# The educational scientist's dream

by Kyle Smith, February 16, 2017

## I never wanted to be in the educational scientist's dream

In my previous post, I included Glenn Fulcher's definition of 'assessment literacy', part of which is

> *The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order in order [to] understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals.*

For those wishing to develop such an ability in themselves or others, Elliot W Eisner's 1976 paper, 'Educational connoisseurship and criticism: Their form and function in education evaluation', is a valuable resource and I'd like to share some sizeable excerpts from it.

> *since the turn of the [20th] century, since the early work of Edward L Thorndike, there has been a strong aspiration among psychologists to create a science of education which would provide educational practitioners – administrators as well as teachers – with the kind of knowledge that would permit prediction through control of the process and consequences of schooling. Laws that would do for educational practitioners what the work of Einstein, Maxwell, and Bohr have done for physicists were the object of the educational scientist's dream. This yearning for prediction through control was, of course, reflected in the desire to make schools more efficient and presumably more effective. Educational research was to discover the laws of learning that would replace intuition and artistry with knowledge and prescribed method. … This aspiration to discover the laws of learning was allied with the efficiency movement in education[described by RW Callahan in his 1962 book, Education and the Cult of Efficiency] that sought to install scientific management procedures in schools … (Reimagining Schools, 2005, p. 38)*

Some questions to consider:

- Do you share this 'educational scientist's dream'?
- How do you feel about the dialectic of prediction and control (i.e., if you can control something, then you can predict it and vice versa) being transferred from physics to education?
- What have the consequences been over the last hundred-odd years of this 'scientific and technological approach' to education?

Eisner describes "four major deleterious consequences".

> *First, because scientific assumptions and scientifically oriented inquiry aim at the search for laws or law-like generalizations, such inquiry tends to treat qualities of particular situations as instrumentalities. The uniqueness of the particular is considered "noise" in the search for general tendencies and main effects. This, in turn, leads to the oversimplification of the particular through a process of*

*reduction aimed at the characterization of complexity by a single set of scores. Quality becomes converted to quantity and then summed and averaged as a way of standing for the particular quality from which the quantities were initially derived. For the evaluation of educational practice and its consequences, the single numerical test score is used to symbolize a universe of particulars, in spite of the fact that the number symbol itself possesses no inherent quality that express the quality of the particular it is intended to represent.*

## The 'measurement fallacy'

Think of our reliance on 'single numerical scores' such as an 'IELTS 4.5' or something that we claim is 'equivalent'. Is a student who scores 4.5 half as proficient as one who scores 9? I suspect, if asked, most people familiar with or involved in IELTS testing would say, 'No, of course not.' But this has an important implication which is perhaps not widely appreciated. To quote Dalziel (1998, p. 353):

*Numerical symbols [e.g. 4.5] can only be used as numbers if there are good reasons to believe that there is a direct correspondence between the empirical property being assessed (student performance on a task) and the 'real numbers' of mathematics.*

For such a direct correspondence to exist, the 'empirical property' we're interested in – perhaps grammatical accuracy – must be quantitative.

*Only some properties are quantitative (such as length, weight, density, etc.), whereas some are not (name, perceived colour, etc.) … for a property to be quantitative, it must exhibit both order and additivity. Order on its own is not sufficient evidence that a property is quantitative, hence the fact that students' work can be ordered from excellent to poor does not in itself provide sufficient evidence that the assessment performances are quantitative, and therefore can be assigned numbers (that is, meaningful numbers). For assessment to be quantitative, both order and additivity must be present.*

*The problem with these requirements is that additivity is only directly observable for a small set of quantitative attributes (such as length and weight), and yet many attributes appear to be quantitative even though no direct test of additivity is available (e.g. density). This problem is particularly applicable to psychological and educational measurement, as many researchers believe that the variables they study are quantitative, but it is unclear how psychological and educational variables could be directly tested for additivity. That is, how can one combine two essays worth four out of 10 together and show that the combination is equal to one essay worth eight out of 10? (p. 354)*

In fact, according to Dalziel, in 1940, "a body of esteemed scientists gathered by the British Association for the Advancement of Science" sought to determine "if measurement was possible within psychology" but concluded that it was not possible because there "was no evidence that any of the variables studied by psychologists were quantitative"; and in the absence, says Dalziel, "there is no basis for the use of numerical scores as numbers" (p. 355).

*Numerical scores are misleading, because they imply that operations such as addition, averages and scaling can be used with the scores when there is not*

> *sufficient empirical evidence to justify these procedures. These problems are most evident in the practice of aggregating marks to determine final scores. (p. 356)*

Following this presentation of the problem of the use of numbers in education and elsewhere, Dalziel presents several 'simulations' to illustrate the fact that "there are many different possible ways of regarding the symbols used to represent assessment performance, and that depending on the assumptions made, very different outcomes can result" (p. 363).

