to the left of) the mean.

*Negative* skewness indicates a distribution with an asymmetric tail extending towards more negative (lower) values: more cases occur above (to the right of) the mean.

Note that the terms *positive* and *negative* skewness may be misleading: positive skewness indicates that most of the scores occur in the lower range; negative skewness indicates that most of the scores occur in the higher range.
Kurtosis is the extent to which, for a given standard deviation, observations cluster around a central point.

If cases within a distribution cluster more than those in a normal distribution (more “peaked”), the distribution is called *leptokurtic*.

If cases cluster less than in a normal distribution (flatter), the distribution is called *platykurtic*. 
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>Frequency is the number of occurrences of a particular value observed, e.g., the number of individuals obtaining any specified score or falling in any specified class interval.</td>
</tr>
<tr>
<td>frequency distribution</td>
<td>A frequency distribution is any orderly arrangement of scores, usually from highest to lowest, showing the number of individuals (i.e., the frequency) attaining each score or falling in each class interval.</td>
</tr>
<tr>
<td>variable</td>
<td>A variable is any trait or characteristic that may change with the individual or the observation.</td>
</tr>
<tr>
<td></td>
<td>More strictly, a variable is any representation of such a trait or characteristic which is capable of assuming different values: e.g., a test score is a variable.</td>
</tr>
<tr>
<td>histogram</td>
<td>A histogram is a representation of a range of observed values subdivided into equal intervals. Each column in a histogram represents the number or proportion of cases with values within the interval.</td>
</tr>
</tbody>
</table>

![Histogram Diagram](image-url)
A boxplot is a display of the summary statistics for the distribution rather than actual values. It plots the median, 25th percentile, 75th percentile, and values that are far removed from the rest.

From the median you can determine the central tendency. From the length of the box you can see the spread, or variability, of the observations. If the median is not in the center of the box, you know that the observed values are skewed. If the median is closer to the bottom of the box than the top, the data are positively skewed: there is a tail with large values. If the median is closer to the top of the box than the bottom, the opposite is true: the distribution is negatively skewed. The length of the tail is shown by the whiskers and the outlying and extreme points.

Boxplots are particularly useful for comparing the distribution of values in several groups, e.g. performance of countries or provinces in international studies of achievement.
charts

There are many types of charts used to display data, including lines, bars, and pies.

There are also many ways of misrepresenting data graphically, including 3-dimensional bars, which distort differences and make it difficult to find exact values. In the following example, can you find the exact score for student # 7?

Samples and populations

The totality of all cases about which conclusions are desired is called the population. The cases actually included in a study are the sample, a general term referring to a group, however selected, assumed to represent an entire population.

Statistics help us draw inferences about populations based on observations obtained from small random samples, or samples in which the characteristics and relationships of interest are independent of the probabilities of being included in the samples. The results serve as estimates of the unknown population values.

If you have the population, no sample statistics are necessary; anything you find can be said to be “true” and reliable. The results describe the population exactly.

The observations included in a study are just one of many random samples that could have been selected from a population. Different samples usually produce different estimates. The theoretical distribution of all possible values of a statistic obtained from a population is called the sampling distribution of the statistic. The mean of the sampling distribution is called the expected value of the statistic. The standard deviation is termed the standard error. Sampling distributions of most commonly-used statistics (like the mean) are tabulated and readily accessible.
Based on the sampling distribution, you can calculate the probability that a difference at least as large as the one observed would occur if the two population means were equal. This probability is called the significance level; if the significance level is small enough, usually less than .05 or .01, the hypothesis that the population means are equal is rejected.

**random sample**
A random sample is a sample selected from a population in such a way that each member of the population has an equal chance of being selected.

**sample size**
One of the first questions arising in planning a survey is how big the sample should be. If the sample is too large, it can be costly in terms of time and money, without an increase in reliability. On the other hand, if the sample is too small, there will be inadequate information obtained and any time and effort will be wasted.

The more people in the sample, the closer you get to a true estimate for the whole population.
A researcher can get a fairly accurate picture from a properly-chosen sample. The difference between this picture and what the results would be for the whole population is popularly called the margin of error.

