Alberta

Freedom To Create. Spirit To Achieve.

Analyzing Diploma Examination Results at the School Level

Government of Alberta

Education

Goals of today

- Provide you with the tools to interpret diploma exam results for your school.
- Provide you with tools to take back to your staff to assist them in interpreting their individual results.
- Develop ways of recognizing successes and identifying fixable weaknesses.
- Help you develop strategies to set reasonable targets.
- Remember: Reports reveal what the performance levels are, not **why** those levels were achieved. To get the full picture, one must also look at the factors that contribute to students' success.

What factors affect student achievement?

- Brainstorm possible factors that you experience in your own professional practice.
- Describe the characteristics of your student population.
- Which factors are within a teacher's or school's control?
- Would you expect your school results to be higher or lower than the provincial average? Why?
- Would you expect participation rates in the various subjects to be higher or lower than provincial rates? Why?

Considerations in Diploma Examination Reporting

- Reliability
- Validity
- Target Setting
- Maintaining Consistent Standards

Reliability and Validity

- "Reliability" means the consistency with which a set of test scores measures whatever they measure.

- "Validity" means the accuracy with which a set of test scores measures what they are intended to measure.

Ensuring Reliability of Diploma Examinations

- Field Testing
  - Sample selection
  - Sample size

- Item Analysis

Ensuring Validity of Diploma Examinations

- Design
  - Curriculum-based blueprint
  - Teacher item-writers
- Field Testing
  - Teacher validation
- External review

Two standards: acceptable and excellent

- The acceptable standard (between 50% and 79%) shows a reasonable understanding of the basic content and process objectives of the relevant program of studies; the exact standards are outlined in each of the subject bulletins.
- The standard of excellence (80% and more) is **not** merely more of the same, with fewer mistakes; the quality of the work, and the complexity of the tasks being accomplished, is of a different character, than work at the acceptable standard.
- See the Social Studies 30-1 rubrics for writing to see the difference between work rated as 3, compared to work rated 4 or 5.

Proficient and Satisfactory in Social Studies 30-1 Evidence Scale

Proficient
Evidence is specific and purposeful. Evidence may contain some minor errors. A capable and adept discussion of evidence reveals a solid understanding of social studies knowledge and its application to the assignment.

Satisfactory
Evidence is conventional and straightforward. The evidence may contain minor errors and a mixture of relevant and extraneous information. A generalized and basic discussion reveals an acceptable understanding of social studies knowledge and its application to the assignment.

Maintaining Consistent Standards

Philosophy

- Fairness to students
- Accurate tracking of achievement over time

Strategies

- Test Equating or Linking (for all examinations except Français 30, French Language Arts 30 and Science 30)
- Secured Examinations
- Delayed Item Release

Test Equating Details (Humanities)

- As the tasks and standards in Part A are essentially the same from year to year, with only the content of the task differing from examination to examination, the Part A marks are not equated, and separate writings for Part A and Part B are feasible.

- Part B marks are equated, with the marks being raised if the examination is harder than the baseline examination, and being lowered if the examination is easier than the baseline examination.

- The results in Tables 1 – 4 reflect equated marks, and the results in Tables 5 – 7 (all parts) reflect raw, unequated marks.

Test Equating Details (1) (Mathematics – Sciences)

- This applies to all mathematics and science examinations.
- The complete examination, both multiple-choice and numerical-response, is equated.
- Total marks were equated, with the marks being raised if the examination was harder than the baseline examination and lowered if the examination was easier than the baseline examination.

Test Equating Details (2) (Mathematics – Sciences)

- Equating was used in Pure Mathematics 30, Applied Mathematics 30, and Biology 30, as these examinations were essentially consistent from year to year.
- Equating has now started for Science 30, as the numbers taking any examination are now large enough for the equating process to be valid.
- Equating **may** be used for Chemistry 30 and Physics 30 because the Program of Studies is fairly new, and a baseline examination is almost ready to be selected.

