Subjectivity, Objectivity, and Teachers' Qualitative Judgments

Discussion Paper 5

Abstract: Qualitative judgments play an essential role in the assessment of student achievements in all subjects. This Discussion Paper contains a definition of qualitative judgments, a discussion of the meanings of subjectivity and objectivity, and a statement of certain conditions that must be satisfied if qualitative judgments are to enjoy a high level of credibility.

Author:
Royce Sadler, Assessment Unit. January 1986
Subjectivity, Objectivity, and Teachers' Qualitative Judgments

Teachers' professional judgments, like the judgments of many other professionals, are about the behaviours, conditions, characteristics, or performances, of persons. For teachers, the persons are of course their students. In assessing student work, not only do teachers engage in a deliberate act, but they are also sensitive to the consequences both for their students and for themselves of making poor judgments.

The ROSBA scheme relies heavily on the professional judgments of teachers, as did the Radford scheme it replaced. This Discussion Paper is about those professional judgments. In particular, it contains

(a) a discussion of the characteristics of a certain class of professional judgments which, for the sake of convenience, are called here qualitative,
(b) an analysis of qualitative judgments in terms of subjectivity and objectivity, and
(c) a consideration of several conditions that make for sound qualitative judgments.

Structurally, a number of general issues are dealt with first. Towards the end of the Paper, it is shown how recent developments in school-based assessments in Queensland, using criteria and standards, fit into the broader conceptualisation. Some readers might find it advantageous to skim-read this last section first, using it as an advance organiser. Throughout the Paper, criteria and standards are distinguished in an important sense. Criteria refer to dimensions or characteristics by which appraisals are made, while standards are fixed points of reference along the criteria. Criteria can exist without standards, but not the reverse.

A measure of confidence in a judgment can be obtained by comparing one judgment with a number of other judgments, all of which are made by suitably experienced members of a profession. That is, a particular judgment can be validated according to how well it agrees with the judgments of other members of the 'guild' of professionals. In practice, it is not normally expected that consensus must be reached on every single judgment. It is often sufficient to take samples of judgments for checking, or to set up procedures by which a particular judgment could, if necessary or on appeal, be compared with those of a group of experts.

External examinations were abolished in Queensland in the early 1970s. Since that time, the professional judgments teachers make about the achievements of their students have formed the basis for summative grading and certification in secondary schools. Teachers have always, of course, been making professional judgments, but until the introduction of the Radford scheme, these were mostly limited to the organisational aspects of teaching, and to providing evaluative feedback to students and parents about the educational progress of students.
Qualitative Judgments Defined

Of special interest in school-based assessment is the particular subset of professional judgments that might be called qualitative. A qualitative judgment is defined for the purposes of this Paper as one made directly by a teacher, in which the teacher's brain is both the source and the instrument for the appraisal. The judgments are not reducible to a formula that can be applied by a non-expert. Here "instrument" is used in a different sense from what is meant by "test instrument", in which case it means a stimulus or task specification to which students respond. An analysis of judgments that are not qualitative in the sense just defined is important philosophically, but a discussion of this is beyond the scope of this Paper. Suffice it to say that non-qualitative judgments are not necessarily quantitative in that numbers or formal measures are involved.

Although technically irrelevant to the above definition, the notion of a qualitative judgment in the context of school-based assessment carries some felicitous overtones, because teachers are engaged in assessing the quality of student performance—deciding whether what a student does is, according to accepted criteria and standards, excellent, good, or bad. But judgments about quality can, under certain conditions, also be made quantitatively and mechanically.

In general, qualitative educational judgments are characterised by some or all of the following:

(a) At least some of the criteria are 'fuzzy' rather than 'sharp'. A sharp criterion contains an essential discontinuity in which there is an abrupt transition from one state to another, such as from correct to incorrect. It is always possible in principle to determine which state applies. By contrast, a fuzzy criterion is one for which there is a gradation from one state to another. Originality, as applied to an essay, is an example of a fuzzy criterion because everything between wholly unoriginal and wholly original is possible. Qualitative judgments usually involve some fuzzy criteria.

(b) There is often no independent method of confirming, at the time when a judgment is made, whether the decision or conclusion is correct. Indeed, in some circumstances, it is meaningless to speak of "correctness" at all. To give an example of methodological independence, suppose that two lamps with potential use as street lights are being compared. One way to determine effectiveness is to ask someone to judge which provides the better illumination, either in the laboratory or in the field. A different method of judging illumination would be to set up appropriate electro-optical instruments. These two methods are independent because they use essentially different means for arriving at a conclusion. Having several persons instead of just one would not qualify as an independent method, although several persons may make judgments without reference to one another, and in that sense work independently.