When a researcher reports that 65% of the people surveyed expressed a certain opinion and that the survey is accurate within 3 percentage points 19 times out of 20:

- the margin of error is 3 percentage points
- the percentage of the population holding that opinion would be between 62% and 68%, (no more than) 3 percentage points away from 65%
- the confidence with which the researcher can say this is 95% (19 times out of 20)

The margin of error and the level of confidence depend on the size of the sample.
# Significance (Hypothesis-testing)

**significance levels**  
Hypothesis testing is used to make a decision on whether to accept or to reject a claim on the basis of sample information.

The null hypothesis is the hypothesis that the sample statistic is the same as the “real” population value. When the sample data indicate that a null hypothesis should be rejected, the observed difference is said to be *significant*.

The significance level of a test is the probability that the test statistic falls within the rejection region when the null hypothesis is true. The .05 level of significance indicates that the risk of rejecting a true null hypothesis does not exceed .05; the .01 level of significance indicates that the risk of rejecting a true null hypothesis does not exceed .01.

**Note:** statistical significance is not necessarily equal to meaningful educational significance. A small difference between results of very large samples may be statistically significant; they may or may not be meaningful (see *effect size*, page 16).
Chi Square is a statistical test used to compare observed and expected values to decide if the original hypothesis is valid – i.e., that two variables of a crosstabulation are independent of each other.

Two variables are, by definition, independent of each other if the probability that a case falls into a given cell is simply the product of the marginal probabilities of the two categories defining the cell.

In the following example, eating or not eating candy was not independent of gender: males were much more likely to eat candy, with a zero probability that this result could have happened by chance.

A t-test is a statistical test used to determine whether two samples are likely to have come from the same two underlying populations that have the same mean.

In the following example, the indication is that the two samples did not come from the same population.
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ANOVA</strong></td>
<td>Analysis of variance (ANOVA) is a statistical test used to determine if means from several samples are the same or different.</td>
</tr>
<tr>
<td><strong>confidence interval</strong></td>
<td>The confidence interval is a range on either side of a sample mean.</td>
</tr>
<tr>
<td><strong>F test</strong></td>
<td>The F-test is a statistical test used to determine if two samples have different variances (levels of diversity).</td>
</tr>
<tr>
<td></td>
<td>In a report, the result of an F-test might appear as: F (1,18) = 6.71, $p&lt;.021$.</td>
</tr>
<tr>
<td></td>
<td>(1,18): represents the degrees of freedom (df) = the number in the sample minus 1.</td>
</tr>
<tr>
<td></td>
<td>$p&lt;.021$: less than 2/100 that this would happen by chance.</td>
</tr>
<tr>
<td><strong>correlation</strong></td>
<td>A correlation is a dimensionless index that ranges from -1.0 to 1.0 inclusive, and reflects the extent of a linear relationship between two data sets; it indicates the tendency for two (or occasionally more) variables to change values concomitantly.</td>
</tr>
<tr>
<td></td>
<td>Note: evidence of correlation is not evidence of causation.</td>
</tr>
</tbody>
</table>
coefficient of correlation

The coefficient of correlation is an index number indicating the degree of relationship between two variables; i.e., the tendency for values of one variable to change systematically with changes in values of a second variable.

A perfect relationship is indicated by +1.0 (positive correlation) or -1.0 (negative correlation).

The most basic type is the Pearson correlation which is used when both variables are continuous, distributed symmetrically, etc.

The following table is typical of the way in which such results are reported. Clearly, the correlation is 1.00 when the variable is correlated with itself. The statistical significance of each correlation is based on the sample size; in this example, correlations above .3 were statistically significant.

