http://dx.doi.org/10.1080/0969594X.2017.1319338

A response to "Assessment and learning: fields apart?"

I welcome the opportunity to respond to this paper (Baird et al., 2017), which covers a number of important and controversial issues relating to the quantification of educational assessment. The section on assessment for learning covers less ground and does not, for me, raise any substantial issues, and my remarks are therefore directed entirely towards the quantitative discussion.

My commentary is in two parts. In the first I offer some general comments on the Baird et al. paper. The authors make a number of default, and indeed common, assumptions about quantitative assessment that on the whole go unquestioned, and yet in my view are highly germane to a proper understanding. Because I consider these to be important, my commentary, in places, may seem harsh, but in fact is designed to identify where I believe more careful consideration is needed.

The second part looks in more detail at the role of international large scale assessment (ISLA) by exploring the issues that the authors mention, as well as others.

The stated objective of the paper is to elucidate the relationship between theories of learning and theories of educational assessment. Yet there is very little in the paper that directly addresses this. There is certainly a discussion about the relationship of curriculum to assessment and, for example, criticism of the PISA approach to curriculum, but one looks in vain for a good example of how a specific theory of how people learn should determine how to construct a quantitative assessment. The authors mention learning theory 'foundations for assessment', but there is no clear example of what this might look like. They also in many places refer to 'assessment theory' but the only places where there seems to be any attempt to describe such theory is in terms of psychological constructs – which are not the same thing, and the discussion of assessment constructs is at a very general level with plenty of references to the literature, but no good exemplars. The authors do have some discussion about what quantities (numbers) mean, but this is not specific to educational assessment, and offers no fresh insights.

My own view is that assessment is essentially a technology that can have no theoretical basis of itself but seeks to be motivated by the area to which it is applied, in this case education. It is therefore somewhat fruitless to pursue a 'theory' of assessment as such in the same sense that one might seek to develop a scientific theory that has sufficient generality so that it applies across all contexts where assessment takes place with a sufficient level of abstraction, and has explanatory power that can be verified/falsified by empirical observation. Rather, what one should be doing, in my view, is attempting to develop assessment instruments which reflect the concepts they are supposed to be measuring. Of course, as a technology, or perhaps a collection of technologies, assessment will have some basic procedures that practitioners will wish to adhere to. These may be guidelines on how to construct assessments or tests, for example on the role of group differences such as whether a test should balance the average responses of males and females. Or they may be to do with how test scores are analysed. Such practices do not however, combine to constitute a theory.

There is of course the point that the use of any particular assessment instrument of itself defines or 'becomes' the thing we wish to measure, and that is a legitimate area of debate and we may very often wish to view assessments as 'tacit'. Just as 'tacit knowledge' may be very useful in certain circumstances, so 'tacit assessment' may have an important role. Thus, suppose we desire a measure of children's understanding of probability. This might be based upon answering a set of questions involving the manipulation of a few basic rules using the context of dice throws. There is no inherent 'assessment theory' involved but the technology of how to present and evaluate the responses can be conceived in various ways. Any 'theory' that might guide this technology, if we

wish to seek a theory, will derive from assumptions about how children come to learn about and apply probability. Nevertheless, it is not at all clear that a theoretical justification of any kind is actually necessary for an assessment to be useful. In society there are many examples of measurements that have only a limited theoretical justification, such as a retail price index, yet perform an important (if contestable) function.

The authors do spend some time discussing psychometric approaches. While this is understandable given the historical dominance that field has had within education, it is in effect a side issue. Moreover, while criticism of certain psychometric approaches in terms of restrictions upon dimensionality is certainly justified, it is a pity that there is not a more general critique in terms of how validity is assessed, how item translation is attempted, how comparability (equating) is carried out and other aspects of the practice of quantitative assessment design. Post second-world-war psychometrics has tended to become staffed by a largely self-serving 'priesthood', producing material opaque to non-specialists. It is not that psychometrics is 'atheoretical' that is the problem, rather that many psychometricians seek to dress up what they do as 'theory'. It is this that Goldstein and Wood (1989) really complained about.

It is almost as if in order to obtain academic or intellectual 'status' many practitioners involved in educational assessment feel a need to justify their practices as 'theoretically' driven.

My major reaction to this paper, therefore, is one of disappointment. While there is some justified criticism of aspects of quantitative assessment, for example, with respect to PISA, the authors not only throw little light on their stated aim, but an amount of what they have to say seems often to be based on misunderstandings of what is involved in quantification. This is unfortunate and does not really move the subject forward in any useful way. In the remainder of this commentary I shall pick up some more of these issues, because I believe that they may give rise to confusion, and do need clarification.

In section 2 the authors talk a about educational assessment as being 'normative' which they suggest is to do with 'affect(ing) the attribute being assessed'. On the contrary, while assessment may be used to affect learning, for example by setting test score 'targets' for teachers, quantitative assessments do not have to be constructed on that basis – they may be designed simply as monitoring devises or instruments to evaluate an educational reform. Indeed there are many other ways to devise targets to create what is taught and learnt.

