Educ Asse Eval Acc (2009) 21:329–345
DOI 10.1007/s11092-009-9083-1

Why school based assessment is not a universal feature of high stakes assessment systems?

Iasonas Lamprianou · Thomas Christie

Received: 13 September 2008 / Accepted: 17 July 2009 /
Published online: 2 September 2009

Abstract Accepting that school based assessment may have the potential to bring additional reliability to the assessment outcomes of an educational system, this research uses Generalizability Theory to address the question "why school based assessment is not a universal feature of high stakes assessment systems"? Three major issues are identified: (a) there is a conflict between the psychometric model and classroom assessment practice; (b) different schools are not equally effective; and, (c) teachers' judgments are frequently accused of being biased. The role of public examination boards is discussed in this context.

Keywords School based assessment · External assessment

1 Introduction

It is widely held that forms of assessment are very powerful determinants of forms of teaching and learning (e.g. Kellaghan and Greaney 1992; Broadfoot 1996), always providing that teachers are both willing and able to change their classroom practices. Radical changes in every day classroom practice cannot be imposed. It follows that an effective strategy for implementing centrally directed educational reforms is to embed them in assessment-led initiatives, especially if the assessment is high-stakes. This proposition accounts in large measure for the immediate attractiveness of the introduction of school- based assessment (SBA) as an engine of educational change. The publication of standardising material to guide teacher judgment acts as a covert

I. Lamprianou
School of Education,
University of Manchester, Manchester, UK

I. Lamprianou (✉)
Department of Education Sciences,
European University-Cyprus, P.O. Box 22006, Nicosia 1516, Cyprus
e-mail: iasonas.lamprianou@man.ac.uk

T. Christie
Aga Khan University Examination Board, Karachi, Sindh, Pakistan
e-mail: thomas.christie@aku.edu

indication of how teaching practices might be extended or redirected. Giving the teacher responsibility for assessment increases 'ownership' of the changed practices with beneficial consequences for willingness to change. And finally there is plentiful meta-analytic evidence that relatively prompt feedback by a teacher contributes to higher achievement (Bangert-Drowns et al. 1991; Kluger and DeNisi 1996) even if the strength of positive results is conditional upon both the nature and the social context (Black and Wiliam 1998) of the feedback.

School-based assessment initiatives are to be found in the public school systems of a number of countries, usually supported by moderation procedures designed to increase the comparability of the reported outcomes. According to the Queensland Board of Senior Secondary School Studies (2003), a major aim of the moderation task is to provide validity, reliability, credibility and acceptability to the SBA. In England, school-based assessment systems in the form of Teacher Assessment (TA) were established more than a decade ago after the introduction of the National Curriculum. TA is considered to be

> 'an essential part of the national curriculum assessment and reporting arrangements. The results from teacher assessment are reported alongside the task and test results. Both have equal status' (QCA 2002, p.10)

It is not the system of choice, however, when it comes to certification, licensure, entrance to universities and public accountability. In the UK, the 'hard evidence' from the National Curriculum tests is usually treated as more 'objective', standardized and, therefore, more trustworthy than Teacher Assessment, i.e. SBA (Reeves et al. 2001).

Although SBA can be the vehicle to promote improvements in the educational system and add reliability to its assessment procedures, SBA is not a universal feature of high-stakes assessments. This study investigates potential reasons why using generalisability theory (Cronbach et al. 1972).

2 The aims of the study

School-based assessment has been widely held to have the potential to bring additional reliability to the assessment outcomes of an educational system. One would, therefore, expect SBA to be a universal feature of high stakes assessment systems. Why is this not happening? From the perspective of an examination board, there are three main barriers to adoption;

- conflict between the psychometric model and classroom assessment practice,
- differences in school effectiveness,
- bias in teachers' judgments.

Each of these barriers is explored below, seriatim. The conclusion synthesizes and summarizes the findings of the study.

3 The dataset

Wherever appropriate, the study calls on a comprehensive data set of all assessments made at Grade 12 (18 year olds) in 31 timetabled subjects in 23 schools in a

European country. In all timetabled subjects teachers returned three assessments of each student, one per term (i.e. there is a three months interval between each one of the three assessments). There were no formal external requirements of how these assessments should be arrived at. The 31 subjects may be theoretically split in two groups: The first group consists of 18 timetabled subjects where there was an externally set and marked end of year examination. The second group consists of 13 timetabled subjects where there was no externally set and marked end of year examination. Other than that the subjects of the two groups are identical; for example, both groups consist of science, mathematics, history, chemistry, biology, languages subjects etc. The same subject in the two groups (with and without external exam) are taught essentially by the same teachers, using the same textbooks, in the same building, sharing the same resources etc. At the beginning of the year, each student would choose to take external exams only on a number subjects.

Both internal and external assessments are recorded as marks on a 20 point scale. Students attempted one of a finite number of predetermined combinations of at least four of these externally examined subjects, according to their career aspirations. Most subjects were on offer in all schools, but there were five timetabled subjects offered by fewer than 20 schools: the least pervasive subject was offered in only eight schools. Similarly, subjects differed in popularity within schools with school entries averaging between 10 and 20 in five subjects and between 40 and 240 in the others. The least stable sample from the point of view of generalisability (Cronbach et al. 1972) is Social Studies with only 279 observations. The median number of observations in the analyses is 1,230 school-based assessments per subject.