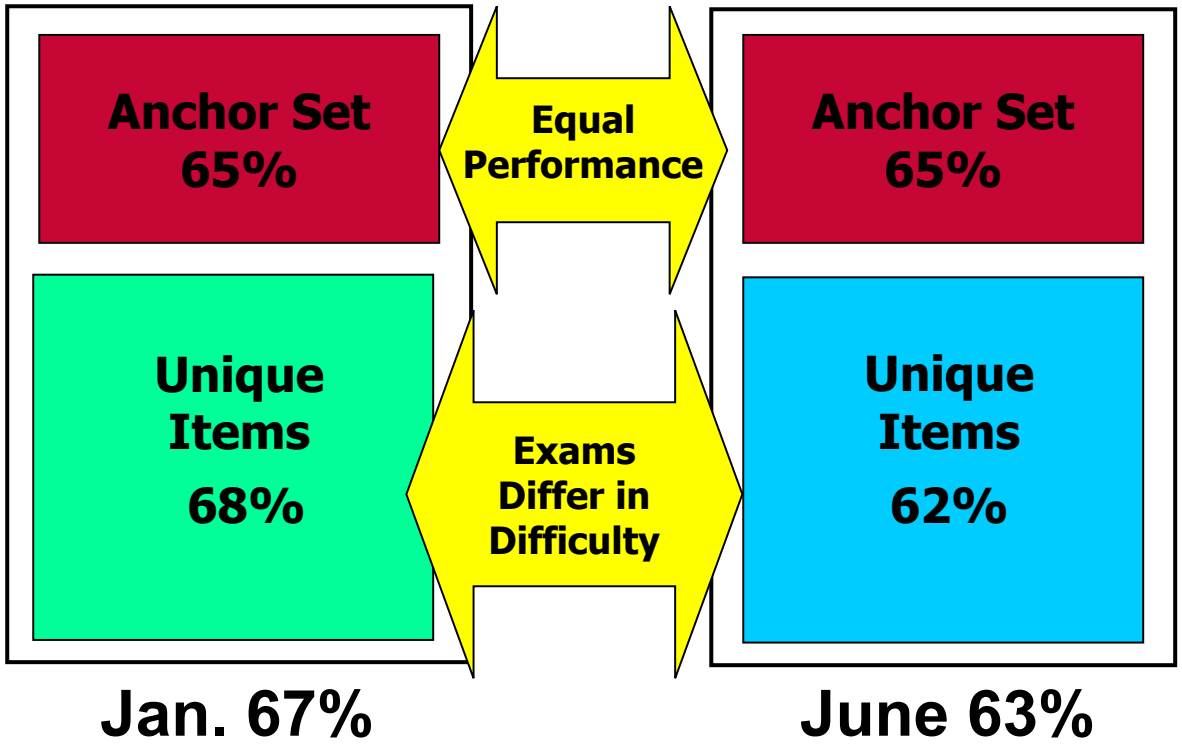
Test Equating Details (3) (Mathematics – Sciences)

The results in Tables 1 – 4 reflect equated marks, and the results in Tables 5 – 7 (all parts) reflect raw, unequated marks.

