

Risking validity: The use of course marks as school effectiveness indicators  
University of Toronto/OISE - Saad Chahine  
Presented at ICSEI 2006

Abstract:

Student performance, credit accumulation and course marks are school effectiveness indicators that are used in research and in practice. Often these indicators are based on teacher interpretation. This is of great concern due to the varying parameters in teacher judgment. This paper provides method to measure the quality of these indicators by examining the validity of teacher-generated exams. The information in this paper was developed through a quantitative study based on two data sets: teacher-generated mathematics exams and an online questionnaire. The findings identify the possibility of validating these indicators and provide a scheme to improve quality.

---

Student performance is a primary indicator of school effectiveness, which is produced by taking an account of test results (Rowe, 2000; Wolford, 2002,). Traditionally testing methods are externally developed, administered and scored through a rigorous process, to ensure a high degree of accuracy. Although accurate there is heated debate over validity of large scale assessment for school effectiveness purposes. Coe & Fitz-Gibbon (1998), describe the rationale for this argument by Carver (1975), “Tests that did not test that which teachers were teaching could hardly be considered a measure of ‘effectiveness’ – at least if a distinction is accepted between aptitude and achievement (p. 422). This junction that unites instruction and testing, is of great concern to nations, states/provinces and districts, as the methods employed often do not meet “instructional remedies” risking the quality decision making process (Baker & Linn, 2004).

Recently we are witnessing a phenomenon, to utilize locally developed assessments and more notably aggregated course marks in school effectiveness research (King, Warren, Boyer, & Chin, 2005). This paper, focus on this recent phenomenon and examines it through a research study based on two interrelated foci. The first is the validity of teacher-generated examinations and the second the validity of locally developed marks as indicators of school effectiveness. This paper will take you through the concept of validity, the use of performance indicators in school effectiveness, validity of locally developed assessment and the potential risk on education.

### Concept of Validity

Validity is often a slippery concept that may be interpreted differently by educators, measurement specialists, policy makers and general public. In this paper I distinguish validity as a quality of interpretations made from student performance results. The evaluation of quality is determined on the extent to which the purposed use of results matches the intended purpose of the performance measure through the construction of a resultant score. This concept is based on Standards for educational and psychological testing, where “Validity refers to the degree to which evidence and theory support the interpretations of test scores entangled by proposed uses of tests” (American Educational Research Association [AERA], American Psychological Association, &

National Council on Measurement in Education, 1999, p. 9). This modern view of validity is largely based on Samuel Messick’s research, where he defines validity

holistically; and asserts that varying facets are parts of validity; this includes evidence of test construction, relevance/utility, value implication and social consequences (Messick, 1995, 1989, 1980). Their presence and cohesion determines the degree to which an interpretation is deemed valid.

To add to the slipper concept, when using non-standardized assessments, the lines between reliability and validity are less defined (Moss, 1994). Thus the new trend in using course marks as effectiveness indicators coupled with a low accuracy and quality is a risk. Decisions based on these results need be carefully considered. To aid the situation, Baker & Linn (2004) present two main questions to be asked in evaluating validity: “Is the definition of the content domain to be assessed adequate and appropriate? Does the test provide an adequate representation of the content domain it is intended to measure” (p. 52)? These questions form structure for this study; where I examine the relationship between content suggested by the ministry of education, content present teacher-generated grade 9 mathematics exams. This exploration provides a method analysis to measure the validity of locally developed examinations and the potential use and risk of the use of course marks as effectiveness indicators.

### Use of performance measures in school effectiveness

Standardized testing for purposes to monitor and measure school effectiveness are globally employed with varying parameters innate to countries, regions and states/province. In the US through the No Child Left Behind Act obligates states to provide an account of student performance, UK examinations are publicly reported in league tables in newspapers, and Australia is utilizing a sophisticated moderation system. There are numerous examples, yet the impact of these methods on society, are slowly evolving and this has brought the validity debate to the forefront.