(c) The performances of students differ from one another on a number of dimensions. Typically there are multiple criteria, which are used simultaneously rather than sequentially. As well as individual elements or features, the total pattern of relationships among those elements is important. Because the criteria interlock, the overall configuration is what matters. In making qualitative judgments of this type, the teacher draws inferences from multiple cues. Similarly because the raw evidence is not unidimensional, the notion of equivalences between different performances is a key concept.

(d) There exists a large pool of potential criteria that could legitimately be brought to bear in particular assessments, but only a relatively small subset are used at any one time. The competent judge is not only able to make an appraisal, but can also substantiate a judgment by appeal to relevant criteria. Professional judgment consists in knowing the rules for using (or occasionally breaking) the rules.

(e) If numbers are used, they are assigned after the qualitative judgment has been made, not the reverse. Consider the situation where an overall conclusion is arrived at by adding together numerical marks (from, say, different test items, or components in a subject) and then looking at only the sheer magnitude of the result. The compounding process in this case is mechanical and therefore, according to the definition above, not qualitative. This is so in spite of the fact that the contributing judgments that give rise to the marks may themselves be qualitative.

Subjectivity and Objectivity

No consideration of the nature of qualitative judgment proceeds far before the matters of subjectivity and objectivity are raised. The definition of qualitative judgment given earlier in this Paper implies that ALL qualitative judgments are, in a restricted sense, subjective. This is simply because the appraising apparatus is the brain of a person (that is, a subject, not an object).

This is not the way the term subjective is popularly used. It is often used in a derogatory sense when referring to judgments that are skewed or unbalanced. That is why teachers sometimes apologise for their own subjective judgments, or tend to dismiss those of others as "mere opinion". But of course any judgment, subjective or otherwise, that is in accord with known facts or accepted contentions is, by definition, a good one. The bad reputation of subjective judgments arises from a failure to distinguish between judgments that are based on sound evidence and reasoning, and those that are based on personal taste or whimsy. They all tend to get tarred with the one brush. Subjective judgments can be arbitrary, biased, or slanted. On the other hand, they can be quite dependable. What must be done is to identify and create those conditions that lead to appraisals that are potentially reproducible and independently confirmable.

The transition from subjective, meaning person-al, to subjective, meaning biased, probably stems from a thought sequence somewhat as follows:

(a) a judgment is subjective, by definition, if a person makes it;
(b) the human brain processes facts in a private, value-laden environment, the resultant judgments often being idiosyncratic; therefore
Now it is a matter of both research and common knowledge that the brain, as
information processor, is not perfect. Indeed, many systematic tendencies
towards distortion have been identified in the literature on human judgmental
processes. But is it a categorical mistake to conclude that because some subjective
judgments turn out to be biased and not in accord with the facts that all such
judgments are therefore biased, or at least suspect. If one were to apply that line
of thinking to daily life, things would quickly grind to a halt. Life is simply
saturated with subjective judgments, most of them being made without much
conscious thought. Even the most savage critic of subjective judgments could not
afford to reject them out of hand. The challenge facing educators is to devise
systems in which subjective judgments are made within a context of checks and
balances so that the probability of unbalanced appraisals is minimised.

In relation to student assessment, proposition (b) calls for some elaboration. The
literature on subjective assessment contains many depressing accounts of the
fallibility of teachers' judgments. A number of effects are well established. These
include unreliability (both inter-rater discrepancies, and the inconsistencies of
one rater over time), order effects (the carry over of positive or negative
impressions from one appraisal to the next, or from one item to the next on a test
paper), the halo effect (letting one's personal impression of a student interfere
with the appraisal of that student's achievement), a general tendency towards
leniency or severity on the part of an assessor, and the influence of extraneous
factors (such as neatness or handwriting). The reasons for these effects are
multiple: the intrinsic difficulty of maintaining cognitive constancy over time
 especially when a large number of essentially similar judgments have to be
made), fatigue and boredom, carelessness, capriciousness, cursoriness, and
personality clashes.

All of these appear to place subjective judgments under something of a cloud.
However, much of the research cited by critics of subjective assessment has been
undertaken with teachers working intuitively, with minimal external control by
way of rules or guidelines over the processes involved. On the other side, there is
unequivocal evidence that teachers who assess within a clear assessment
framework, and who have been trained in the use of the appropriate tools, make
consistently reliable and valid appraisals. There are no grounds at all for being
wary of teachers' subjective judgments, provided they are carried out according
carefully worked out procedures. Teachers' qualitative judgments continue to
lie at the heart of good teaching and good assessment. The issue of appropriate
external controls is taken up in the next section of this Paper.

At this point, something needs to be written about objectivity. Again, two
meanings may be distinguished. According to the dictionary, something is said to
be objective if it is real, belongs to the external world, is observable or verifiable,
and exists independently of the mind.

