Outside/Inside: Criterion-Referenced Assessment and the Behaviourist-Constructivist Dilemma

Dennis Cato, Lachine, Quebec

For an activity so central to education as the assessment of student performance, the absence of consensus about the nature of the activity is a matter of concern. At the level of commonsense, assessment is unproblematic and consists of determining the degree to which a student has "mastered the material," but in determining what it means to have "mastered" something, consensus founders. Assessment presupposes some prior conception of knowledge and how it is validated, and at this level assessment *does* become problematic. On the one hand, emphasis is on the production of knowledge in objective and specifiable behaviours and a standard or scale which determines the level of achievement. From this perspective there can be no measurement of student performance in the absence of overt and specified behaviours. On the other hand, emphasis is placed on the formation of individual constructs of such diversity that they cannot, with any precision, be captured by a standardized or objective scale. Assessment is of the quality of those constructs.

This dichotomy is not a matter of emphasis but of the nature of knowledge and understanding, and it shapes the current behaviourist-constructivist dilemma. For the behaviourist, the product is central. For the constructivist, the process of construct development is central: The product cannot be assessed independently of the process. Assessment for the behaviourist risks being unrelated to the deeper processes of understanding. On the other hand, the constructivist's assessment risks becoming lost in individual and incommensurable constructs. The tasks are to give an account in which construct development may coherently be assessed and to characterize the product in constructivist terms.

The behaviourist-constructivist dilemma has arisen in educational philosophy most clearly in the dispute over the legitimacy of "criterion-referenced assessment." An example of such assessment is to be found in the current demand by the International Baccalaureate Organization (IBO) that the evaluation of students' "Personal Projects" be "criterion-referenced" rather than "norm-referenced." In line with its general assessment philosophy,

> IBO will not pursue a norm-referenced approach to assessment in the Middle Years Program: instead it will aim to achieve a more criterion-referenced approach. That is to say, students will be assessed against defined assessment criteria and not

against other students. ("Assessment Details", nd. p.2)

The "Personal Project" is a research essay, the last step before the awarding of the International Baccalaureate for those completing Secondary V (16-year-olds). Researched and written under the guidance of a teacher "mentor" over a period of two years, it is presented before and assessed by a panel of teachers selected for their familiarity with the subject matter. It is assessed using eight "defined assessment criteria" ranging, in ascending order of complexity, from "Planning and Development" to "Analysis of Information." Each criterion contains four graded levels of "descriptors" which, in the case of the most basic, "Planning and Development," extend from a simple identification of the goals of the project and an outline of how the student aims to achieve them (1 or 2 marks) through the identification and description of those goals (3 or 4 marks) and how the student aims to achieve them (5 or 6 marks) to the determination that the development of the Personal Project is consistent with that description (7 or 8 marks).

There is, however, considerable ambiguity in the descriptors. Where, for example, does a "simple outline of how he/she aims to achieve this purpose" (3 or 4 marks) end and a "coherent description of how he/she aims to achieve this purpose" (5 or 6 marks) begin? Where does the Project being "generally consistent with this description" (5 or 6 marks) stop and being "totally consistent with this description (7 or 8 marks) start? What, exactly, is added to a "coherent description" of aims (5 or 6 marks) by it being "coherent and thorough" (7 or 8 marks)?

To assess the Personal Project in a general or impressionistic sense presents no insurmountable problems. However, since this is precisely what the IBO wants to avoid, two additional criteria, this time for the teachers, are incorporated. Not only is the assessment to be standardized among all teachers on each panel but they are also required to justify their application of the assessment criteria on the basis of that standardization.

> Where several teachers are involved in the assessment of the same subject, it is essential that schools carry out their own processes of internal standardization to ensure that similar standards have been applied to all students [since] teachers will need to be able to justify their application of the Personal Project assessment criteria. ("Assessment Details", p.2)

The demand for the standardization and justification of the application of the Personal Project assessment criteria constitutes a concrete example of the behaviourist-constructivist dilemma. Valid assessment of the Personal Project must be based upon the production of knowledge, upon objective and specifiable performance as embodied in the criteria. However, the application of the criteria to individual performance viewed as the culmination of a unique process of construct

development is not something which could be standardized in an equally objective and specifiable manner. While their views have never been linked with this dispute, the claims of W. James Popham and Andrew Davis, major proponents of the behaviourist and constructivist viewpoints respectively, clarify, if they do not resolve, the dilemma.