I believe, the relationship between social implications of testing and the purpose of testing are often not in agreement, according to Messick (1980) using of a Singerian system of inquiring discussed by Churchman (1971), validity is at risk. One test is not capturing how individual school are effective. Wolford (2002) discusses this notion in light of the situation in the UK "Quite simply, while league tables purport to reflect the effectiveness of schools, this is actually masked by the effects of the intake characteristics of the students to each school" (p.48).

In Ontario, Canada student performance is measured by the Education Quality Assurance Office (EQAO), an arms length agency to the Ministry of Education, with multiple proposes of testing. A recent report identified EQAO mandate as:

...[developing] tests... [and] undertake the administering and marking of tests of pupils in elementary and secondary schools...[and] report to the public and to the Ministry of Education and Training on the results of tests and generally on the quality and effectiveness of elementary and secondary school education and on the public accountability boards (Wolfe, Childs & Elgie, 2004; also see Education Quality and Accountability Office Act, 1996, S.O. 1996, c. 11).

Based on this mandate, Wolfe et al (2004), state there are three main purposes of standardized testing that emerge: 1. Report on results of the test, 2. Report of the quality and effectiveness of education, 3. Report to accountability boards. In Ontario there are

not overt consequences from testing with exception the Ontario Literacy Test, which is a requirement of graduation. Wolfe et al (2004), clearly state, "...the most important purpose of the grade 9 assessment seems to us to be the production of results to help schools plan instruction" (p. 9). According to Newman, King and Rigdon (1998), we have an incomplete system of accountability as we do not identify who is held accountable, under what consequences.

Understanding Canadian society, values and beliefs that have evolved to produce a unique system of accountability is crucial to appreciating the social implications of the testing results. Although not spelled out one group of people who are affected by these results are teachers. However there is lacking evidence in how the exam results are useful. As well, the results from tests should help teachers develop more coherent and capable instruction and assessment practices. Yet, there are not overt consequences to the assessment, but consequences for students based on assessment that occurs in the class. This is where our system of assessment differs as there are low stakes for standardized assessments but high stakes for non-standardized assessment. It is ironic that our accurate assessments have little impact on students and the test which are not measured for accuracy, have a great impact on student. A disharmony will exist when we begin to use course marks as effectiveness indicators with a disregard to the measurement concepts that ensure EQAO results are generalizable.

#### Validity of locally developed assessment and course marks

Local and provincial systems of assessment are bound by issues that are beyond the scope of education. Our society's value system dictates a large part of what we consider to be the function of assessment. Broadfoot (1996) states, "Even in the most simple societies, children must be trained and subsequently demonstrate competence in the appropriate forms of behavior and skills required by all members of that society" (p. 26). Consequently, testing is not only based on the prescriptions of external organizations and schools. Its function goes deeper. Testing is a manifestation of our need as a society to ensure competency has been measured. The current approach is not meeting the needs of individual schools, yet there are methods that may be employed to capture school effectiveness at the classroom level.

A method that uses classroom assessment as school effectiveness indicators is evident in Nebraska: "Districts would provide information to an appropriate agency (e.g. state department of education, office of accountability) that describes student performance on the district assessments" (Buckendahl, Impara, & Plake, 2002, p. 8). Nebraska's system of accountability is not test-based but rather assessment-based. Districts present to State departments evidence of: "students' performance on a district's assessments, the technical quality of that district's assessments, and selected noncognitive indicators of student performance" (Buckendahl, Impara, & Plake, 2002, p. 8). Many pieces of evidence are taken into account before passing judgment on a school's or a district's effectiveness. As well a similar yet, less structured occurrence is developing in the Los Angeles Unified School District (Baker & Linn, 2004). These developments in the US, is what might be considered a hermeneutic approach to assessment. Moss(1994) describes:

A hermeneutic approach to assessment would involve holistic, integrative interpretations of collected performances that seek to understand the whole in light of its parts, that privilege readers who are most knowledgeable about the context in which the assessment occurs, and that ground those interpretations not only in the textual and contextual evidence available, but also in a rational debate among community of interpreters (Moss, 1994, p. 7)

Thus the educational system is facing new, exciting and different methods to measure school effectiveness with a much more depth in meaning. However, both Nebraska & the district in California have a structure designed to ensure comparability between school without hindering on individual teacher instruction or assessment. This initiative to ensure standardization is missing within Ontario. Thus, invalid interpretations are made comparing course marks for school effectiveness purpose. The follow provides evidence to the variability that exists between schools.