<table>
<thead>
<tr>
<th></th>
<th>Number of students in rental housing</th>
<th>Estimate of child poverty for school area</th>
<th>Student mobility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of students in rental housing</td>
<td>1.00</td>
<td>.71</td>
<td>.47</td>
</tr>
<tr>
<td>Estimate of child poverty for school area</td>
<td>.71</td>
<td>1.00</td>
<td>.10</td>
</tr>
<tr>
<td>Student mobility</td>
<td>.47</td>
<td>.10</td>
<td>1.00</td>
</tr>
</tbody>
</table>

regression

A regression analysis predicts values based on relationships in existing data.

Linear regression seeks to fit a linear function to a data set.
factor (1) strictly and technically, an element or variable presumed to exist because of its ability to help explain some of the interrelationships noted among a set of observations

(2) equally properly, the ability or characteristic represented by a factor

(3) loosely, anything which may be partially responsible for a result or outcome (e.g., "study is an important factor in obtaining good grades")

factor analysis Factor analysis is one of several complex statistical procedures for analysing the intercorrelations among a set of tests (or other variables) for the purpose of identifying the factors (1 and 2 above), preferably few in number, that cause the intercorrelations.

Factor analysis is widely used in efforts to understand the organization of intelligence, personality, and the like, as well as to understand the relationship among different tasks, e.g. reading and writing.
## How important are the results?

The preceding section dealt with the statistical significance of results. As already noted, statistical significance does not always indicate meaningful or important results.

Usually, we are interested in finding out how well something works and how it compares to other ways of doing things.

<table>
<thead>
<tr>
<th>meta-analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta-analysis is a method for comparing different approaches, e.g. reading interventions. The intention is to find out how much difference some kind of intervention makes. Sometimes, meta-analysis is done comparing the results of several or many studies done by a variety of researchers, but this kind of analysis can also be done within one study.</td>
</tr>
</tbody>
</table>
effect size

A key concept in meta-analysis is **effect size**. Effect size is derived from the means of the experimental (intervention) group compared with that of a control (comparison) group. Using this method, different interventions can be compared using pre- and post-tests, even if these tests are on different marking scales.

Effect size, standardized to allow for comparisons of different interventions, is calculated by determining the difference between the post-test means for the experimental and control groups and dividing the difference by a “pooled standard deviation.” This statistic can be corrected for bias, restrictions in range, small samples, etc.

The effect size is the best estimate of the magnitude of the effect if the experiment were repeated in a series of classes, for example.
In the example below, Intervention 1 involved a very small (N=5) group of students; the results would be unlikely to be statistically significant—in fact, most statistical tests cannot be done on such a small sample.

Intervention 2 involved more students (N=12), but the effect was smaller.

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Intervention 1</th>
<th>Intervention 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental Mean</td>
<td>6.00</td>
<td>5.91</td>
</tr>
<tr>
<td>Control Mean</td>
<td>4.20</td>
<td>5.17</td>
</tr>
<tr>
<td>Pooled SD</td>
<td>1.24</td>
<td>1.86</td>
</tr>
<tr>
<td>Effect Size</td>
<td>1.45</td>
<td>0.40</td>
</tr>
<tr>
<td>Error</td>
<td>0.75</td>
<td>0.42</td>
</tr>
</tbody>
</table>

The effect size might be shown graphically as:
## Testing terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>criterion v norm-referenced</td>
<td>The main purpose of norm-referenced testing is to compare students with a defined population (see “norms”). <em>The Canadian Cognitive Abilities Test</em> (CCAT) is an example of a norm-referenced test. The main purpose of criterion-referenced testing is to determine how well the students are doing in relation to specified standards of performance – regardless of how many other students achieve the same level of proficiency. The Grade 3 testing by the Education Quality and Accountability Office (EQAO) is an example of criterion-referenced testing.</td>
</tr>
<tr>
<td>analytic scales v holistic scoring</td>
<td>Analytic scales, such as those used in the Ottawa-Carleton District School Board Senior English Evaluation Projects (SEEP), specify the criteria for marking various aspects of the construct under consideration. In SEEP, essays are marked according to explicit criteria in relation to focus, support, organization, style, mechanics. Other categories, such as creativity and humour, might also be used in marking essays. Holistic scoring is used to mark the product as a whole. For example, an essay would be marked for its overall impression.</td>
</tr>
</tbody>
</table>
validity  

Validity is the extent to which a test does the job desired of it.