Criterion-Referenced Assessment

"Norm-referenced assessment" measures student achievement relative to that of others in the same test group; "criterion-referenced assessment" measures achievement against a continuum from absence to perfect performance. According to Robert Glaser, the American psychologist credited with introducing the concept of criterion-referencing in the 1960's,

> The point is that the specific behaviors implied at each level of performance can be identified and used to describe the specific tasks a student must be capable of performing before he achieves one of the knowledge levels. It is in this sense that measurement of proficiency can be criterion-referenced.[1]

Entailed in criterion-referenced assessment is the specification of both the content and of the relationship between task and behaviour, on the one hand, and the knowledge level within a particular domain, on the other. Each knowledge level is characterized by specifiable tasks that are identifiable with and describable in terms of the specifiable behaviours. The tasks constitute the *criteria;* the behaviours constitute the *performance* used to measure proficiency. For the supporters of criterion-referencing, this specifiability bestows the virtue of articulating clear learning objectives in contrast with norm-referenced tests. According to W. James Popham,

> Prior to the early Sixties most educators had framed their educational objectives in remarkably general and often opaque fashion. An example of such gunky objectives was that 'The student will become conversant with key events in U.S. history.' One of my favorites, encountered in a state's language arts curriculum syllabus, was that 'The student will learn to relish fine literature.' (I never heard of an objective that 'the student will mayonnaise mathematics,' but it sounds equally plausible.) Advocates of behavioral objectives, however, abandoned such generalities in favor of objectives that actually spelt out what behaviors a student should be able to exhibit after instruction.[2]

Popham's objection to "gunky" objectives derived from his view that

> any new tests imposed on educators as part of the educational accountability movement would, of necessity, dramatically influence what was taught. ... If high-stakes tests were going to have an impact on instruction, I committed myself to devise tests that would make assessment's impact as beneficial as possible.[3]

He would make assessment's impact as beneficial as possible by constructing criterion-referenced high-stakes tests.

Rich Knowledge

For some, Popham's vision of behaviouristic, criterion-referenced assessment designed to meet the demands of educational accountability is misconceived in principle. For example, Andrew Davis maintains that to make either teachers or schools accountable for the development of such specifiable behaviours is to misconstrue learning and assessment, pointing out that "school curricula often cannot develop in pupils specific cognitive skills, precisely identifiable, which will directly serve the needs of industry."[4] According to Davis, a "mythology associated with assessment itself" arises from a failure to appreciate the interconnectedness of beliefs. Rather than being discrete stand-alone items, beliefs "are defined in terms of their interactions with other beliefs and with other aspects of rational agents such as desires and intentions."[5] As a consequence, rather than absorbing discrete pieces of information or smoothly acquiring specific cognitive skills, one encounters them with a prior "belief set," a network of interconnected beliefs which both renders such information or skills meaningful and is itself changed as a result of that encounter. This is not just a matter of "empirical Piagetian assimilation and accomodation" but rather

> follows automatically from a simple holism about belief, and might be characterized as a low-level form of constructivism. The latter does not make any radical claims about the non-existence of an independently existing reality; it basically consists of the intuition that pupils learn by building on what they already know; that knowledge, if it is to be understood, cannot be put into minds as though they were empty rucksacks (choose any hackneyed image you prefer).[6]

This "low-level form of constructivism," however, *does* make radical claims about the validity of criterion-referenced assessment. Because pupils build on what they already know, any valid assessment must necessarily engage this process, not focus on the external product of that process. Criterion-referenced assessment is therefore invalid in constructivist terms. "Criterion-referenced assessment linked in any sense to progression through levels," Davis points out, "is incompatible with an appropriate 'constructivist' perspective on children's acquisition of rich knowledge and understanding."[7] Rather than a progression through levels, the appropriate constructivist perspective is rather on "rich beliefs which, when true and held with justification and understanding, count as rich knowledge."[8] For rich beliefs to be transformed into rich knowledge, it is not sufficient that those beliefs be independently true and justified by reasons. They must also satisfy the

understanding condition. For Davis, to possess rich knowledge one must possess "true justified beliefs which are connected in appropriate ways, and the owner of these beliefs [must have] an appreciation of these connections."[9] One might possess rich beliefs but fail to possess rich knowledge if one did not also possess an appreciation of the appropriateness of the connections between the beliefs. Such an appreciation ensures that constructs are not simply the products of learning by rote or by external authority.