In section 3 the paper discusses 'constructs' and suggests that the field of quantitative assessments is 'riddled with assumptions'. That is quite so, as are all fields, and they need to be in order to get anywhere at all! They also say that there is 'remarkably little reflection on the assumptions underpinning them (constructs)'. It seems to me however, that tests routinely make claims for the constructs they are attempting to measure, and it isn't clear what the authors are aiming at here. They also misquote my 2012 paper which discussed Galton (who invented the correlation coefficient, not Pearson) where I specifically argued that in fact Galton was a poor scientist, preferring his data to fit his theories rather than framing his theories so that they could be empirically substantiated, or not, by the data. The authors' introduction of 'policy-driven' constructs is useful, but they fail to point out that in a broader sense, constructs also reflect social and cultural norms. What this implies is that different societies and cultures will generally assume different constructs and hence assessments. It is this that is discussed as a major criticism of the internationalisation of assessment such as seen in PISA, and it is a pity that it gets no mention; I shall return to this.

I find section 5 unclear. In the first paragraph the authors seem to conflate 'unidimensionality' with 'model fit'. Yet the first is an assumption underlying a particular set of psychometric models and the second is a statistical device for testing an assumption about any model. The second paragraph appears to be making a similar point about statistical power. This distinction matters, since most

readers will not be familiar with the technical terms used. The third paragraph makes little sense to me and I wonder if the authors would like to expand on this in their response. The fifth paragraph conflates 'unidimensionality' with 'local dependence', and the sixth paragraph seems to imply that 'constructs' are 'variables', which they are not.

In section 6 in the first paragraph the authors in their example suggest that a 'first class degree is any score over 70 and a second class degree is 60-70". They then pose a question: "what if the attainment required to move from 66 to 71 is greater than to move from 60 to 65?" First, what is meant here by 'attainment'? Secondly they seem implicitly to be arguing that on the chosen test score scale, converting a score to a grade should only be done for equal score ranges, yet they provide no justification for this. Moreover, and typically, test score scales are arbitrary and can be, and often are, transformed non-linearly while preserving the score order. The authors seem to find the combination of scores over several dimensions troubling, presumably because there is an element of judgement in how to 'weight' the components. While it does seem to be generally true that test constructors claim 'objectivity' for their techniques, in practice this is a chimera and judgements are always involved, explicitly or implicitly (Goldstein, 1989). Thus, the problem disappears once one recognises the role of judgement in quantitative assessments. In paragraph 4 the authors say "The question then, is not whether educational attainment exists as quantity in people, but whether attempts at quantifying are useful." If educational attainment does exist how can we possibly know whether it does except by some way of measuring it, even if only crudely?

In section 7 paragraph 5 the authors seem to equate 'assessment theory' with 'assessment model' and the remainder of the argument of this paragraph gets lost in trying to distinguish assessment from psychometrics. In the following paragraph they appear to be making a claim for assessment being criterion referenced, malleable, focussed on validity etc, and argue that "learning theories and measurement model paradigms are in opposition", which is a very strong claim, made with no supporting evidence, and also, surely, not true?

Section 8 deals with international large scale assessments (ILSA), and does indeed make many useful points. The authors rightly criticise the unavailability of the actual items used in the test and discuss the fundamental problems associated with comparability across educational systems. They somewhat miss the point, however, about the use of plausible values, which may have certain technical drawbacks, but are a perfectly legitimate means of carrying out statistical analyses and are a particular case of the use of 'multiple imputation' for handling missing data. They also miss one of the most important limitations of PISA and similar studies, which is that they are essentially cross-sectional and hence unsuitable for attempting to draw causal conclusions, and it is perhaps this fact that best explains the limited use of PISA by researchers, and I will return to ILSA studies below.

More generally, I believe that there is, throughout the paper, an unwillingness to engage with the meaning of 'quantitative assessment'. At a basic level, one can describe all forms of assessment as quantitative, since assessment implicitly assumes that a judgement is being made, even if it only a binary one, such as 'has understood/not understood a concept' or 'has reached/not reached a target' or 'is ready/not ready to move to another stage of understanding'. From these one can move to more elaborate assessments such as 'has reached a satisfactory/unsatisfactory/very satisfactory…level'. More elaborate still an assessor may decide to combine (average) assessments over several observed responses to produce a scale: perhaps just an ordinal one. A statistical analyst may then wish to relate such a scale, to other factors such as ethnic and social identities or measures of poverty.

From what they say, the authors appear to regard much of quantitative assessment as consisting of such things as item response models, but this is to miss the point about these. Item response models, which are special cases of more general factor analysis models, are attempts to combine in a small number of summary measures, a large number of separate item responses, typically found in

standardised tests. This is a perfectly legitimate activity, although they are, of course, criticisable both technically and substantively. They should, however, not be regarded as the only kind of quantification that takes place. The act of obtaining a response, as I have argued, is an attempt at quantification that all assessors are involved in. The act of combining and relating these responses is not solely the province of psychometricians, but should be of interest to all assessment practitioners, including those involved in assessment for learning. The latter activity may be focussed largely on formative processes but that does not release it from a concern with quantification and analysis of its measurements.