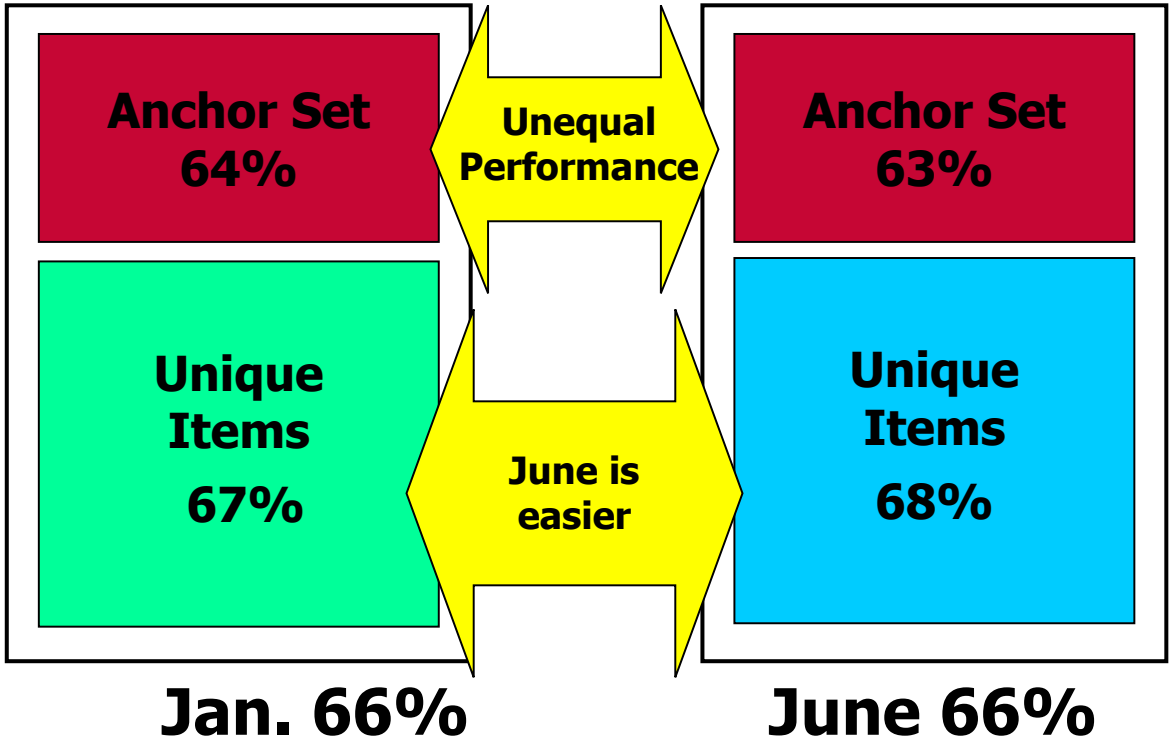
Characteristics of an "Anchor Set"

- Represents a cross-section of curriculum content
- Forms at least 20% of the entire machine-scored (MS) examination
- Has an average difficulty which is the same as the average difficulty of the entire MS examination
- Has a range of difficulties similar to the examination
- Appears in relatively the same order in subsequent examinations
- Is statistically "sound" (meets our statistical parameters)

Equating or Linking at the Macro Level



Equating at the Macro Level (2)



Equating at the Macro Level (3)

- Anchor items enable the direct comparison of students across current and previous examinations.
- Equating adjusts for differences in overall examination difficulty as well as relative differences in difficulty across the range of scores.
- As a result, equating ensures that a consistent standard can be maintained over time.
- Equating ensures that an 80% on the current examination is equivalent to an 80% on the previous examination.

School Reports

- School reports are best analyzed by the school staff as a group and can be used as the basis for:
  - Identifying practices that seem to be working and those that may not be working
  - Identifying areas of the Program of Studies that are well covered and those that may require greater coverage

School and examination marks (1)

- It is not expected that there would be a fixed relationship between school and examination mark at the level of the **individual** student.
- Individual student marks tend to be within 15%, only two-thirds of the time, so there will be one or two 20-point differences occurring in almost every instructional group.
- Classroom assessment covers a different and broader range of activities than can any paper-and-pencil external assessment.
- The school-awarded mark for diploma examinations also includes assessments of those Program of Studies outcomes that cannot easily be measured using paper-and-pencil instruments.

School and examination marks (2)

- The standard being reflected in both the school and the examination should be approximately the same.
- A sign of concern is if one average mark (school or examination) is well below or somewhat below provincial average, and the other average mark is well above provincial average.
- This might indicate a difference in standards.