4 Methodology

Considering the nature of the dataset and the aims of the study, Generalizability Theory was the most appropriate tool to use. Generalizability Theory has a very rich conceptual framework and is equipped with powerful statistical tests in order to address measurement issues like the ones handled in this study. Brennan (2001) illustrated how Generalizability Theory may be used to quantify and explain the consistency of observed scores and especially how useful it can be as a theory in the case of multi-faceted designs. In the simple case of a single-facet measurement design, persons (p) are usually crossed with items (i) and this is denoted by p×i. This study will start considering a single-facet design where persons (p) are crossed with occasions of assessment (o) which denotes the three teacher assessments, one for each term. Any person-occasion score that might be observed (an observable score) can be expressed using a linear model.

$X_{po}=$	μ	Grand mean
	$+\mu_p-\mu$	Person effect (ν_p)
	$+\mu_o-\mu$	Occasion effect (ν_o)
	$+ X_{po}- \mu_p-\mu-\mu_o+\mu$	Residual effect (ν_{po})

A one-factorial ANOVA model can be used to describe how the observed score of person (p) under condition (o) can be partitioned into a grand mean (μ), an effect for

each person (ν_p), an effect for each occasion (ν_o), and a residual (ν_{po}). More formally, the model can be written as

$$X_{po} = \mu + \nu_p + \nu_o + \nu_{po}$$

In the above equation, the grand mean (μ) is a constant value for all persons, positioning the score for all persons on the particular scale of measurement. The person effect (ν_p) describes the difference between a person's universe score and the grand mean indicating interindividual differences in universe scores. The occasion effect (ν_o) indicates the difference between an occasions 'harshness' (the teacher assessment for each term) and the average occasion 'harshness'. The residual combines the influence of the person-by-occasion interaction, systematic sources of error other than items, and random events.

According to this model, specific observations X_{po} vary due to differences between persons (i.e., person effects), differences between occasions (i.e., occasion effect), and other unaccounted for sources (i.e., residual). The variance of a set of observed scores ($\sigma^2_{X_{po}}$) may be partitioned into three components: persons (σ^2_p), occasions (σ^2_o), and the residual (σ^2_{po}). The residual combines two sources of variance that cannot be distinguished from each other in this model: they are the person-by-occasion interaction and chance fluctuations in scores.

All of the effects (except μ) are called random effects because they are associated with a process of random sampling from the population and universe. It is important to note that in generalizability theory, these random effects are defined in terms of mean scores for the population and universe. Therefore,

$$\underset{p}{E}\,\nu_p = \underset{o}{E}\,\nu_o = \underset{p}{E}\,\nu_{po} = \underset{o}{E}\,\nu_{po} = 0$$

All effects in the model are assumed to be uncorrelated. More specifically, denoting a prime designate a different person or occasion, this means that

$$E(\nu_p \nu_{p'}) = E(\nu_o \nu_{o'}) = E(\nu_{po} \nu_{p'o}) = E(\nu_{po} \nu_{po'}) = E(\nu_{po} \nu_{p'o'}) = 0$$

and

$$E(\nu_p \nu_o) = E(\nu_p \nu_{po}) = E(\nu_o \nu_{po}) = 0$$

At this point it should be noted that the above model has been specified without any normality assumptions, and without assuming that score effects are independent (which is a stronger assumption than uncorrelated score effects).

Using ANOVA procedures, the variance components and the % of variance attributed to each facet (or interactions of facets) may be computed. Progressively, additional facets like the school, the subject and format of the assessment (internal/external) will be added in order to meet the aims of the research. Figure 1 shows the model of the variance in a pupils (p) within schools (s) by subjects (c) by internal/external assessment (f) analysis of variance with p, s and c random and f fixed which is the most complex model used in this study (for more information see Brennan 2001). The estimations for the Generalizability Theory and the significance tests were carried out using SPSS version 11. In every case, the assumptions of the

Source	Mean Square*	Variance Components** (σ^2)
Subject	c	$\sigma^2_{cfps,e} + n_f\sigma^2_{cps} + n_f n_p \sigma^2_{cs} + n_f n_p n_s \sigma^2_c$
Format	f	$\sigma^2_{cfps,e} + n_c\sigma^2_{fps} + n_p\sigma^2_{cfs} + n_c n_p \sigma^2_{fs} + n_p n_s \sigma^2_{cf} + n_c n_p n_s \sigma^2_f$
School	s	$\sigma^2_{cfps,e} + n_f\sigma^2_{cps} + n_c n_f \sigma^2_{ps} + n_f n_p \sigma^2_{cs} + n_c n_f n_p \sigma^2_s$
Subject by Format	cf	$\sigma^2_{cfps,e} + n_p\sigma^2_{cfs} + n_p n_s \sigma^2_{cf}$
Subject by School	cs	$\sigma^2_{cfps,e} + n_f\sigma^2_{cps} + n_f n_p \sigma^2_{cs}$
Format by School	fs	$\sigma^2_{cfps,e} + n_c\sigma^2_{fps} + n_p\sigma^2_{cfs} + n_c n_p \sigma^2_{fs}$
Subject by Format by School	cfs	$\sigma^2_{cfps,e} + n_p\sigma^2_{cfs}$
Pupils within School	p(s)	$\sigma^2_{cfps,e} + n_f\sigma^2_{cps} + n_c n_f \sigma^2_{ps}$
Subjects by Pupils within School	cp(s)	$\sigma^2_{cfps,e} + n_f\sigma^2_{cps}$
Format by Pupils within School	fp(s)	$\sigma^2_{cfps,e} + n_c\sigma^2_{fps}$
Subject by Format by Pupils within Schools	cfp(s)	$\sigma^2_{cfps,e}$

Note:

*Mean Squares estimated using ANOVA procedures for each facet or interaction of facets

**The variance components provide a decomposition of the so-called "total" variance i.e. the variance of scores (Brennan, 2003).