There are certain aspects of assessment in some subjects (such as mathematics
and sciences) where objectivity is understood quite naturally in more or less its
dictionary sense. There is not a shadow of doubt when chemical equation:
are correctly balanced, mathematical problems solved, or theorems proved. There is
no need to seek confirmation from a chemist, mathematician, or teaching
colleague. That one aspect of assessment is objective does not, however, make
assessment in, say, the sciences and mathematics generally free of subjective
judgments. It is incorrect to extrapolate backwards and claim that because of the
existence of intrinsic objectivity at the end point that all the processes of
assessment are similarly objective and therefore superior to assessment in, say, the
humanities.

At every stage in the design and administration of any objective test constructed
by a teacher, subjective judgments are involved. The teacher has to decide on the
subject matter to include, the behaviours to sample, the complexity and difficulty
of proposed tasks, the item format, and the wording and mode of presentation.
The process is objective only at the very last stage, which is deciding on the
correctness of an answer. So-called objective assessment consists of a chain of
subjective decisions, with one final objective link. Unfortunately the essential
objectivity of the end point, and the fact that the outcome of the final step is often
expressed in numerical form (which to many people is the hallmark of
objectivity) obscures the subjectivity inherent in all the steps leading up to it. So
even in areas of the curriculum traditionally associated with objective
assessment, the issue of subjective judgments needs to be taken seriously.

The second meaning for objectivity is derived from the first, although it is not
identical with it. For qualitative judgments, objectivity has no automatic or self-
evident meaning, in that everyone recognises it as factual, immediately, once it is
pointed out to them. Whatever definition is adopted becomes stipulative. That is,
objectivity for a particular purpose is defined into existence. Sometimes this
concept of objectivity manifests itself in a convention which, in itself, is arbitrary
but which becomes normative (and even binding) through widespread
acceptance and common usage.

For example, an agency which produces standardised tests typically has quite
strict rules and criteria for constructing, trialling, and including items in a test.
The agency may decide to reject all items that have discrimination indices below a
certain value. The fact that such an unambiguous rule exists does not alter the
fact that the critical cut-off value could have been set higher or lower.

The most common definition of objectivity in the social sciences generally, and
certainly in relation to qualitative assessment, is in terms of consensus, or
inter-rater reliability. This accords somewhat with the dictionary definition, because
when a number of competent judges agree on an assessment, the judgment is
made observable, verifiable, and exists independently of (one) mind. This does
not necessarily imply that every single appraisal has to be made, or at least
checked, by a group of teachers. In teaching, it is reasonable to trust a single
qualitative judgment provided that the person making the judgment has been
"calibrated". If a particular assessor's judgments are refined until they are found
to agree consistently with those of a number of other competent assessors, and if
by appeal to the consistency and character of the assessor there is no reason to
believe that the person shows one pattern of behaviour during the calibration
exercise and another when it is all over, the assessment can be usually accepted
with confidence. The fact that a person directly makes a judgment has, of itself,
little bearing on the matter. It should now be possible to see how a single appraisal made by a teacher (even in isolation) could be both subjective and objective at one and the same time!

Improving Qualitative Judgments

At this point, it is convenient to think of teachers-as-qualitative-judges as operating in a dual mode,

(a) as custodians of in-the-head standards, and
(b) as experts in making complex comparisons.

In-the-head standards rely to a great extent on memory, and the degree to which the various levels of performance are internalised by the teacher. Standards internalised by experienced teachers tend to be fairly stable, and for teachers who have participated in discussions and decisions at moderation meetings, to be much the same from teacher to teacher. On the other hand, in-the-head standards for inexperienced teachers tend to decay, or to undergo metamorphosis during evaluative activity. Certain so-called 'serial' effects, for example, are well established in the literature. The standard that an inexperienced teacher uses for grading the first in a set of student essays is often different from the standard used towards the end. Not only may high expectations at the start be unconsciously lowered if it is discovered that very few students perform at all well (a drift in level), but the form of the expectation may change as well (if, for instance, the students do not answer, or misinterpret, the question or nature of the task). It is this malleability of in-the-head standards that makes periodic moderation or recalibration of teacher judgments necessary for inexperienced teachers.

Given that the decision as to a student's level of achievement should be based on the fullest information available, be relatively unaffected by small aberrations in teacher judgment or student performance, and be comparable from school to school, consider the following three conditions.

**Condition A** provides opportunity to make well-considered, grounded assessments with a maximum of reflection and evidence and a minimum of time constraints. This is achieved by spreading the decision over both time and tasks, so that no single piece or testing episode is crucial.

**Condition B** provides mechanisms by which teachers can come to a consensus as to what constitutes performances of different qualities. This is achieved by setting up procedures for consultation, multiple marking, or cross validation, to enable teachers to test their own judgments against those of their peers.