The "mythology associated with assessment" arose from the illusion

> that behind behaviour are minds populated by specific and identifiable beliefs, giving rise to the idea that if only we could probe effectively enough we could find out what is there. The reality is a more complex and elusive situation in which interpretations are made of the mind-states of others. These interpretations require many assumptions which it would be difficult to make wholly explicit.[10]

Criterion-referenced assessment focuses on specific behaviours, low-level constructivist assessment on determining the pupil's possession of rich beliefs and an appreciation by their owner of the appropriateness of their connections. These interpretations require many assumptions difficult to make wholly explicit, principal among which is a simple holism about belief. But how will the determination of appropriateness in a constructivist assessment be made?

The Well-Defined Behavioural Domain

For Popham, "a particularly perplexing issue facing criterion-referenced test specialists involves a decision about the generality of the behavioral domain being measured."[11] To assess domains beyond low-level routine performance successfully, test specialists must "spell out" specific items eliciting appropriate behaviours. "Complex behavioral domains," however, render the specification of such appropriate behaviours unmanageable. By way of a solution, Popham proposes "derivative homogeneity:"

> Try to set down such constraints in the fields of history or literature, to choose only two of many, and you'll either fall short or be obliged to create enclyclopedic specifications. No, while the items for a good criterion-referenced test should possess derivative homogeneity, they need not possess functional homogeneity in the sense that examinees answer them all correctly or incorrectly.[12]

Derivative homogeneity requires specific behaviours derived from the particular domain and reflecting the appropriate level. But in complex behavioral domains such as literature and history, the difficulty is to explain how derivative homogeneity would itself be derived.

Popham's first attempt involves turning the problem back to front.

> Given most people's willingness to tolerate descriptive information, it makes more sense to write test specifications so they subsume all or most of the difficult levels of test items associated with that behavioral domain, *then subsequently judge* the extent to which the items are derivatively homogeneous.[13]

Willingness to tolerate descriptive information in the form of a proliferation of test specifications depends, one supposes, on whether that descriptive information makes sense, but in the present case it does not. If constraints for complex domains like history and literature could not be set down *before* the test specifications, it does not make sense to judge the extent to which those items are derivatively homogeneous *after* they have been written. To write the test specifications presupposes prior possession of the criteria to judge which items are derivatively homogeneous, and it is these criteria that Popham fails to reveal.

Popham's next attempt consists of a "limited-focus strategy" in which

> we attempt to isolate a small number of high-import behaviors to be measured, even though such behaviors turn out to be quite complex. This means that we must think of truly significant examinee behaviors that subsume more elementary behaviors.[14]

But how would he identify a few "high import behaviors" which subsume those which are more elementary independently of criteria of derivative homogeneity? His "limited-focus strategy" still presupposes prior possession of the criteria in terms of which those "high-import behaviors" might initially be identified. The consequence is that ambiguity in determining the generality of the domain being measured extends from descriptions of test specifications to the method of selecting among competing test items:

> For practicality's sake, test specifiers usually have to make their best guess as to the generalizability of competing measurement tactics [and] it seems that test specification folks are going to have to engage in some pretty shrewd estimating of a potential measurement tactic's generalizability.[15]

Without criteria of derivative homogeneity it is not clear how to make that "best guess" or engage in that "pretty shrewd estimating." This is not a minor difficulty. Popham's "well-defined behavioral domain" which would serve to spell out those behaviours a student should be able to exhibit after instruction is empty, and his program of behavioral criterion-referenced assessment collapses. Popham appears to concede as much.

> Unfortunately, perhaps because of the recency of our work with such specifications or perhaps because of the nature of the task itself, we do not yet possess a refined and tested set of rules to guide those who must create criterion-referenced test specifications. Measurement folks have been thrashing around, trying to get a fix on the kinds of test descriptions that will prove effective.[16]

The absence of refined and tested rules to guide criterion-referenced test specifications has less to do with the recency of the work than with the nature of the task itself. There is no such refined and tested set of rules which will exhaustively spell out those behaviours corresponding to the tasks which mark out the progression of ability within complex behavioral domains. Popham's attempt to reduce understanding in such domains to its overt and objectively specifiable manifestations misses that which gives such manifestations their meaning: the context in which they are embedded and out of which they emerge. Any attempt to specify the particulars of that context in behavioral terms leads not to meaning but to incoherence.