Fig. 1 Model of the variance in a pupils (*p*) within schools (*s*) by subjects (*c*) by internal/external assessment (*f*) analysis of variance with p, s and c random and f fixed. *Mean Squares estimated using ANOVA procedures for each facet or interaction of facets. **The variance components provide a decomposition of the so-called "total" variance i.e. the variance of scores (Brennan 2003)

statistical tests were investigated and whenever in doubt, non-parametric tests were also run in order to compare the results. In all cases, only the results of the parametric tests are reported here because they were found to be reliable and valid and their interpretation is more intuitive. In Table 1, the non-parametric equivalent of the independent samples t-test was run in order to compare the average percentages of the variance estimates for the subjects with and without external examination. The Mann-Whitney U test agreed in all cases with the results of the t-test and gave statistically significant results whenever the t-test gave significant results (see Table 1).

Table 1 Averaged variance estimates (%) over 13 subjects where SBA is not followed by an external examination and 18 subjects where an external exam follows

Source	No external exam		External exam follows		t	p
	Mean[a]	S.D.[b]	Mean[a]	S.D.[b]		
o%	1.15	1.22	0.28	0.35	2.88	0.017[c]
s%	11.16	5.85	7.84	6.48	1.47	0.154
os%	2.57	2.92	2.18	1.61	0.48	0.390
p(s)%	69.18	13.74	78.86	9.61	2.31	0.028[c]
op(s),e%	15.94	8.02	10.84	4.19	2.31	0.029[c]
Total (raw)	5.11	3.81	7.6	1.73	2.45	0.020[c]

[a] Mean is the average of the % of raw score variance estimates attributed to each source (facets and interactions) over all 18 subjects

[b] S.D. is the standard deviation of the % of raw score variance estimates attributed to each source (facets and interactions) over all 18 subjects

[c] Mann-Whitney U test also gave statistically significant results at the 0.05 level

The t and p values on the table refer to the difference between the means (averaged variance estimates) of the 13 subjects where SBA is not followed by an external examination and the 18 subjects where an external exam follows. Values of p smaller then 0.05 indicate statistical significance

5 The first barrier: The conflict between the psychometric model and classroom assessment practice

Torrance and Pryor (2001) identify two social constructions of assessment. "Convergent" aims to discover if the learner knows, understands and can do a predetermined thing, an activity which would commend itself to Gipps et al's (1995) "systematic planners". This approach is characterised by an analysis of the interaction of the learner and the curriculum from the point of view of the curriculum - the curriculum comes first. An alternative approach espouses a more constructivist approach to learning and a "divergent" approach to assessment, aiming to discover what the learner knows, understands and can do. These assessors are focused on future development rather than current achievement, very much in the manner of Gipps et al's "intuitives" who developed a kind of gut feeling about the "whole child" based on their unaided recollections of critical incidents. Gipps et al's third category, "evidence gatherers", do not figure in Torrance & Pryor's survey analysis. It may be that presentation and discussion of task and quality criteria are now so taken for granted as fundamental classroom processes that the pressure to conform to government policy has swept the "evidence gatherers" into the convergent camp. Certainly the teachers involved tended to the linear sequence of activities — test, interpret, plan — implicit in the convergent approach "but how to actually use assessment data to help plan for the needs of all the children in their class remains unclear to them" (Torrance and Pryor, p. 621). This is a powerful corroboration of Paul Black's (2001) analysis. He asserts that it is merely a "dream" to suppose that requiring teachers to conduct SBA without adequate in-service support will in itself embed formative assessment in everyday classroom practice.

Our data set suggests that what does get embedded is a reporting strategy. In each subject, three SBA scores are available for each pupil in each school. In generalisability terms, there are three sources of variance; the occasion of assessment, o, which is random (i.e. irrelevant) in classical theory, the school, s, which includes the impact of the teacher as rater, and the pupil within the school, p(s), which includes both the variance due to the pupil and the pupil's interaction with the school. (He/she might have fared differently elsewhere). These three sources of variance interact and the complete design, random throughout, can be seen in the row entries of Table 1. In this and the following tables to ease comparison the row entries are reported as percentages of the total variance in each subject analysis. The subject analyses of course are computed from raw data (ratings on a 20 point scale). The columns refer to two conditions of assessment, one in which the teacher is final arbiter of grades and one in which school-based assessment (SBA) is followed by an external examination. In both cases, the data analysed are the SBA results alone showing the influence on teachers of the existence of the external test. These conditions coexist in each school and even in some subjects, i.e. languages may be offered both as examined subjects of study and as non- examined service subjects, rounding out the curriculum.

Each condition encompasses a wide variety of subjects. The examined group includes subjects drawn from the humanities, languages, sciences and practical activities such as typing and technology. The group of subjects without external examination is equally various (in fact it consists of exactly the same subjects) with the addition of the performing arts in the shape of art and music. In fact, the two conditions consist of the same type of subjects, taught by the same teachers, using the same textbooks in the same buildings. For example, science is offered both as a subject with and without examination — all the other aspects remain the exactly the same. The external examination at the end of the year is the only obvious bias in the comparison of the two groups. It would seem, therefore, that the two assessment regimes are in themselves able to so condition assessment activity in different subject classrooms that significant differences emerge between the regimes. It should be noted at this point that the significant differences in Table 1 exist even without the inclusion of the results of arts and music — therefore the differences between the two regimes may not be attributed to the different nature of the two groups of subjects — the examination again seems to be the only factor that makes the difference. Sufficient homogeneity has been created in the subject classrooms of each group to elicit significant contrasts between the two assessment regimes in the same schools. The first condition for SBA as a vehicle of educational change has been met. It standardises not only the recording of achievement but also the main influences on recorded achievement in a diverse range of subjects.