**Condition C** externalizes the standards, removing them from the private knowledge structures of teachers. This makes standards accessible not only to other teachers, but also to non-members of the guild of professionals, namely parents, employers, and most importantly, the students themselves.

Condition A is intended to reduce the discrepancy between what are referred to in the literature as performance and competence. What a student actually does or produces (and therefore what the teacher assesses) is called a performance. Competence is the name given to the underlying degree of proficiency that gives rise to particular performances. Competence cannot be assessed directly, but must be inferred from a series of performances. Reasonably strong inferences can be made by considering a multiplicity of performances. A single performance may not reflect competence accurately for a variety of reasons, among them poor task specifications, sub-optimal conditions under which the students works, and imperfect teacher judgments. Condition A is designed to minimise both random fluctuations in the judgments of individual teachers and irregularities in the performances of students.

But there is a more fundamental philosophical reason for preferring evidence from multiple performances. Competence is broader in concept than a mere 'envelope' of performances, and by definition can be assessed only by giving students opportunity to perform a variety of tasks, under a variety of conditions. Among other things, this provides a better simulation of production conditions in the world outside the school than can ever be achieved under, say, formal (and particularly external) examinations. School-based assessment makes it feasible to think of assessing not only performance, but competence.

Condition B is obviously intended to minimise differences among the intuitive or 'base-level' standards of different teachers, which ordinarily stand a good chance of being idiosyncratic. Condition B corresponds to the definition of objectivity as agreement among competent assessors. The safety-in-numbers principle implies that the greater the number of judges who concur, the greater the objectivity.

As for Condition C, it corresponds more to the dictionary definition of objectivity, as relating to something observable and external to the mind. Although an educational standard is, strictly speaking, always an abstraction (and never an object), the concrete referents associated with it are explicit, and can be discussed and argued about. The most effective way of specifying and promulgating standards is a matter beyond the scope of this paper. In most subjects, however, there are sound theoretical reasons for defining standards by a combination of both verbal descriptions and carefully selected examples (called exemplars).

Earlier it was suggested that in many contexts teachers function in the dual roles of 'carriers' of standards, and as experts in making comparisons. If Condition C is achieved, some of the effort that is required to set up and maintain consensus for in-the-head standards (as under Condition B) can be rechanneled, allowing for more concentration on the actual processes of comparison.

On examination, it will be seen that Condition C is an indirect way to accomplish the same inter-rater reliability expressed in Condition B. Consequently, only one of the two is strictly necessary in an ideal situation. For sound, reproducible qualitative judgments, therefore, the requirements can be summarised as Conditions (A + B) or Conditions (A + C). (In a situation that is less than ideal, it is still fairly safe to place most of the emphasis on either Condition B or Condition C, provided that the other is not ignored completely.) Which one of
Conditions B and C is preferable depends on a number of outside considerations. Initially, Condition C is almost certainly more difficult to realise than Condition B; but there are nevertheless some strong arguments for pursuing it as the preferred option. One reason is simply public accountability. If the standards in a subject are defined somewhere, they can if necessary be made available to persons not involved in teaching or the education system generally. A second strong reason is pedagogical. If students know what the standards are, they know what they are aiming for.

Using a clearly defined system of criteria and standards as outlined under Condition C is unlikely to be more labour-intensive than conventional marking, once teachers become familiar with the process. Certainly one would not expect experienced teachers to be constantly referred to a book of standards for each appraisal they make. Once teachers get used to the idea of assessment by criteria, and accumulate experience in the practice of it, the standards will gradually become almost fully internalised. But the external standards specifications, rather than colleagues, would constitute the ultimate reference framework for assessment decisions. Inexperienced teachers would, of course, make intensive use of the standards specifications. It is likely that the use of explicit standards would considerably shorten the period of time it takes for beginning teachers to become competent assessors.

Assessment policies in Queensland

How do these conditions relate to Queensland secondary schools? Condition A has been available ever since the abolition of external examinations. Under the Radford scheme, Condition B was achieved as district and state moderation committees worked towards achieving consensus on grades. At the back of the moderation procedures, however, was the technical requirement of achieving specified proportions of the seven grade levels over the state as a whole. Note that consensus can be achieved without the need to be explicit about criteria and standards. The ROSBA scheme, on the other hand, marked a move away from in-the-head standards (for these did exist despite the overt policy of norm-referencing, and had to exist for the scheme to work at all) towards explicit criteria and standards. In principle, the replacement of Radford by ROSBA corresponded to a shift from (A + B) to (A + C).
Library Digitised Collections

Author/s:
Royce, Sadler D

Title:
Discussion Paper 5. Subjectivity, Objectivity, and Teachers' Qualitative Judgments

Date:
1986

Persistent Link:
http://hdl.handle.net/11343/115732