School and examination marks (3)

- Causes of **class average** school marks being significantly higher than the corresponding examination marks may include the following:
  - too few questions set at the standard of excellence
  - too many marks given for participation and for completeness of work, rather than for quality of work
  - lack of awareness of the provincial standard for the examination, whether at acceptable standard or at standard of excellence

School and examination marks (4)

- Causes of **class average** school marks being significantly lower than the corresponding examination marks may include the following:
  - too many tasks set at the standard of excellence (approximately 20% of the items on a diploma examination are at the standard of excellence)
  - the desire to rank students in a very strong class or a very strong school makes it harder to give school marks over 85% - 90%
  - excessive penalization of students who do not hand in work regularly (more often seen in 30-2 and 33 classes)

Description of Tables (1)

Table #	Description of Table
Table 1	Final course mark
Table 2	A, B, C & F % distribution
Tables 3 & 4	Breakdown by gender

Description of Tables (2)

Table #	Description
Table 5	Exam breakdown by parts; written-response and multiple-choice in humanities, multiple-choice and numerical-response in math-science
Table 6 (humanities)	Part A % distribution by question

Description of Tables (3)

Table #	Description
Table 6 (math-sciences) Table 7.1 (humanities)	Machine-scored raw scores by reporting category
Table 7 (math-sciences) Table 7.2 (humanities)	Machine-scored item descriptions and results

Blue, yellow and pink highlights on the reports

- Use a blue highlighter to indicate results that are **well above** provincial norms.
  - Clustered blue highlights indicate strengths of an area.

- Use a yellow highlighter to indicate results that are **somewhat below** provincial norms.
  - Isolated highlights, especially yellow ones, do not normally result in action being required.

- Use a pink highlighter to indicate results that are **well below** provincial norms.
  - Clustered pink highlights, especially if this pattern continues for more than one examination, usually indicate a need for action.

Proportions

- Some results are expressed in terms of **proportions.**
- Examples of proportions are found in
  - Tables 1 and 3: proportions at acceptable and excellence
  - Tables 2 and 4: proportions getting A, B, C or F grades
  - Table 6 (humanities): proportions getting each score in the written response
  - Table 7.2: proportions getting each machine-scored item correct
- These are analyzed in the slides that follow

Criteria for blue, yellow and pink highlights (1)

- The criteria set in the following slides are based on sample sizes in the range from 20 to 80, so they work for average class sizes and mid-sized schools:
  - analysis for sample sizes below 10 is extremely risky, and is not encouraged
  - with more than 80 students, less items may be highlighted than should be

Criteria for blue, yellow and pink highlights (2)

- For school or exam marks (Tables 2 and 4 top portion only), use the following:

Score:	Criteria for **blue**	Criteria for **yellow**	Criteria for **pink**
A (80+%)	At least 10% more than province	5% to 10% less than province	At least 10% less than province
B (65%-79%)	At least 10% more than province	5% to 10% less than province	At least 10% less than province
F (0% - 49%)	At least 10% **less** than province	5% to 10% more than province	At least 10% **more** than province

Criteria for blue, yellow and pink highlights (3)

- For school results relating to the humanities written-response scales (Table 6), use the following:

Scale descriptor	Criteria for blue	Criteria for yellow	Criteria for pink
Excellent (4.5 and 5)	At least 10% more than province	5% to 10% less than province	At least 10% less than province
Proficient (3.5 and 4)	At least 10% more than province	5% to 10% less than province	At least 10% less than province
Limited and Poor (zero,1.0, 1.5 and 2)	At least 10% less than province	5% to 10% more than province	At least 10% more than province

Criteria for blue, yellow and pink highlights (4)

- For school results for individual machine-scored items (Table 7 or 7.2) use the following:
  - At least 10% higher than the province; **blue**
  - Between 5% to 10% lower than the province; **yellow**
  - At least 10% lower than the province; **pink**