### Validity of teacher generated mathematics exams

I recently studied the validity of teacher generated grade 9 mathematics exams in Ontario. This study was developed through an amalgamation of various concepts and literature in the assessment field to create a methodology that operationalizes validity. I collected a data set through an alignment analysis of teacher generated grade 9 mathematics exams. The information was gathered from in three independent schools in Ontario who teach and test the same mathematics course: MPM1D – Principle in Mathematics.

#### *Data gathering*

Mathematics exams were coded and an alignment analysis was conducted to measure the content of the exams. Alignment models are more popular in the US for large scale testing purposes and may be defined as “...how well all policy elements in a system work together to guide instruction and ultimately, student learning” (Rothman, Statter, Varnek & Resnick, 2002, p. 5; Webb , 1996). Presently, there exist three models of alignment that are accepted and advocated by the Council of Chief States School Officers, ([CCSSO], 2002). The first model is the Webb, named after the developer Norman Webb of the Wisconsin Center for Educational Research, with the assistance of the CCSSO. The second model is the Surveys of Enacted Curriculum, (SEC), which was developed by Andrew Porter and John Smithson through the Wisconsin Center for educational research and also with assistance from CCSSO. The third model of alignment is the Achieve model, designed by a non-profit educational leadership organization in Washington DC (CCSSO, 2002).

All the models are based on common principles, the SEC model, created by Porter and Smithson, is most adaptable to the classroom environment. This is because it is founded on a commonality between policy, teachers, and testing, versus the other models that only consider policy and testing. The SEC model also stems out of three phases of

curriculum: intended, enacted, and assessed curriculum which may be more familiar to teachers , (Smithson &Porter, 2004, p. 110).

### *Findings & Discussion*

#### *Construct Underrepresentation & Irrelevance*

Construct Underrepresentation refers to the degree in which a test does or does not fully represent what was learned and/or what was supposed to have been learned. This often occurs because the items chosen to be on the test are not a complete sampling of content “types of content, [or] engage some psychological processes, or elicit some ways of responding that are encompassed by the intended response”(Webb, 1996). Using the SEC model a frequency count was conducted to measure the number of times a strand appeared on the examination. There are four strands within the grade 9 mathematics course: Number Sense and Algebra (NSA), Relationships (RTL), Analytic Geometry (AG), and Measurement Geometry (MG). The following table displays the percentages of these strands covered within the exams of three schools.

*Table 1*

*Percentages of content present on MPMID exams*

Strand	Red School	Blue School	Yellow School
NSA	41	53	54
RTL	26	16	12
AG	30	25	27
MG	3	6	7

The table show an order of that is common in all exams: NSA, AG, RTL and MG. Notably MG may be considered a underrepresented strand of these exams, when compared to the other strands. The most frequently tested strand on all the exams is NSA. Summative exams are intended to represent a sampling of course content throughout the year. This basic descriptive analysis raises the question of “balance” and strands that may exist with the instruction of the course.

None of the exams identified questions that would incorporate various factors that may give some student advantage over others within the mathematics curriculum. Thus construct-irrelevance, is not an issue of concern. However, there can be several reasons for this. For example, the rater did not find any irrelevance in the testing or the items applied a basic understanding and thus were applicable to most populations taking the grade 9 mathematics course.