Validity is expressed as a correlation coefficient. Validity coefficients generally range from .00 to .60, with some as high as .80, but rarely are they higher. A validity coefficient of .40 is considered acceptable.

• content validity  

Content validity is the logical evidence that the item content of a test is suitable for the purpose for which the test is to be used.

To establish content validity, it is necessary to examine systematically the content of the test items to determine whether they are representative of the skills and knowledge within the domain being tested. In evaluating content validity, the appropriateness of the type of items, the completeness of the item sample, and the way in which the items assess the content of the domain must be considered.

The concept is used principally with achievement tests.
• criterion-related validity
  Criterion-related validity is the relationship between test scores and some type of criterion or outcome, such as ratings, classifications, or other test scores.

There are two types of criterion-referenced validity: concurrent and predictive.

  – Concurrent validity refers to whether or not test scores are related to some other measured and available criterion. For example, if standardized math test scores for a class correlate with the teacher’s assessment of the pupils’ knowledge of math, the test can be said to have concurrent validity.

  – Predictive validity refers to the correlation between test scores and future performance on a relevant criterion. It answers the question: *Is the score obtained on a test an accurate predictor of future performance on the criterion?*

• construct validity
  Construct validity is the extent to which the test measures a theoretical construct.

For example, the construct validity of an intelligence test would be evaluated by examining how the items relate to a specific theory of intelligence.

Test validation based on a combination of logical and empirical evidence of the relationship between the test and a related theory; it is concerned with the psychological meaningfulness of the test.
• face validity  Face validity is superficial appearance of validity; i.e., a test looks as if it should measure what is intended (regardless of the presence of other relevant data indicating that it is actually valid for some purpose).

reliability  Reliability is the consistency or stability of a test or other measuring instrument in repeated testing; it is necessary for, but not sufficient for, validity.

It is commonly expressed as a reliability coefficient ranging from 0 (completely inconsistent) to 1.0 (indicating perfect consistency). There should be evidence of at least one type of reliability, with a coefficient greater than or equal to .80.

• internal reliability  Internal reliability is the extent to which similar items in a test produce similar scores.

• test-retest reliability  Test-retest reliability is the extent to which the same instrument administered to the same students produces similar scores over different time periods.

• equivalent-form reliability  Equivalent-form reliability is the extent to which different instruments generated by the same procedure produce similar scores for the same students.
item analysis  

Item-analysis is the act or process of examining a test item empirically to determine:
(a) its difficulty value (a test item's difficulty, usually expressed as the percentage of individuals in a given group answering the item correctly), and
(b) its discrimination value (any of several statistics used to express the extent to which a test item shows a difference between high-ability and low-ability students).

Note: such values will differ somewhat from group to group and from time to time and according to the particular statistic used.

standardization  

In the general sense, standardization is the act or process of developing a standardized way of testing, by giving the same test in the same way.

In the more specific sense, standardization refers to developing standardized (see below) commercial tests. Many stages are involved in careful standardization, among these: tryout of items, item analyses, validation studies, reliability studies, development of norms, etc.

standardized test  

A standardized test is an empirically developed test, designed for administration and scoring according to stated directions, for which there is evidence of validity and reliability, as well as norms.

The OCDSB has developed a procedure dealing with the use of standardized tests in the Board; this procedure appears in Appendix B.
Norms are test scores which have been transformed into a normal distribution with a specified mean and standard deviation.

Norms can be expressed as percentiles, standard scores, stanines, or grade equivalents, indicating how well the student performed compared to a defined population of the same age or in the same grade.

Defined student population refers to the large number of students whose test results were used for the purpose of standardization. Test developers generally collect and provide information about these students, such as the size of the population, the age of the students, tested, the language spoken at home, etc. The defined student population used in the norming should be similar to the population to be tested; if the populations are very different, the results may not be valid or reliable.
A standard score is one of several derived scores based on the number of standard deviations between a specified raw score and the mean of the distribution.