Scores

- Some results are expressed in terms of scores on a test, subtest, or reporting category (i.e., the mean and standard deviation are also given then)
- Examples of scores are found in:
  - Tables 2 and 4: school, examination and blended mark averages (bottom of the table)
  - Table 5: scores on written-response, multiple-choice, and numerical-response
  - Table 7.1: scores on different reporting categories
- These are analyzed in the slides that follow

The use of *z*-scores

- *z*-scores tell how many standard deviations away from the mean a score resides.
- *z*-scores can be positive or negative:
  positive (+) *z*-score = value is above the mean
  negative (–) *z*-score = value is below the mean

Why we need to use *z*-scores

- Subtests and reporting categories are of different lengths and difficulties; *z*-scores make all subtests and reporting categories of equivalent lengths and difficulties.
- We compare apples with apples, with the school and the province being compared on the same measure.

z-score as a criterion for subtests (1)

- To measure school values for subtests, the best way is to use *z*-scores. The *z*-score is defined as follows:

$$z = \frac{(\text{school mean} - \text{provincial mean})}{(\text{provincial standard deviation})}$$

- *z* will be **negative** when school averages are below provincial, and **positive** if school averages are above provincial.

z-score as a criterion for subtests (2)

- For school results for reporting categories and examination components (Tables 2, 4, 5, and 7.1), use the following:
  - *z* greater than or equal to 0.50; highlight blue
  - *z* between –0.25 and –0.50, including –0.25 ; highlight yellow
  - *z* less than or equal to –0.50; highlight pink

z-score as a criterion for subtests (3)

- The critical values for *z* depend strongly on the sample size, being higher for smaller classes. For a sample size of 1, the critical value would be 1.96 for significance. These are shown in Sections V and VI of the workbook.
- The rule of thumb given in the previous slide will have the pink and the blue significant in a sample of 16, and the yellow significant in a sample of 64.
- Alberta Education does not calculate these *z* values in the school reports, as they are rules of thumb, not the full significance tests needed in the publication of legal documents like the school and jurisdiction reports. Teachers are encouraged to calculate them.

Notes from Case Study #1 (Applied Mathematics 30): *n* = 33

- The school is a suburban high school with most of its students either coming in by school bus or driving themselves.
- It serves a group of suburbs, but is located outside any particular suburb, with few students within walking distance, and with no public transit to the school.
- Most students have the choice of going to this school or to other high schools that are part of either suburban centres or metropolitan districts.

Notes from Case Study #1 (Applied Mathematics 30): *n* = 33

- **Table 2** shows very similar results on the school mark and the exam mark, indicating that the standards at school and exam were similar, and that according to either mark, this class was very slightly below provincial average. The exam results showed a bimodal effect, with very few marks between 65% and 79%.

- **Table 4** shows two gender effects. The girls did better than the boys on both school and exam marks. However, the boys' marks improved on the exam, while the girls' marks declined from school to exam.

Notes from Case Study #1 (Applied Mathematics 30): *n* = 33

- **Table 5** shows a marked difference in performance between the student-selected multiple-choice questions and the student-constructed numerical-response questions. Performance was much better on the selected responses.

- If this pattern has been seen before, it is something to be concerned about, but if this is a one-off, it is not worth taking action on.

Notes from Case Study #1 (Applied Mathematics 30): *n* = 33

- **Table 6** shows considerable variation from one unit to another, with great success on the matrices and pathways unit. There was less success on the cyclic patterns unit, and on the statistics and probability unit.

- **Table 6** also shows greater success in conceptual understanding and problem-solving, with less success in procedural knowledge.

Notes from Case Study #1 (Applied Mathematics 30): *n* = 33

- **Table 7** amplifies the results found in Table 6. The slight weakness of the performance can be seen in the presence of only 6 blue highlights, as compared to 4 yellow and 9 pink highlights.

- The distribution of blue highlights is irregular, with more blue highlights in the strongest unit (matrices and pathways) and fewer in the weaker units (statistics and cyclic patterns).