Let me turn now to ILSA studies, notably PISA. It is a pity that the authors do not provide a more comprehensive critique of PISA. There is a growing literature on how the results have been used by policymakers to achieve certain goals, and a study of these has a great deal to say about how assessment results are used in general. In particular the use of assessment as a form of social control ought to be one of the most important concerns of assessment professionals. We see this clearly, for example, in the use of school league tables within a competitive educational marketplace, to the detriment of the educational process itself. We have seen a similar phenomenon happening in the higher education sector, where the UK is a leader in attempting to establish a crude marketplace where 'consumers' are expected to choose institutions on the basis of a small number of poorly conceptualised indicators of performance (Hazelkorn, 2011).

The authors point out that PISA results have had relatively little research exploitation. I suggest that this may in part be to do with the fact that the data are cross sectional so that there is no satisfactory way of measuring 'progress' through secondary schooling. Over the last few decades researchers into the effects of schooling has come to accept that a minimum requirement for drawing inferences of a causal nature is to have longitudinal data. Yet, OECD has not encouraged the use of longitudinal data as part of its routine data collection. While this would clearly involve extra resources, the scientific gain would be large, and one is tempted to infer that OECD has relatively little interest in pursuing such scientific knowledge in comparison with its desire to provide crude international comparisons. Indeed, OECD has been described as attempting to establish itself a 'world ministry of education', controlling what is taught by controlling a common assessment set of instruments and their use for the comparative evaluation of educational systems.

Another reason why PISA may have had relatively little exploitation is that access to the items used in the tests is restricted, with only up to 15% or so released to secondary analysts. The reason given for this is so that some of the items can be used to measure trends over time using equating procedures, and hence should remain 'secure', that is to be publicly unavailable. This is in fact rather a weak argument. First, it is doubtful whether meaningful comparisons over time are actually possible when curriculums are changing and hence also are the relative characteristics of the test items such as their difficulty. Denying users access to the test items makes it difficult, if not impossible, properly to assess what is being measured, in particular whether any given item is appropriate for any given country. It denies the analyst, unless they happen to have privileged access within their own country, the opportunity for any detailed comparisons based upon individual items, which is presumably a further reason for its under-utilisation. Most national assessment systems do in fact release the items in order to comply with demands for transparency, and there is no sound reason why OECD should also not do this.

With respect to PISA, I would suggest that another barrier to widespread research utilisation lies in the relative complexity of handling the data. OECD provides 5 'completed data sets' for each area and sub-topic area. Each data set contains a different set of imputed (plausible) values and any statistical model has to be fitted to each data set and the parameter estimates combined at the end to provide the final results. For an analysis of any real complexity this can be time-consuming.

Furthermore, 5 such datasets is almost certainly not sufficient when fitting multilevel models (Goldstein, 2011) and where a multilevel model becomes complicated involving random coefficients, the imputed values themselves will generally be inadequate, resulting in biased estimates. It would be technically sounder to release the data without any imputations, but of course indicating where the appropriate values were missing, either by design or inadvertently, so that the analyst could carry out their own, more valid, imputation. Of course, this would involve yet more complexity, but there are software tools suitable for handling complex imputation based analyses, for example STATA (Statacorp, 2016) and STATJR(Centre for Multilevel Modelling, 2016). OECD could provide guidance and exemplars to assist users, and ultimately this would be a more useful course to take.

Finally, I agree that it is certainly highly desirable, within education, that assessment fully specifies what is being assessed in terms of what is being learnt and what students are being encouraged to learn. If assessment is to have any 'theoretical' content, however, this should derive from its relationship with learning and is not something that needs to be pursued in its own right. Indeed, psychometrics has done a great disservice to assessment by claiming that in some sense it's procedures are theoretically based, as in the term 'Item Response Theory' (Goldstein and Wood, 1989). If educational assessment is to find an appropriate niche within educational thinking, it needs to emerge fully from the shadow of psychometrics and eschew any pretensions to own a 'theory' in its own right.

Harvey Goldstein,

University of Bristol

March 2017

Reference

Jo-Anne Baird, [a] David Andrich,[b] Therese Hopfenbeck[a] and Gordon Stobart. (2017). *Assessment and learning: fields apart*? Assessment in Education (to appear)

References

Centre for Multilevel Modelling (2016). http://www.bristol.ac.uk/cmm/software/

Goldstein H & Wood R. (1989). Five Decades of Item Response Modelling. . British Journal. of Mathematical and Statistical Psychology, **42** 139-167.

Goldstein H. (1989). Equity in Testing After Golden Rule. Paper read to American Educational Research Association meeting San Francisco, March 27-31.

Goldstein, H. (2011). Multilevel Statistical Models. Fourth edition. Chichester, Wiley.

Goldstein, H. (2012). Francis Galton, measurement, psychometrics and social progress. Assessment in Education, 19,2.147-158

Hazelkorn, E. (2011). *Rankings and the reshaping of higher education*. Basingstoke, Palgrave MacMillan.

Statacorp (2016). www.stata.com