Appropriate Content Appropriately Connected

For Davis, assessment consists of interpreting the pupil's "rich knowledge," that is, possession of true justified beliefs distinguished by understanding and appreciation of the appropriate ways in which the rich beliefs are connected. However, the determination of what is *appropriate* without reference to an independently existing reality creates difficulties. Where Popham's dilemma arose from his behaviourist focus on the *outside* performance without reference to understanding, Davis' dilemma arises as a result of his constructivist focus on the *inside.*

To illustrate the necessary imprecision of written, criterion-referenced assessment of rich knowledge, Davis asks how we determine what a pupil knows about Faraday's theory of electromagnetism. Davis' short answer is that we cannot, at least with the degree of precision required by the imperatives of educational accountability because "we cannot attribute Faraday *beliefs to our pupil unless we assume her possession of an indefinite number of other beliefs with appropriate content."*[17] It is not their correspondence with the independent reality of electromagnetism that is of primary concern but the assumption that, in addition to her new Faraday beliefs, she possesses an indefinite number of other beliefs with appropriate content. If she really understands Faraday's theory rather than having learned it by rote, "these other beliefs are appropriately connected to her Faraday *beliefs*, and... she correctly identifies the nature of these connections."[18] Indeterminacy in assessment therefore derives from three assumptions: that she possesses an indefinite number of other beliefs with appropriate content, that her new Faraday beliefs are appropriately connected to those other beliefs, and that she has an appreciation of the appropriateness of, or can identify the connection between, her appropriate other beliefs and her new Faraday beliefs.

Difficulties initially arise in connection with the relation of other beliefs to new beliefs, specifically the criterion which will distinguish appropriate from inappropriate other beliefs. In the absence of an independent reality, in respect to the existence of which low-level constructivism makes no radical claims, other beliefs can only be judged appropriate either by virtue of their being elicited by the new beliefs or by sanctioning beliefs among the pupil's other beliefs.

If the new beliefs provide the criterion for selection they are automatically self-validating, accepted on authority and not rich knowledge, or meaningless. Where new beliefs simply elicit supporting beliefs from the pupil's other beliefs, they are automatically "appropriate." Correspondingly, the failure of new beliefs to elicit warranting beliefs would make those other beliefs "inappropriate." The pupil who has no beliefs about magnetism or electricity embraces Faraday without understanding and fails to attain rich knowledge. Such new "beliefs" are necessarily meaningless. The relation between a set of new beliefs and the pupil's other beliefs in the tripartite construct is simply one of tautology. Other beliefs are appropriate when they endorse the set of new beliefs since that is what it means to be "appropriate." Where they do not, they are not "inappropriate" but do not arise at all. They are irrelevant. The consequence is either tautologically self-validating or the rejection of new beliefs.

Davis' pupil either automatically accepts new beliefs or automatically rejects them. In either case she does not acquire rich knowledge. She could not, by definition, inappropriately connect them as the appropriateness of her other beliefs has been antecedently endorsed. She could not appropriately connect inappropriate other beliefs to her set of new beliefs since they are irrelevant to those new beliefs. Connecting other beliefs with appropriate content to her set of new beliefs can only be unerring where it is not impossible.

To transform her rich beliefs into rich knowledge, the pupil must appreciate the appropriateness of the connections between her set of new beliefs and her other beliefs possessing appropriate content, thereby completing the tripartite construct. Viewed from the inside, such appreciation must necessarily flow from what it means to be appropriate. Possessing other beliefs with appropriate content and having appropriately connected them to her set of new beliefs, Davis' pupil could not fail to appreciate the appropriateness of their connections or to correctly identify at least some of the connections between them. Such identification and appreciation stand in tautological relation with the appropriateness of her achievement.