- The large number of highlighted items (19 out of 40) shows that either the whole group gets a concept or very few get the concept; this is consistent with the small number of marks between 65% and 79%.

Notes from Case Study #8 (Pure Mathematics 30): *n* = 65

- This school is a large high school in a metropolitan area, situated in a somewhat lower-income area of the city.
- All courses are offered in both semesters, with several sections of each course being offered in each semester.

Notes from Case Study #8 (Pure Mathematics 30): *n* = 65

- **Table 2** shows a very small difference provincially between school and examination marks. The average drop is 6.7% provincially. At this school, the gap is much narrower at 2.5%.

- The school-awarded marks were significantly lower than the provincial average at this school, with the examination marks coming in somewhat higher than could be predicted from the school mark. This shows an understanding of the standards implied by the program of studies and the assessment standards.

Notes from Case Study #8 (Pure Mathematics 30): *n* = 65

- **Tables 3 and 4** show a poorer performance from girls, especially on the school-awarded mark.
- The drop from school mark to exam mark was very similar for girls (2.6%) and for boys (2.3%).
- At the acceptable standard, the difference between boys' performance and girls' performance was especially marked, both for school marks and for exam marks. For both school and exam marks, there were double the percentage of failures for girls compared to boys.

Notes from Case Study #8 (Pure Mathematics 30): *n* = 55

- **Table 5** shows that, at the provincial level, the numerical-response questions are somewhat harder than the multiple-choice questions. This difference (66% MC, 49% NR) is partly due to the use of NR questions to replace what was covered in the former WR questions.

- At this school, the drop from MC to NR was equally marked– 59% for MC and 41% for NR.

Notes from Case Study #8 (Pure Mathematics 30): *n* = 55

- **Table 6** shows much stronger results on the transformations and conics units, and weaker on the permutations and combinations unit and the statistics unit. This could be the result of difficulties in completing the course in the time allotted, leaving insufficient time for providing sufficient practice, especially in permutations and combinations.

- The same weakness showed up in the lower results in procedures and in problem-solving, as compared to conceptual understanding . The two stronger units are more attuned to conceptual understanding than are other units.

Notes from Case Study #8 (Pure Mathematics 30): *n* = 55

- **Table 7** amplifies the results shown in **Table 6**, and gives further details.
- The results in transformations and in conics showed 1 blue and 2 pink highlights out of 11 questions, while permutations and combinations and statistics showed 7 pink and 2 yellow out of 11 questions. What was relatively successful in the transformations and conics, and why did it not occur for permutations and combinations and statistics?
- Was it because transformations and conics are more visual and less abstract?

Producing Narrative from a Report

- The numbers alone, even when the highlights have been added, are only the first step in the analysis process.
- The data in any report should be converted to a narrative, table by table, in the same way as was shown in the case studies included in this presentation.
- The narrative should include both areas of strength and areas of improvement.
- If a table (often Tables 3 and 4 on gender) shows nothing worthy of comment, do not stretch the data to manufacture narrative.

Going from Narrative to Explanations and Hypotheses

- There is almost always a reason for any significant point in the narrative, and here teachers use their knowledge of their students, together with reflections on their experiences with the course, to look for reasonable explanations.

- Some explanations may be quite tentative, and require no immediate actions; other explanations would need to form the basis of an immediate action plan.

Going from Explanations to Action Plans: Example 1

- A small school that taught English and Social Studies 20/30 in alternate years had poor results in Social 30 Part A
  - Grade 11 students with only English 10-1 performed poorly
  - Grade 12 students with English 30-1 were above acceptable standard

Going from Explanations to Action Plans: Example 1

- Immediate action was taken by having a combined Social 20/23 taken in Grade 11, followed by a combined Social 30/33 in Grade 12.
- The same was done for English, thus ensuring that students taking Social Studies had sufficient practice in extended writing (from English) to be able to apply these skills to the issues and contexts of Social Studies.