Furthermore, although all subjects were marked out of 20, a significant difference emerged in the total mark variance. Where there is no external examination to constrain the teachers' judgments the variance in scores is significantly smaller. Inspection of the raw data reveals that in every subject regardless of condition full marks were awarded, sometimes liberally. In just over half of the subjects without external examination however no mark awarded was below ten, a truncation of range that occurred in only one subject where there was an external examination. The significant reduction in total variance and the reduced percentage of this reduced

variance attributable to the individual pupil, p(s), are an extreme example of a "pressure to reward" often observed in educational systems.

Existing literature around the world has warned about the above phenomenon where teachers award grades using assessment criteria that do not comply perfectly with the official assessment guidelines. For example, Davison (2004) reported that teachers in Australia did not base their assessment of student ability solely 'ticking the boxes according to published assessment guidelines' but also used other parameters emerging from their professional judgment. On the other hand, in Honk Kong the authors found much more variability in the underlying assessment criteria of the teachers. The teachers reached consensus on the assessment criteria through reference to community norms rather than explicit statements of performance. Moerover, in American schools McMillan et al. (2002) report that American teachers conceptualise two major components of classroom performance when grading: actual academic performance and an amalgam of effort, ability, improvement, work habits, attention and participation. McMillan et al. categorise as those teachers as "enablers to academic performance" (p. 209) and Filer and Pollard (2000) might recognise those teachers *as* a platform for assessment within a more constructivist paradigm. Also, Brookhart (1993) noted that in the American grading system below average students get a bonus if they have tried hard, while above average students get the grade earned with no additional penalty for working below their expected level.

Table 1 however also suggests a countervailing pressure. There is significantly less occasions variance, o%, and pupil by occasions variance, op(s)%, where an external examination follows the SBA. Teachers faced with an external examination appear to be importing psychometric assumptions into their classrooms rather than a model of progress in learning. The teachers would no doubt express it as reporting on each occasion on the basis of an expectation of final grade predicted from the current performance, rather than feeding back on current performance against an invariant standard.

If SBA is to be used in the service of school improvement, the fundamental issue, one which Boards are aware of but mostly prefer not to confront, is the denial of change in repeated measures upon which the whole edifice of classical reliability theory is built. The true score of a person is the average of the observed scores of an infinite number of administrations of the same test to the person, — without impacting on the person's performance. Impact is conceptualised not as learning but as "practice effect" and that is a form of error. Thus perfect reliability is achieved when essentially the same question is asked over and over again with precisely the same result each time. That is far from the outcome that teachers expect. They only repeat a question verbatim in the hope that intervening explanation will have lead to a different response. Teachers are in the business of creating classical unreliability.

Learning trajectories assume that the learner is in a state of flux, a direct conflict with the classical assumptions of psychometrics in which repeated measures reveal a central tendency which is "true". In Table 1 the influence of external examining seems to tend in the psychometric direction. To this group of subjects the variance in pupils scores from occasion to occasion (op(s),e) is low. The main effect of occasions is also negligible, as is the interaction of occasion and school. In this assessment regime, the same standard is observed at the beginning as at the end of

the school year and that consistency is observed in school after school. It looks very like the tendency noted by Harry Black (1993) in Scotland a few years after the introduction of the National Curriculum. Teachers were not recording pupil achievement in terms of the particulars of the curriculum. Instead they were following the practice of the public examination boards and were using SBA to record performance in slices of the curriculum tested in summative fashion.

Teasdale and Leung (2000) explore the conflict between assessment for learning and assessment as classification in some depth from the teachers' point of view, concluding that psychometric approaches may not provide an adequate response to recent pedagogical developments. Examinations board practice in communicating standards through actual samples of students' work for each of the levels of the system, annotated to illustrate important points, has been found to be successful (Black and Wiliam 1998) but in the constructivist paradigm a learning event points to a future state of being rather than a current state of knowing with feedback directing the students' attention to particular performance dimensions where change is desired/expected. There is assistance aplenty in capturing the current performance but it is left to the teachers to negotiate a compromise between the current achievement and the likely credentialled outcome.

Stability is antithetic to the learning objective of the assessment as Rea-Dickins and Gardner (2000) have demonstrated through interviews with second language teachers. They note a strong informal dimension to SBA. The knowledge that a teacher has about individual pupils, and about English as Additional Language (EAL) development in general, plays a significant part in the teacher's decision making about the language development, attainment and ability of individual pupils. "It follows that valid inferences about the learner's future are only to be made from evidence which is interpreted in the light of a theory of the social nature of learning" (Black 2001, p.80).

Here we are concerned with the Board's point of view. One powerful argument for the introduction of SBA is that examination reliability will be increased thereby. But if a constructivist view of knowledge acquisition is adopted (Filer and Pollard 2000), it is only the most recent observation that is "true" and even that is only provisional. Should Boards then ask for a single summative judgment from their teachers referred to the student's grasp of the subject at the end of the course of instruction or should they average multiple teacher reports as classical theory demands? Table 1 suggests that, left to their own devices, teachers will orient their assessments more towards what it is best to feed back to the child at any point in time but once examination boards raise the assessment stakes, reporting has the board as its audience, rather than the student. Consistency of standard then becomes the badge of the teacher's professionalism, at the expense of formative feedback.