Standard scores are a measure of the relative ranking of a mark, assuming a normal distribution of marks. One must keep in mind, however, that since standard scores represent results relative to system means, absolute improvement (e.g. raising the mean score from 64% to 70%) cannot be reported; only improvement relative to others can be reported.

i.e., if a school’s mean increases from 64% to 70%, but the same improvement occurs for most of the other schools in the Board, the school’s mean may still be lower than the system mean.

The use of standard scores facilitates comparisons of results from different years. We can also combine standard scores to give one mean score for a semestered school.

A z-score is the basic standard score widely used in test-related research.

The z-score measures how many standard deviations an individual score is away from the mean.
percentile  A percentile is the value below which a given percentage of cases fall.

A percentile represents any of the 99 points along the scale of score values that divide a distribution into 100 groups of equal frequency; e.g., P73 is that point below which fall 73 percent of the cases in a distribution.

Percentiles may be local (e.g., for a one-off testing) or one of the scores published for commercial tests.

Note: small differences in raw scores (i.e., one or two more items correct) can make large differences in percentile ranks.

quintiles  By definition, for each range of scores given for a quintile, 20% of all students tested scored in that range.

ranking  Ranking gives the position of a value in a sorted range of values.
A grade equivalent score is the grade level at which the typical pupil makes a particular raw score.

This statistic takes the form of a normal distribution with the mean set for a particular age/grade level at the time the test is taken. The mean grade-equivalent for a grade 6 test taken half-way through the year would be 6.5. Empirically, or simply by extrapolation, other scores are given a higher or lower grade equivalent.

On the Canadian Tests of Basic Skills (CTBS), for example, raw scores may be converted into grade equivalent scores between K2 and 81. For those grade equivalents, the first digit represents the grade and the second digit the month within the grade in which the typical pupil makes the corresponding raw score. (Grade equivalents outside this range were derived statistically.) A year’s growth is considered to be 10 points by definition.

This method of presenting results is deceiving, corresponding, as it would seem, to a recognizable school category. It should be noted that a “grade equivalent” is purely a name for a statistic and is in no way equivalent to a grade level. Also, grade-equivalent scores earned on different tests do not necessarily represent equally good performance.
levels of performance

The concept of standards is meant to change the way performance is reported for students. These standards describe performance at different levels which cannot be averaged; rather, one would report what percentage of students was performing at each level.
# Choosing Tests / Assessments

**fairness**  
*Principles for Fair Student Assessment Practices for Education in Canada* (Appendix B) sets out guidelines for both developers and users of assessments. The guidelines cover developing and choosing methods for assessment, collecting assessment information, judging and scoring student performance, summarizing and interpreting results, and reporting assessment findings.

<table>
<thead>
<tr>
<th>purposes</th>
<th>What is the purpose of the test/assessment?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>What do you want to find out?</td>
</tr>
<tr>
<td></td>
<td>Some of the common purposes are:</td>
</tr>
<tr>
<td></td>
<td>• evaluating program effectiveness</td>
</tr>
<tr>
<td></td>
<td>• diagnosing student strengths and weaknesses</td>
</tr>
<tr>
<td></td>
<td>• public accountability</td>
</tr>
<tr>
<td></td>
<td>• screening for specific purposes (e.g. gifted, special education, remediation)</td>
</tr>
<tr>
<td></td>
<td>• assigning report card marks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>form of the test/assessment</th>
<th>Does the form of the test/assessment reflect your teaching practices?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Is the form of the test/assessment fair to all of your students?</td>
</tr>
</tbody>
</table>
content Is the test/assessment valid?

Does it measure what you teach?

Does it measure what is important?

bias Is the test/assessment content free from bias (cultural, gender, racial, etc.)?

marking How is the marking done?

Is it reliable? if marked by a different person, would the results be the same?

Is it based on clear criteria?

reliability Is the test/assessment reliable?

If the students were given the same test again, would the results be the same?

Is there a published coefficient of reliability?
Is it acceptable?

results Will the results be in a form useful for the purposes you wish?