Going from Explanations to Action Plans: Example 2

- One school had much stronger results in the critical/analytic response to literature in English 30 – 1, both in the long essay and the analytic questions in the machine-scored.

- Results were far weaker in the personal response, both in the shorter essay and the connection of literature to self in the machine-scored.

Going from Explanations to Action Plans: Example 2

- The teachers found that they employed peer editing freely as part of the writing process for critical/analytic essays, but did not use peer editing for personal essays, citing privacy concerns.

- They then worked with their students on the concept of a learning community, so that students accepted that information gained during the peer editing process remained confidential. As a result, they introduced peer editing to personal, as well as critical essays.

Going from Explanations to Action Plans: Example 3

- Physics 30 had a new curriculum, which included relatively minor changes to content, but major changes to process. The first large-scale exam was in January 2009.

- A school that usually was somewhat above provincial average in Physics 30 found that their January 2009 results on questions that related to the changed content and process outcomes were poor; the results on the carry-over outcomes were very good.

Going from Explanations to Action Plans: Example 3

- The teachers came to the conclusion that they had underestimated the process changes in the new Program of Studies, and worked as a team to remedy this.

- Their results were good in June 2009, and they have shared their experiences with the Social Studies teachers in their school, so that their students will be on top of the new Social Studies examinations in January 2010.

Locating the multi-year reports

- Unlike the school and jurisdiction reports for a particular examination, the multi-year reports for each school and jurisdiction are publicly available from Alberta Education's website.

- The pathway is as follows:

  Alberta.ca>Education>Administrators>Provincial Testing>Diploma Examination Results – Multiyear Search Feature

- The multi-year feature is **not** available until 5 years of data has accumulated for a particular Program of Studies.

Using the multi-year reports (1)

- Local school jurisdictions may define their Grade 12 population differently from Alberta Education; Alberta Education considers all students in their third year of high school as first-year Grade 12s.
- The participation rates are based only on third-year high school students so that the presence of repeaters or grade 11s does not affect either the school data or the provincial data. The participation of Grade 11 students is counted when they reach their third year of high school.
- Participation rates for a typical school should be fairly similar to provincial norms.

Using the multi-year reports (2)

- A relatively high participation rate does **not** necessarily imply a lower performance in 30 and 30-1 subjects; many successful schools are able to have high (but realistic) participation combined with high performance.
- Anomalies to look for include unusually low participation rates in English 30-1 and/or unusually high participation in English 30-2 and Social 30-2.

Using the multi-year reports (3)

- The subjects with the most stability in enrolments tend to be English 30 – 1 and Chemistry 30; fluctuations tend to be greatest in Physics 30, Science 30, and Applied Mathematics 30.

- Trends to watch for in individual subjects are:
  - Ratio of exam marks over 80% to school marks over 80%: in most subjects this ratio is about two-thirds, except for English 30 – 2, Social Studies 30 – 2 , Applied Mathematics 30 and Science 30, where it is closer to 1:1, and English 30 – 1 where it is about one-third
  - Exam pass rates: normally about 75% - 85% in most subjects

Using the multi-year reports (4)

- Trends to watch for in a school or a jurisdiction are:

  - Percentage of third year students completing graduation requirements in English Language Arts and Social Studies. Provincially, these are 86% in English and 84% in Social Studies, and they have been trending upwards over the last five years.

  - Any sudden increases or decreases in enrolment in any particular course.

Ways to Get Involved

- Working Groups – item writing, technical reviews, development for new Program of Studies, etc.
- Marking (Humanities and Achievement)
  - Nominated through the superintendent's office
  - You must be nominated for each list separately as the criteria for involvement are different

Other Consortia Workshops

Sources of further information

- For further information contact:

  Assessment Sector at 780-427-0010.
  To call toll-free in Alberta dial 310-0000.

- Internet address education.alberta.ca

we engage
engager