6 The second barrier: Obduracy of the School effect

One final feature of Table 1 is not immediately apparent. Cronbach et al. (1972) introduced generalisability theory as an aid to test construction. By first identifying and measuring distinct sources of error it then allows an optimum test to be devised by predicting the size of these errors for longer or shorter tests using the Spearman-Brown

formula. The variance estimates in Table 1 refer to the component influences on a single score. If we wish to know the reliability of the sum of the three SBA scores we can multiply the "true" score, p(s), by three because it is perfectly correlated with itself and leave the other variance components as they are, since error is uncorrelated and so does not aggregate. Using only p(s) and op(s),e from the SBA only column of Table 1, the average reliability of the sum of three teacher assessments is

$$\begin{aligned}\text{Reliability} &= \frac{\text{true}}{\text{true} + \text{error}} \\ &= \frac{69.18}{69.18 + 15.94} = 0.81 \text{ for a single observation} \\ &= \frac{3 \times 69.18}{3 \times 69.18 + 15.94} = 0.93 \text{ for the sum of 3 independent observations}\end{aligned}$$

The same outcome will be obtained by dividing the error rather than multiplying the true which is how Cronbach et al calculate their delta coefficient of reliability. While theirs is a fully worked out theory it is also of its time, a time when schools made no difference according to the Coleman (1966) report. Things which make no difference have no place in a generalisability study, so although Cronbach et al deal with items nested within tests they do not deal with persons nested within institutions. However if we extrapolate from the analysis so far, every time we observe the pupil we are also observing the person's interaction with the school and so must include the school in the analysis. The school effect will then behave as the pupil effect. It is not random in the sense that error is. It is a stable component of the person's achievement and so each time the pupil variance is multiplied so is the School variance. It does not go away.

Attempting to establish whether school effects are true or error is the most salient problem in the use of SBA for public examination purposes in the developed world. It will be even more so in developing countries where studies of school effectiveness reveal a school variance estimate of 30–40% in contrast to the 10–15% found in industrialised countries (Scheerens 2001). Table 2 explores the status of the school

Table 2 Averaged variance estimates over 18 subjects where SBA is followed by an external examination before and after taking out the regression of the final examination

Source	Raw score variance estimates		Residual variance estimates		% variance explained by external exam
	Mean[a]	S.D.[b]	Mean[a]	S.D.[b]	
o%	0.019	0.029	0.033	0.027	0
s%	0.518	0.304	0.223	0.157	57
os%	0.146	0.100	0.001	0.003	99
p(s)%	6.130	1.707	3.111	1.399	49
op(s)	0.787	0.235	0.011	0.014	99
Total (raw)	7.601	1.702	3.379	1.449	55

[a] Mean is the average of the % of raw score variance estimates attributed to each source (facets and interactions) over all 18 subjects

[b] S.D. is the standard deviation of the % of raw score variance estimates attributed to each source (facets and interactions) over all 18 subjects

effect as true or error by using the external examination as an independent estimate of achievement and regressing it on each occasion of SBA in turn, a form of moderation. That takes out the occasions variance other than the main effect, o, which is created by the discrepancy between the overall mean and the mean of school means when the schools are of different sizes.

The occasions variance, however, is not the problem. It can be dealt with by increasing the number of independent observations. It is the school effect that has to be minimised and here the moderation technique is much less successful. The correlation between internal and external school means in these schools is stronger than the correlation between internal and external pupil scores, but in each case only half the variance is accounted for. There remains a school component of candidate scores which measures something other than the results of the external tests and which cannot be minimised by these means. Differences in school effectiveness which cannot be accounted for are a threat rather than an opportunity in public examining.

There is also a component of the pupil variance which cannot be accounted for by the external examination. Can we therefore take the pupil variance as true or is half of that error too?

7 The third barrier: Biases in teacher judgement

Teachers mediate the external pressures upon them through the 'filter' of their own professionalism″ (Yung 2002, p. 99). Professional decision makers, teachers among them, are engaged in a four stage process; (a) identify relevant cues from the array of information, (b) assess the amount or intensity of the cue, (c) cluster the cues into a smaller number of dimensions and finally (d) weight the dimensions to reach an overall decision (Einhorn 2000).

Two of the above steps, (a) and (d), are somewhat problematic in the context of teacher assessment; for example, Raveaud (2004) suggested that routine assessment in the classroom by teachers "constitutes a prism through which one can examine teachers' beliefs and values", implying that teachers in different contexts may identify different cues as "relevant" to the assessment criteria and may apply different weights on different assessment dimensions. Along similar lines, Hall and Harding (2002) suggest that the interpretation and application of assessment criteria by teachers is a difficult task — usually there is too much information about pupil performance to select from and it is difficult to identify relevant cues from this array of information that much the assessment criteria, especially if they are very general.

Moreover Feiler and Webster et al. (1999) found that teachers making literacy predictions for an incoming grade 1 cohort were prepared to do so on the most partial of information gleaned even before the children themselves were encountered in the classroom. These judgments were strongly influenced by perceived social class stereotypes, such as home address or parental occupation. Nevertheless initial representations of children as "likely to fail" became more permanent with time. This is the "anchoring-and-adjustment" heuristic, a tendency for judgements to be biased towards an initial value arrived at from only partial or even no detailed consideration (Baron 1994). The mechanism that increases the stability of first

impressions has been identified by Kahneman and Tversky (1982) and Arkes (1986) among others. They have shown that information which is confirmatory of an initial hypothesis tends to be accorded more weight than evidence which is contradictory. Such selective attention to behaviour which reinforces intuitive first impressions will tend to increase the internal consistency of teacher assigned scores. SBA may be stable rather than reliable.