If the test is normed, are the norms appropriate to your student population?
Explaining Results and Planning Improvement

questions to ask *

1. Given what you know (from other sources) about this school, were there any unexpected results?

   If so,

   a) What were they?

   b) Are they really different from what you expected?

   e.g. Minor fluctuations in results (year by year) are to be expected.

   e.g. Fluctuations are more likely to occur when the population taking the test is small.

* Specific questions for EQAO results can be found in Appendix D (students) and E (classes/schools); questions for board-wide exams can be found in Appendix F.
c) How would you account for the results?

Were the results a consequence of or a combination of:

the teaching?
  e.g. to what extent did the teachers provide the opportunity to learn (OTL) what was to be tested?

the students?
  e.g. was the test normed or developed with this student population in mind?

the test?
  e.g. was the test different from previous years in any significant way?
  e.g. how well did the test match the curriculum (Test-Curriculum Match: TCM)?

d) What other information do you need?

2. In what area(s) did the school do very well?

3. In what area(s) did the school not do well?

4. What strategies should you consider to improve student achievement?

5. What did you do as a consequence of last year’s testing?

6. Was there any improvement on this year's test?
value-added results  In the past few years, there has been increasing interest in analysing results to show how well students are performing compared to how well we would have expected them to perform based on previous testing of either ability or achievement. In other words, how much value has the school/class added? Is it less than, equal to, or more than what we would have expected?

For example: we might wish to analyse results on two OCDSB math tests given in different grades.
The most important use of test and questionnaire results is planning for school improvement.

An excellent resource on using data for school improvement is the booklet published by the Training and Assessment Consortium:

*Linking Achievement Data to School Improvement.* 1996

**participants**

Planning can involve different groups of people at different times and in different ways. For example:

- Principals and teachers may meet to discuss the results.
- A cross-section of teachers from the school may meet to discuss the results and make recommendations to the principal.
- The principal may involve the members of the School Council and Student Council (as appropriate) in either creating the plan or vetting it.

**timelines**

Planning can be for the short-term and/or for the long-term.

**formats**

There are many formats available for doing such plans. Two such formats follow.
Example 1: School Action Plan

<table>
<thead>
<tr>
<th>School:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Major Issue:</td>
<td></td>
</tr>
</tbody>
</table>

**Exit Outcome:** By May 31, 2001, ...

**Enabling Outcome:** By May 31, 1999, ...

<table>
<thead>
<tr>
<th>Activity</th>
<th>Responsibility Centre</th>
<th>Deadline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 2: School Action Plan

1. Improvement Outcome:

<table>
<thead>
<tr>
<th>Evidence of problem (How do you know it is an issue?)</th>
<th>Hypotheses (What are possible explanations?)</th>
<th>Strategies (How will you get there?)</th>
<th>Timelines (When will you get there?)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Responsibility</strong> (Who will do it?)</td>
<td><strong>Resources</strong> (What will you need to do it?)</td>
<td><strong>Monitoring</strong> (How will you know if you are there?)</td>
<td><strong>Communication Plan</strong> (How will you let others know you are there?)</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>---------------------------------------------</td>
<td>-----------------------------------------------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Reporting Results

school comparisons  Comparing schools, systems, provinces, or countries is extremely difficult to do in a meaningful way.

There are numerous variables to be considered, including differences in:

• *population*
  e.g. proportion of ESL or special needs students, proportion of students still in school or taking that subject

• *program options*
  e.g. gifted, French Immersion

• *test-curriculum match* (TCM)

• *opportunity to learn* (OTL).

ranking  Rankings can be misleading, with differences between rankings being very small: e.g., the difference in mean score for the 1st-place school/country may not be significantly different from the mean score of the 5th-place school/country.

Test scores are not exact measurements; they are estimates. Using confidence intervals or quintiles or some other measure of dispersion helps to indicate this estimate.