- Simulation 1: Scores on 17 individual assessment items (e.g. essay, report, tests) are combined "in a weighted sum calculation of the final raw mark" which is then scaled to fit in with a grading system (High Distinction, Distinction, etc.)
- Simulation 2: Scores are not truly quantitative and cannot be aggregated in any way to produce an overall grade
- Simulation 3: Scores are not truly quantitative but a rule such as 'choose the median grade across grades for all 17 assessment items' can be applied to extract a final grade
- Simulation 4: Scores are considered quantitative and can be added but, unlike Simulation 1, the distances between grades (e.g. High Distinction and Distinction, or Distinction and Credit) is not assumed to be equal; an arbitrary rule was applied to determine the distances between the different grades
- Simulation 5: Same as Simulation 1, but measurement error for some assessment items is factored into the calculation of the overall grade.

When compared to Simulation 1, the percentages of students with a different final grade were as follows:

- Simulation 2: Not comparison was possible because of the assumption that the scores are not quantitative
- Simulation 3: 24%
- Simulation 4: 8%
- Simulation 5: 12.6%

> *The point of these simulations is to illustrate just how tenuous any one system of determining final marks is and that without evidence that the variables involved are quantitative, any number of potential systems for producing marks and grades are equally valid (and, potentially, equally wrong, in that sense that they do not represent anything quantitative in the first place). Hence a student may score on mark or grade under one assessment system, but may receive quite a different mark or grade under another system, and the critical determining factor in which mark or grade is received is not just the student's academic performance but also the assumptions of the assessment system used by the course administrator. The seriousness of this situation should not be overlooked, and the decisions of course administrators should be recognised as powerful influences on the assessment outcomes of students.*

In response, Knight (2002) advises us to "beware of the numbers created by summative assessment and mistrust conclusions based on the transformation or manipulation of those numbers" (p. 281) and suggests "placing psychometrics under erasure while revaluing assessment practices as primarily communicative practices" (p. 285).

Similarly, Yorke (2011) writes:

> *A belief in assessors' capacity to measure student achievements implies that the assessor can utilise the kind of measurement that is typical of science (say, with respect to mass, length and time) in which numerical measurement has over centuries demonstrated its power. This belief might be termed the 'measurement fallacy'. Attempts to measure student achievement are scientistic rather than scientific … [and] any pretensions to measurement (such as those implicit in the computation of grade-point averages …) should be abandoned. (pp. 253-254)*

## Eisner's other three major deleterious consequences

Back to Eisner's 1977 paper.

> *Second, the technological orientation to practice tends to encourage a primary focus on the achievement of some future state and in the process tends to undermine the significance of the present. (p. 38)*

Think here of the 'washback' of the black-boxed, summative, psychometric, MCQ approach to language assessment: how much frustration does it cause students and teachers as they try they guts out every day in the classroom only to have their hopes and motivation dashed on the rocks of yet another tedious standardised test? Think also of the amount of effort put into designing and validating the test before it's used.

> *Third, scientific and technological approaches to schooling lead, as I have already said, to the attempt to "objectify" knowledge. Objectification almost always requires that at least two conditions be met. First, the qualities to which one attends must be empirically manifest, and second, they must be convertible to quantity. In this way both reliability and precision can be assured, hence conclusions about a state of affairs can be verified. (p. 39)*

To Eisner, this ultimately means that "the opportunity to understand empathically and to communicate the quality of human experience diminishes" (Ibid.).

> *Fourth, when one seeks laws governing the control of human behavior, it is not surprising that one would also seek the achievement of a common set of goals [you could also think of 'learning outcomes' here] for that behavior. When one combines this with the need to operationalize such goals quantitatively, the use of standardized tests becomes understandable. The standardized test is standard; it is the same for all students. It not only standardizes the tasks students will confront, it standardizes the goals against which they shall be judged. These tests, de facto, become the goals. When this happens, uniformity becomes an aspiration; effectiveness means in practice that all students will achieve the same ends. Individualization, regardless of what it might mean, becomes defined in terms of providing for differences in rate; differentiation in pace rather than in goal, content, or mode of expression is the general meaning of individualization. … In a technological orientation to educational practice, the cultivation of productive idiosyncrasy … becomes a problem. (Ibid.)*

To paraphrase Peter Knight, our assessment practices are in disarray and the more we obsess about correlations and coefficients, the longer they will remain in disarray and the more frustration we will cause our teachers and students at the expense of their learning.