Predictive validity also tends to be enhanced, albeit inadvertently. Feiler and Webster (op cit) found shades of the self-fulfilling prophecy: teachers tended to provide less support for children expected to perform poorly, further confirming their initial expectation. The antidote to these biases is the requirement for evidence and comparability of treatment fundamental to public examination processes. Care in specifying the domain to be assessed as well as clarity of and consistency in assessment criteria are important components in large scale teacher assessment. It is essential to sound decision making that the process is articulated in this way. While ordinary people as well as experts are good at picking out cues which predict a criterion, ordinary people are hopeless at integrating the dimensions of the decision into a single variable.

In support of this conclusion Elander and Hardman (2002) cite a very large body of research where the statistical combination of separate items of information, usually by multiple regression, has been found to be superior to a single overall judgment. Their own research started from previously developed dimensions for the evaluation of university essays; addresses the Question, covers the area, understands the material, evaluates the material, presents and develops arguments, structures the answer and orders material and clarity in presentation and expression. Each dimension was backed up by criteria related to seven levels of degree achievement. Over 500 psychology essays were scored on each of the dimensions and given an overall grade by a first marker who had set the essay and a second marker less familiar with the field. Four of the seven markers both first marked their own essays and second marked essays they had not set. These four marked the less familiar material more highly and used a smaller mark range. This is a common reaction. Markers who are not sure of their subject matter tend to generous but undifferentiated decisions. For all second markers, the simple sum of the dimension scores added significantly to the accuracy of the correlation with the first marker. “The results appear to indicate that, for second markers, separate ratings of specific aspects of answers that were not incorporated in the overall mark awarded could help to explain discrepancies in marks between markers" (p. 320). Both raters were aware of the same attributes but differed in the weight they placed upon them.

These conflicts must be played out time and again between the external examination and the teacher-qua-assessor. Traditional notions of assessment tend to focus rigorously on control over the conditions of assessment on grounds of equity. Classroom-based performance assessment lays its claim to validity in the rich diversity of behaviour which is its subject matter, without however any good grounds for belief that the behaviour is either stable over time or over task. Which of these approaches holds out the best hope of improving teaching? Holmes (1998) cites the philosophical rejection by many influential educators of the clear finding that effective skill teaching depends on direct instruction with a clearly sequenced curriculum. The adoption of an integrated curriculum, common in the primary phase,

generally reflects wider beliefs about learning and educational processes which tend to run counter to psychometric theory with its focus on construct and domain specification. The distributed decision processes of SBA with teachers responsible for the first two stages of Einhorn's (op cit.) decision process and the boards responsible for the last two, can therefore be seen as steering the assessment system away from the more holistic emphasis which Filer (2000) sees as an implicit (and international) response to the problem of educating pupils for the knowledge society.

Table 3 is an attempt to throw light on that holistic emphasis. It brings together eight separate assessments of each pupil, four from external examinations and four from intenal assessments of a coherent group of four subjects. It attempts to determine the major sources of turbulence in reaching an overall view, or, in generalizability terms, the extent to which the pupil component in SBA can be held to be true. The analysis follows the model of Fig. 1 which subsumes the simple occasions, schools, persons model employed so far. The new components are the subject of study and the form of assessment, external or internal. The latter contrast is between the external score and the average SBA over three occasions.

The subject combinations are Latin, history and philosophy in the humanities group with 26 schools and 1,068 pupils, mathematics, physics and chemistry in Sciences with 27 schools and 1,228 pupils, and accounting, economics and mathematics for economists in the business group with 28 schools and 2,180 pupils. The fourth subject in each group is the national language.

The first point to be made about Table 3 is that the main effect of format arises from the systematic tendency across all subjects and schools for teachers to give their students the benefit of the doubt, both in the interests of classroom hygiene but also in response to external pressures. This same effective outcome is discussed by

Table 3 Subjects, formats and pupils: absence of a school effect in three subject combinations

Source	Economics	Sciences	Humanities
		variance components[a] (σ^2)	
c[subjects]	0.356	1.364	0.501
f[format]	6.967	4.517	8.309
s[schools]	0.115	0.025	0.012
cf	0.203	1.127	1.849
cs	0.111	0.215	0.227
fs	0.380	0.478	0.152
cfs	0.371	0.556	0.555
p(s)[pupils]	8.223	7.755	8.880
cp(s)	1.329	2.273	0.776
fp(s)	1.162	0.994	1.336
cfp(s)	3.450	3.308	4.215
Total	-	-	-

[a] The variance components provide a decomposition of the so-called "total" variance i.e. the variance of scores (Brennan 2003)

Woods and Levacic (2002) as "positional performance focus", special attention to borderline students who have the potential to improve the school's relative position on key performance indicators.

The second point qualifies the first. In Table 2 we noted the obduracy of the school effect when repeated measures came from the same source. Here there are eight distinct accounts of the effect and they appear to be essentially random in that they cancel each other out. The school effect is nullified when looked at from a range of perspectives, a finding by no means entirely at variance with the literature where the characteristics of effective schools are different for different outcomes, i.e. different schools count as effective according to the outcome chosen for measurement. The Table 3 outcome is not only close to a jaundiced reading of the literature (c.f. Thrupp 2001), it provides positive support for Harris's (2001) view that it is the subject department in secondary schools that should be used as leverage in school improvement efforts. These departments are represented by the cs and cfs effects in Table 3, which have picked up the missing schools variance.