Rankings do not show how well schools or systems affect student progress.
the place of achievement results from group testing

- achievement test scores do not provide comprehensive information about achievement (e.g., achievement scores are reported only for certain subjects at certain grade levels on certain tests on certain test items)

- there are other aspects of schools that are important, including special programs/services, school environment, etc.

- in order to understand achievement test results, one should have an understanding of the population in the school context

When achievement results are reported, therefore, they should be reported within context and with other indicators of success.
school profiles  

School profiles are reports on individual schools. Information commonly reported in school profiles includes characteristics of students, achievement, environment, facilities, program offerings, services, staff, and safety.
QUESTIONNAIRES

Before You Begin...

- What do you want to know?
- What evidence will you accept?
- What will you do with it once you have it?

- Who should you ask? [students, parents, staff?]
- How many of them? [all vs. sample]
- Will they be able to give you the information you want?

Do...

Ask the questions that will get the information you want.

Ask all of the groups of people who should be included.

Look at other questionnaires to get ideas.

Ask enough people to get a reliable result.

*Look ahead to design, distribution, analysis, and reporting of results!*

Don’t...

Ask questions you don’t want the answers to.

Make the process so complicated or time-consuming that very few people will answer.
**Designing the Questions**

- What do you want to know?
- Should the questions for all groups be the same? [level of understanding, point of view]
- How should the questions be presented? [format]
- What type of answer will work best? [options? open-ended?]

**Do...**

Ask the questions that will get the information you want.

Ask some questions which can be compared across groups, if applicable.

Make questions easy to understand.

Give people a chance to give both positive and negative feedback.

Ask fair, balanced questions.

*Look ahead to distribution and return, analysis, and reporting of results!*
Don’t...

Ask questions you don’t want the answers to.

Make questions too complicated.

Deal with more than one issue in a question.

Ask questions in the negative (if possible).

Ask questions that are too personal.

Ask questions that are biased or embarrassing.
**Distribution and Return**

- What is the best way to ask the questions? [phone? questionnaire? focus group?]
- What is the best way to contact respondents? [newsletter? mail?]
- When is the best time to do this? [time of day, time of year]
- Do you need special versions for some people? [translations?]

**Do...**

Clearly identify who you are, why you are asking the questions, and what you plan to do with the results.

Ensure confidentiality.

Let respondents know the deadlines: when you need their answers and when you think the report will be ready.

Make it easy and cost-efficient to distribute.

Ask the questions at a convenient time (if at all possible).

Have translations if a large number of your respondents would find it helpful.

Make it as easy as possible for people get the information back to you.

*Look ahead to analysis and reporting of results!*
Don’t... Call people at dinner time.

Give people too little time to respond.
### Analysing the Results

- How will you record results? [who will do it? how?]
- What statistics do you want to use? [number? percentage?]
- Do you want to separate out groups?
- What resources do you have to do this? [people, software, time]

- What do the results tell you?
- What recommendations should you make?

**Do...**

Ask the questions in a way that will give you data you can record easily.

Record data in a spreadsheet which can be manipulated to give you a variety of reports.

Make sure your data are recorded accurately.

Note what is working well and what isn’t.

Make recommendations for what to keep, what to change, what to stop

*Look ahead to the reporting of results!*

**Don’t...**

Ask questions you can’t analyse.

Make the analysis too complicated.

Note only negative or positive aspects.
Reporting the Results

- Who is your audience?
- What do they want to know?
- What format will work best? [on questionnaire itself? summary?]
- Do you need more than one version? [summary vs. whole report]
- What is the best way to distribute the report? [newsletter? web? mail?]  

Do... Report all results.

Report clearly and concisely.

Report back to the people who participated.

Report to others who might be interested.

Make recommendations based on your results.

Check on your progress and report it.

Don’t... Report results which will identify individuals, unless they wish it and have given you permission.

Conceal results you don’t like.
Sources and Resources


*Microsoft Excel Function Reference* (Version 4.0 for Macintosh), Microsoft Corporation, 1992


*Principles for Fair Student Assessment Practices for Education in Canada*. Joint Advisory Committee, Edmonton, Alberta, 1993