Assessing the presence of rich knowledge is equally problematic. How will Davis' assessment detect the pupil's identification and appreciation of appropriate

connections linking her set of new beliefs with her appropriate other beliefs? It is not clear how he has avoided his own "mythology of assessment," that behind behaviour are minds populated by specific beliefs giving rise to the idea that if only we could probe effectively enough we could find out what is there. If his assessment, as he has claimed, is more complex and elusive - involving as it does interpretations of the mind-states of others, interpretations which require assumptions difficult to make wholly explicit - it falls to Davis to give some account of the nature of those assumptions and the process of such interpretation. Davis never does this. His pupil's acquisition of rich knowledge flowed unerringly from his "holism of belief" and his intuition that pupils learn by building on what they already know but only because the elements of that holism were tautological.

Conclusion

Davis is, as far as I am aware, the only constructivist, "low-level" or otherwise, to have directly engaged criterion-referenced assessment, the principal reason for my interest in his views. He claims that his "low-level constructivism" is not to be considered even agnostic in respect to the existence of that independently existing reality and its role in assessing rich knowledge, that he was never a "radical constructivist" in the fashion of Ernst von Glasersfeld[19] for whom knowledge can never be of an observer-independent reality but only of our own individual cognitive structures. The consequence is that Davis is committed to a constructivist pedagogy while adhering to a traditional realist ontology and epistemology. While this position was dubbed "trivial constructivism" by von Glasersfeld[20] an even stronger criticism might be made on the grounds of theory-practice incoherence since there is apparently no connection between his ontology and epistemology on the one hand and his pedagogy on the other.

Assessment becomes problematic where the question as to what it means to have mastered the material arises. On the one hand, where the activity is viewed from the outside in terms of overt behaviours which are judged to have satisfied standardized criteria, the difficulty arises in giving an objective account both of those criteria and the manner in which they are to be applied without reference to the inside, to some account of the process of understanding distinctive of a particular domain. On the other hand, where assessment is viewed from the inside in terms of the pupil's process of understanding and the construction of new beliefs in the context of other beliefs without reference to an independently existing reality, the result can only be either a simple endorsement of the pupil's belief system or incoherence. Any future assessment must bridge the gap and incorporate both the

outside and the inside.

Acknowledgement

I am grateful for the scrupulous attention given to this paper by an anonymous reviewer at *Paideusis*.

Notes and References

1. Glaser, Robert (1963) "Instructional technology and the measurement of learning outcomes: some questions." *American Psychologist*, 18.2, 519-521. p. 519.
2. Popham, W. James (1998) "Farewell curriculum: confessions of an assessment convert." *Phi Delta Kappan*, 79.5, 380-394. p. 381.
3. Ibid. p.384.
4. Davis, Andrew (1995) "The limits of educational assessment." *Journal of Philosophy of Education*, 32.1. p.9.
5. Ibid. pp.64-65.
6. Davis, Andrew (1995) "Criterion-referenced assessment and the development of knowledge and understanding." *Journal of Philosophy of Education*, 29.1, 3-22. p.6.
7. Ibid. p.20.
8. Ibid. p.6.
9. Ibid. p.7.
10. Ibid. p.68.
11. Popham, James W. (1978) *Criterion-Referenced Measurement.* Englewood Cliffs, N.J., Prentice-Hall. p.94.
12. Ibid. p.98.
13. Ibid. Italics in original.
14. Ibid. p.117.
15. Ibid. p.120.
16. Ibid. p.131.
17. "The limits of educational assessment," op.cit., p.68.(Italics in original).
18. Ibid. p.117. (Italics in original)
19. See Ernst von Glasersfeld (1995) *Radical Constructivism: A Way of Knowing and Learning.* Washington. The Falmer Press. Andrew Davis was present when "Outside/Inside: Criterion-Referenced Assessment and the Behaviourist-Constructivist Dilemma" was given at the annual conference of the Philosophy of Education Society of Great Britain, April, 2000. His views are those expressed at the conference and in subsequent correspondence.
20. See Ernst von Glasersfeld (1989) "Knowing without metaphysics: aspects of the radical constructivist position." Kitchener (Ontario). *Kitchener-Waterloo Record.* Also available from *ERIC REPORTS*, Washington, D.C. #304 344.

Author

Dennis Cato is a retired secondary school History teacher living in Lachine, Quebec. He received his Ph.D. in Philosophy of Education from the University of Ottawa and has contributed to a number of journals in the field, primarily in the areas of epistemology, learning theory, and pedagogy.