Finally, we do indeed now have a more rounded picture of the individual pupil, but only at the expense of adding in a deal of extraneous variance in cp(s) and fp(s). The relative magnitude of these two is in itself surprising. That pupil performance varies from school subject to school subject is well attested and the entire baccalaureat (group certificate) approach to selection for higher education is designed to minimise these variations. That the format effect arising from the difference between school-based and external assessment is equally powerful has not to the writers' knowledge been remarked upon in the literature. If this result were to be replicated in another school system then, on grounds of equity alone, SBA should become universal practice. Why should pupils who please teachers be accorded less worth than pupils who please tests?

It is interesting, however, the fact that there are marked differences between the results of Sciences and the other subject groups as shown on Table 3. Although much could be said about these differences, one of the most striking results is that the format of the assessment accounts for less than 20% of the total variance for the Sciences; the corresponding percentage in the other subject groups is above 30%. Moreover, the % of variance explained by the subject in Sciences is four times the corresponding percentage on the other subject groups. It is obvious that the nature and the structure of the science subject reflects to these results — it seems that the external assessment and SBA come closer.

8 Conclusion

There is a strong positive relationship between frequency of classroom tests and pupil achievement both in industrialised (Bangert-Drowns et al, op cit) and developing nations (Scheerens, op cit) but not for all pupils. Some pupils please teachers rather than tests and tend to do so across all subjects. If this result is found to be general, it follows that an element of SBA in every subject could contribute positively on the fairness of the assessment system.

Unfortunately such a move cannot be guaranteed to increase the effectiveness of schools. Hill (1998) advocates system-wide monitoring and accountability studies as

the next step in integrating school effectiveness and the accountability agenda with inquiry into the mechanisms of school improvement. But there is a prima facie case that "standards can only be raised by improved teaching. Testing can only help if ways can be found to resolve the tension between the demands of accountability testing and the requirements for tests to be valid in reflecting and reinforcing good pedagogy" (Black 2001, p.73).

Ross et al. (2002, p.82) go even further in assuming that evaluation data do not influence achievement directly but rather that the relationship is mediated by student cognitions. There is plenty of supporting evidence. Torrance and Pryor (2001) p. 616) observed *"great differences between children in the same [primary] class, dependent on their perceptions of the implicit social rules of the classroom and their orientation to achievement goals" (p. 616).* Ames (1992) found *"When evaluation is normative, emphasises social comparison, is highly differentiated, and is perceived as threatening to one's sense of self- control, it contributes to a negative social climate" (p.265).* And even when absolute standards are the scale against which achievement is reported, students in high performing classes underestimate their own performance through social comparison (Marsh et al. 1995) while those at risk of early leaving find that all departures from expectation are their responsibility and theirs alone (Smyth and Hattam 2002) — a peculiarly pernicious version of *"blame the child".*

Nevertheless the tension between assessment for learning and assessment in the interests of the wider society can be mitigated by the external examination boards. Public examination boards have to recognise these polarities and be explicit in where they stand. At the moment they are well placed to provide evidence of school effectiveness but at the price of classrooms "*dominated by staccato forms of the old end-of- session examinations. Continuous assessment in action [can mean] continual examination for reporting...*"(Black, H. D. quoted by Black, P. 2001, p.74). The boards can however take explicit steps to redress the balance. One such would be to give much heavier weight to assignments set late in the course as an encouragement to teachers to feedback performance as it is, not as it might be by the time a result is issued. Another and more radical step would be the introduction of SBA in every subject backed up by moderation against the mean of all the student's SBA scores rather than against the external examination score in the subject at issue. The peculiar contribution of classroom behaviour and learning, something like civility, would then be recognized not as a source of error but as an equally valid but different perspective on the student's achievements. Such a strategy would have the disadvantage of rendering the overall mean of SBA scores for the school — potentially its effectiveness — even more problematic. The advantage would be that this source of variance would now be relatively uncontaminated by competing and more complex sources and might then begin to be better understood — a research agenda for a post-internal-consistency phase of public examining.

References

Ames, C. (1992). Classrooms, goals, structures and student motivation. *Journal of Educational Psychology, 84*(3), 261–271. doi:10.1037/0022-0663.84.3.261.

Arkes, H. R. (1986). Impediments to accurate clinical judgement and possible ways to minimise their impact. In H. R. Arkes & K. R. Hammond (Eds.), *Judgement and decision making: An inter disciplinary reader* (pp. 582–592). Cambridge: Cambridge U.P.

Bangert-Drowns, R., Kulik, J., & Kulik, C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research, 85*(2), 89–99.

Baron, J. (1994). *Thinking and decision making* (2nd ed.). New York: Cambridge U.P.

Black, H. (1993). Taking a closer look. Key ideas in diagnostic assessment. Edinburgh: The Scottish Council for Research in Education. Reached at http://www.scre.ac.uk/pdf/taking/key.pdf on 11 May 2003.

Black, P. (2001). Dreams, strategies and systems: Portraits of assessment past, present and future. *Assessment in Education, 8*(1), 65–85. doi:10.1080/09695940120033261.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74. doi:10.1080/0969595980050102.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (2003). Coefficients and Indices in Generalizability Theory. Center for Advanced Studies in Measurement and Assessment, CASMA Research Report, Number 1.

Broadfoot, P. (1996). Liberating the learner through assessment. In G. L. Claxton, T. Atkinson, M. Osborn & M. Wallace (Eds.), *Liberating the Learner: Lessons for professional development in education* (pp. 32–44). London: Routledge.