#### *Evidence based on test content*

Evidence based on test content is a component in determining the quality of interpretations that may be made from results. Test content analysis is a “...logical or

empirical analysis of the adequacy with which the test content represents the content domain and the relevance of the content domain to the proposed interpretation of test score” (AERA et al., 1999, p. 11). Two aspects were examined, how well do the exams represent what is expected to be tested by the ministry of education and how well may they represent the instruction that occurs in schools. According to policy, the ministry of education expects a “balance” of the strands to be represented in the examinations.

It is expected that in developing detailed courses of study from this document, teachers will weave together related expectations from different strands, in order to create an overall program that integrates and balances concept development, skill acquisition, and applications (Ministry of Education and Training, 1999, p. 5)

In curriculum documents, educators are provided 3-4 overall expectations per strand. Yet there is very little structure around how much each strand should be emphasized. I suspect this is purposefully done as to not make the curriculum narrowed and prescriptive. However, they have left little structure hindering the comparability from one school to the next.

I conducted two types of chi-square goodness-of-fit tests. The first examines the “balanced” of strands. The second examines how well the distribution of strands fits two separate teachers instructional distribution of strands based on findings of Ben Jaffar (2006). When the test was performed with equal proportionality for each strand the findings showed that an equal distribution does not exist: Red School  $\chi^2 (3, N = 69) = 21.03, p < .05$ , Blue School  $\chi^2 (3, N = 64) = 31.50, p < .05$ , and Yellow School  $\chi^2 (3, N = 41) = 21.34, p < .05$ . This may show that all the schools do not identify that they have an equal distribution amongst strands and thus if “balanced” was intended to be defined as equal representation in assessment, this is not the case.

When compared with the distributions presented in Ben Jaafar (In press) the findings were mixed. Ben Jaafar (2006) identified teacher 1 to have a distribution of: NSA=47.7%, RTL=7.6%, AG=21.9%, and MG=18.1% and teacher 2 with a distribution of: NSA=46.6%, RTL=8.1%, AG=24.5%, MG=14.2%. the Red school(  $\chi^2 (3, N = 69) p < .05$ ) all showed that teacher 1 instruction portions describe the proportions on the exams (Blue  $\chi^2 (3, N = 64) = 7.11 p > .05$ ; Yellow: (  $\chi^2 (3, N = 41) = 2.39 p > .05$ ). However when comparing the school exams to teacher number 2, except for the Yellow School ( $\chi^2 (3, N = 41) = 4.23 p > .05$ ) all showed that teacher 2 instruction proportions do not describe the exams (Red:  $\chi^2 (3, N = 69) = 40.70 p < .05$ ; Blue:  $\chi^2 (3, N = 64) = 10.41 p < .05$ ). The findings are very interesting, particularly in the case of the Yellow School, since that it has a similar distribution to the instructional practices to two teachers in two separate schools. This means that there some commonality exists between what teachers instruct & test in different school. Nevertheless, there remain significant differences in the other cases.

This is very exciting, if we are aiming for introducing locally developed assessment as an indicator of school effectiveness, we have a chance, as there potential of commonality in practice between schools, in relation to instruction and end-of year examinations. However, evidence based on test content is only one part of validity and we could not conclusive validate assessment based only in this measure. As well without

the clarity of expectations from the ministry of education, we cannot state that these exams measure the “balance” that the ministry is expecting.

*Evidence based on internal structure*

Internal Structure of a test analyzes the degree to which the construction accurately denotes what the test is attempting to measure. SEC calls this Range and Breadth and is comparison of the relationship between “matter range identified in the content standards and the range of topics represented by a particular test” (Porter, 2004). This is a concept that examines whether the test underpinnings are the same as those of the ministry expectations, by examining the breadth of coverage and the weight of certain topics in comparison to others. This may be a better indicator of what is important on the exam as the less evident content may possess a greater mark. The table below identifies the distributions in table 1 and incorporates a weighing of the marks of each strand. This was accomplished by identifying how many marks each strand is worth then developing a percentage score of comparability with distribution. After I developed the weighing, I subtracted the weighted score from the un-weighted score and if the difference appears positive, the construct or strand is tested more often, but worth less marks, and the case is the opposite when difference is negative.