Brookhart, S. M. (1993). Teachers'grading practices: Meaning and values. *Journal of Educational Measurement, 30*, 123–142. doi:10.1111/j.1745-3984.1993.tb01070.x.

Coleman, J. (1966). *Equality of educational opportunity*. Washington: United States Government Printing Office.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *Theory of generalizability for scores and profiles. The dependability of behavioral measurements*. New York: Wiley.

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing, 21*(3), 305–334. doi:10.1191/0265532204lt286oa.

Dweck, C. S. (1999). *Self theories: Their role in motivation, personality and development. Philadelphia*. Philadelphia: Psychology.

Einhorn, H. J. (2000). Expert judgement: some necessary conditions and an example. In T. Connolly, H. R. Arkes & K. R. Hammond (Eds.), *Judgement and decision making: An interdisciplinary reader* (2nd ed., pp. 324–335). Cambridge: Cambridge U. P.

Elander, J., & Hardman, D. (2002). An application of judgement analysis to examination marking in psychology. *British Journal of Psychology, 93*, 303–328.

Feiler, A., & Webster, A. (1999). Teacher predictions of young children's literacy success or failure. *Assessment in Education, 6*(3), 341–356. doi:10.1080/09695949992784.

Filer, A., & Pollard, A. (2000). *The social world of pupil assessment: Processes and contexts of primary schooling*. London: Continuum.

Filer, A. (ed). (2000). *Assessment: Social practice and social product*. London: Falmer.

Gipps, C. V., Brown, M., Mccallum, B., & Mcalister, S. (1995). *Intuition or evidence?*. Buckingham: Open University.

Hall, K., & Harding, A. (2002). Level descriptions and teacher assessment in England: Towards a community of assessment practice. *Educational Research, 44*(1), 1–15. doi:10.1080/00131880110081071.

Harris, A. (2001). Department improvement and school improvement: A missing **link?**. *British Educational Research Journal, 27*(4), 477–486. doi:10.1080/01411920120071470.

Hill, P. W. (1998). Shaking the foundations: Research driven school reform. *School Effectiveness and School Improvement, 9*, 419–436. doi:10.1080/0924345980090404.

Holmes, M. (1998). Change and tradition in education: The loss of community. In A. Hargreaves, A. Lieberman, M. Fullan and D. Hopkins (Eds.) *International Handbook of Educational Change* (p. 242–246) Dordrecht; Kluwer A.P.

Kahneman, D., & Tversky, A. (1982). On the psychology of prediction. In D. Kahneman, P. Siovic & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 48–61). Cambridge: Cambridge U.P.

Kellaghan, T., & Greaney, V.(1992). *Using Examinations to Improve Education: a Study of Fourteen African Countries*. World Bank Technical Paper Number 165. The World Bank; Washington, D. C.

Kluger, A. N., & Denisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284. doi:10.1037/0033-2909.119.2.254.

Marsh, H., Chessor, D., Craven, R., & Roche, L. (1995). The effects of gifted and talented programmes on academic self-concept: The big fish strikes again. *American Educational Research Journal, 32*(2), 285–319.

Mcmillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*(4), 203–213.

QCA.(2002). *Assessment and reporting arrangements*: key stage I, 2002. London.

Queensland Board of Senior Secondary School Studies.(2003). *Moderation of achievements in school-based assessments*. Reached at http://www.qsa.qld.edu.au/publications/senior/files/ModerationOfAchievements.pdf reached at 10 May 2003.

Raveaud, M. (2004). Assessment in French and English infant schools: Assessing the work, the child or the culture? Assessment in Education: Principles. *Policy & Practice, 11*(2), 193–211.

Rea-Dickins, P., & Gardner, S. (2000). Snares or silver bullets: Disentangling the construct of formative assessment. *Language Testing, 17*(2), 215–244.

Reeves, D. J., Boyle, W. F., & Christie, T. (2001). The relationship between Teacher Assessments and Pupil Attainments in Standard Test Tasks at Key Stage 2, 1996–1998. *British Educational Research Journal, 27*(2), 141–160. doi:10.1080/01411920120037108.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (2002). Influences on student cognitions about evaluation. *Assessment in Education, 9*(1), 81–95. doi:10.1080/09695940220119201.

Scheerens, J. (2001). Monitoring school effectiveness in developing countries. *School Effectiveness and School Improvement, 12*(4), 359–384. doi:10.1076/sesi.12.4.359.3447.

Smyth, J., & Hattam, R. (2002). Early school leaving and the cultural geography of High Schools. *British Educational Research Journal, 28*(3), 375–398. doi:10.1080/01411920220137458.

Teasdale, A., & Leung, C. (2000). Teacher assessment and psychometric theory: A case of paradigm crossing? *Language Testing, 17*(2), 163–184.

Thrupp, M. (2001). Recent school effectiveness counter-critiques: Problems and possibilities. *British Educational Research Journal, 27*(4), 443–459. doi:10.1080/01411920120071452.

Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal, 27*(5), 615–632. doi:10.1080/01411920120095780.

Woods, P. A., & Levacic, R. (2002). Raising school performance in the league tables (Part 2): barriers to responsiveness in three disadvantaged schools. *British Educational Research Journal, 28*(2), 227–247.

Yung, B. H. W. (2002). Same assessment, different practice: professional consciousness as a determinant of teacher's practice in a school-based assessment scheme. *Assessment in Education, 9*(1), 97–117. doi:10.1080/09695940220119210.