*Table 2*

*Weighed proportionality and difference*

Strand	<u>Red School</u>			<u>Blue School</u>			<u>Yellow School</u>		
	Un-weighted (x)	Weighed (y)	Difference (x-y)	Un-weighted (x)	Weighed (y)	Difference (x-y)	Un-weighted (x)	Weighed (y)	Difference (x-y)
NSA	41	28	13	53	44	9	54	41	13
RTL	26	37	-11	16	18	-2	12	11	1
AG	30	33	-3	25	28	-3	27	33	-6
MG	3	1	2	6	11	-5	7	15	-8

*Note:*

*If Difference is positive more frequent occurrence of content worth less marks*

*If Difference is negative less frequent occurrence worth more marks*

*The closer difference is to 0 the more the weight is proportional to the un-weighted strand*

This analysis identifies commonalities between schools based on two main strands, NSA & AG. NSA is continuously being measured more frequently but worth less then occurs and AG is tested less and worth less marks then NSA, yet it is worth more then it appears. While the remaining two strands showed mixed results between groups. I believe there may be a few things at play here. From my experience as a mathematics teacher, I have often found that courses are heavily focused on NSA, and often cater to lower level cognitive abilities in NSA, items based on AG often are a main priority of the course and are often based on higher cognitive level of thinking, making them worth more marks. This is my personal speculation from experience teaching the course,

empirical evidence is needed to examine evidence of the response process, this is where research needs to expand in the future, to grasp a clearer understanding of the validity of teacher-generated testing.

### Potential Use & Risk on Education

The original idea for this research was to be utilized within my own teaching practices to see how other schools “do it.” After understanding measurement concepts and reading school effectiveness research. There needs to be measures of how schools are “differentially effective.” I agree with Wolford (2002), “If the differential effectiveness of schools is to be assessed (and this may actually vary with different types of children), then some measure of school or student progress has to be used and meaningful comparisons can be drawn” (p. 48). Moreover I also believe that there will be tension in attempting to make a cohesive system of instruction tied to standards and aligned to assessments.

What are potentially facing is what Katz, Earl and Olson (2001) describe as a *Paradox in classroom Assessment*; where teachers are always in struggle to assess the what is expected by a central authority, which may create a struggle with their own professional ethos to test individual students. Assessment in the classroom can not be considered a closed system, students are not the same and do not travel from one junction to the next learning what is expected. There is also a greater problem at risk here and it is not only the construction of examinations, aggregation of marks to develop a final mark for a student.

Through my experience teaching, and as a student, have witnessed a statistical anomaly within classrooms. A good example comes from one of my former students, Lincoln. Lincoln has some behavioural issues climbing walls, jumping up and down and being very distractible for his fellow students. It took me the first semester of a school year to finally break through to Lincoln and make sure that he has completed home work on time, writes his tests, finishes projects ect, which he would never do previously. I found out that Lincoln was an exceptional mathematics scoring in high 90’s on much assessment towards the end of the year. However when it came time for me to give him his final grade, I averaged all of his marks and he had failed. His mark was not representative of his capability. I found this to be the case with many of my colleges who at the end of the year average a mark & hesitate to present it to a student as the consequences are sometimes detrimental to their belief in self capacity.

I am not sure how averaging a course assessment has become the preferred method of aggregation, to develop a final mark for a student. What is alarming is that, if this was research and a mean was taken of sample size of 10-30 participants, it would be considered poor research. In Ontario we have gone a step further and aggregate 30 or so students’ marks to develop a class average and plot the distribution of marks for all students in Ontario. How has it come to be that we as educators and researchers accept the arbitrary aggregation of inaccurate testing as an indicator of school effectiveness for the decision making process? Decision making based on classroom results can be greatly enhanced by simply questioning the quality of these results, I urge you to do this, otherwise we are taking great risks with children’s futures. It is obvious that more research within the area of classroom assessment is needed as it is a relatively unknown territory in accountability structures.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Baker, E. L., & Linn, R. L. (2004). Validity issues for accountability system. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 47-72). New York, NY: Teachers College Press.
- Ben Jaafar, S. (2006). An alternative approach to measuring opportunity-to-learn in high school classes, *Alberta Journal of Education Research*, 52(1).
- Broadfoot, P. (1996). *Education, assessment and society*. Buckingham, United Kingdom: Open University Press.
- Buckendahl, C. W., Impara, J. C., & Plake, B. S. (2002). District accountability without a state assessment: A proposed model. *Educational Measurement, Issues and practice*, 21(4), 6-16.
- Buckendahl, C. W., Plake, B. S., & Impara, J. C. (2004). A strategy for evaluating district developed assessments for state accountability. *Educational Measurement, Issues and Practice*, 23(2), 17-25.
- Council of Chief States School Officers, (2002). *Models for Alignment Analysis and Assistance to States*: <http://www.ccsso.org/content/pdfs/AlignmentModels.pdf>
- Churchman, C. W. (1971). *The design of inquiring systems*. New York, NY: Basic Books Inc.
- Coe, R. & Fitz-Gibbon, T. (1998). School effectiveness research: criticisms and recommendations. *Oxford Review of Education*, 24(4), 421-38.
- Earl, L. (1999). Assessment and accountability in education. *Education Canada*, 39(3), 4-8.
- Education Quality and Accountability Office. (2005). *Grade 9 assessment of mathematics, framework*: [http://www.eqao.com/pdf\\_e/05/05P014e.pdf](http://www.eqao.com/pdf_e/05/05P014e.pdf).
- Katz, S., Earl, L., & Olson, D. (2001). The paradox of classroom assessment: A challenge for the 21st century. *McGill Journal of Education*, 36(1), 13-25.
- King, A., J.,C., Warren, W.,K., Boyer, J.,C., & Chin, P. (2005). Double cohort study: Phase 4 report. Ministry of Education/Ministry of Training: <http://www.edu.gov.on.ca/eng/document/reports/phase4/report4.pdf>
- Messick, S. (1980). Test validity and the ethics of assessment. *American psychologist*, 35(11), 1012-1027.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5-11.

- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741-749.
- Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, 45, 35-44.
- Ministry of Education. (2005). *The Ontario curriculum grades 9 and 10: Mathematics (revised)*: <http://www.edu.gov.on.ca>
- Ministry of Education and Training. (1999). *The Ontario curriculum grades 9 and 10: Mathematics*: <http://www.edu.gov.on.ca>
- Moss, P., A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement, Issues and practice*, 22(4), 13-25.
- Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23(2), 5-12.
- Porter, A., C., Chesterer, M.,D. & Schlesinger, M., D. (2004). Framework for an effective assessment and accountability program: The Philadelphia example. *Teachers College Record*, 106(6), 1358-1400.
- Rowe, K. J. (2000). Assessment, league tables and school effectiveness: Consider the issues and 'let's get real!' *Journal of Educational Inquiry*, 1(1), 73-98.
- Smithson, J. L., & Porter, C. A. (2004). From policy to practice: The evolution of one approach to describing and using curriculum data. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: The 103rd yearbook of the national society for the study of education, part 2* (pp. 105-131). Chicago, Il: University of Chicago Press.
- Webb, N. L. (1997). Research monograph no. 6: Criteria for alignment of expectations in mathematics and science education. National Institute for Science Education University of Wisconsin-Madison & Council of Chief State School Officers Washington, DC:  
<http://facstaff.wcer.wisc.edu/normw/WEBBMonograph6criteria.pdf>
- Wolfe, R., Childs, R., & Elgie, S. (2004). *Final report of the external evaluation of EQAO's assessment processes*. Ontario Institute for studies in education of the University of Toronto.
- Wolford, G. (2002). Redefining school effectiveness. *Westminster Studies in Education*, 25(1), 47-58.