BSRLM Working Group: Using statistics in mathematics education research

Andy Noyes[1], Jeff Evans[2], David Pepper[3], Jeremy Hodgen[3]

[1]*University of Nottingham,* [2]*Middlesex University,* [3]*Kings College London*

This working group has been meeting for about a year at BSRLM day conferences. We focus on two surveys discussed at the March 2014 conference: PIAAC (aka Survey of Adult Skills) and PISA. Jeff Evans outlines some basic points to look for in the methods used in educational surveys, and illustrates these in relation to the PIAAC adult skills survey results. He argues that we need to distinguish three aspects of validity: construct validity, internal validity, and external validity. David Pepper and Jeremy Hodgen outline the OECD validation of PISA and argue that it is not sufficient for the proposed high stakes use of the assessment. They focus on PISA's assessment of student confidence in mathematics.

Keywords: statistics, social survey, numeracy, confidence, self-efficacy, validity, construct validity, internal validity, external validity

Using statistics in mathematics education research – Andy Noyes

In October 2012 the Royal Statistical Society hosted a small seminar to discuss the use of statistics in mathematics education research. Harvey Goldstein offered a critique of the American Statistical Association's 2007 report 'Using Statistics Effectively in Mathematics Education Research' and a few education researchers talked about their experiences of using statistics in their work. In that meeting, the participants explored some of the ways in which the mathematics education research community might develop expertise and 'build capacity' in this area. One suggestion was the establishment of a working group at the termly conferences of the BSRLM. Thus, in March 2013, four seminar attendees (Evans, Monaghan, Noyes & Pope) initiated the Using Statistics in Mathematics Education Research working group.

The first session brought together a mixed group of around twenty participants. Andy Noyes presented some of his paper from the society's journal Research in Mathematics Education (RME 14(3)) which made use of survey data from the Geographies of Mathematical Attainment and Participation project. Andy also reflected on the process of moving from a predominantly qualitative approach to a more quantitative one. This generated some interesting discussion and led to a broader conversation about how the working group might develop in the future. With such a positive response we committed to reconvening the working group at subsequent BSRLM conferences.

The second meeting of the working group in June 2013 in Sheffield focused on experimental methods and drew on papers published in a special issue of RME (RME 15(2)) (It is hoped to report on this more fully in a future report). The most recent meeting of the group focused on the use of surveys (March 2014) in mathematics education research.

Reading the PIAAC Results: what to look out for - Jeff Evans

Since October 2013, results from PIAAC (Programme for the International Assessment of Adult Competencies) have been available (OECD, 2013a, 2013b). PIAAC aims to provide information as an international comparative survey, successor to IALS (during the 1990s) and ALL (2000s), and it has many similarities with national studies such as Skills for Life in the UK. Unlike the school level surveys (TIMSS, PISA), which gain access to "captive populations" in schools, PIAAC needs to use a combination of household survey and educational testing methodologies.

The first round covers a greater range of countries (24, two thirds of which are EU members, with the rest from North America, East Asia and Australia) – though all are advanced industrial economies. It focuses on three domains or "competencies" – Literacy, Numeracy, and Problem Solving in Technology Rich Environments (PSTRE). It uses computer administration, which in particular allows *adaptive routing* (allocation of items of "appropriate" difficulty to each respondent, based on performance in several "trial" items), and implements tighter specification and regulation of sampling and fieldwork standards than in previous international surveys (OECD, 2013b).

PIAAC's main objectives were presented by Andreas Schleicher (2008) of the Education Directorate at OECD – as helping the participating countries to:

> Identify and measure *differences* between individuals and across countries in key competencies
>
> Relate measures of skills based on these competencies to a range of economic and social outcomes relevant to participating countries, including *individual outcomes* such as labour market participation and earnings, or participation in further learning and education, and *aggregate outcomes* such as economic growth, or increasing social equity in the labour market
>
> Assess the performance of education and training systems, and clarify which policy measures might lead to enhancing competencies through the formal educational system – or in the work-place, through incentives addressed at the general population, etc. (Schleicher, 2008, pp. 2-3).

Numeracy is defined for the purposes of designing the items for PIAAC as:

> the ability to access, use, interpret, and communicate mathematical information and ideas, in order to engage in and manage the mathematical demands of a range of situations in adult life (OECD, 2013b, p.26).

This is put forward as a basis for conceptualising mathematical thinking in context. However, in order to *produce measures* of numeracy, the idea of *numerate behaviour* is developed, which "involves managing a situation or solving a problem in a real context by repsonding to mathematical information and content represented in various ways" (OECD, 2013b, p.26).

This led to specifying the following dimensions of "numerate behaviour" that can be used to guide the construction of assessment tasks:

- *context* (four types): everyday (or personal), work, society and community, further learning
- *response* (to mathematical task - three main types): identify / locate / access (information); act on / use; interpret / evaluate.
- *mathematical content* (four main types): quantity and number, dimension and shape, pattern and relationships, data and chance
- *representations* (of mathematical / statistical information): e.g. in text, tables, and / or graphs.

Each item can be categorised on these four dimensions, plus its estimated difficulty.

PIAAC also aims to produce affective and other contextual data that can be related to the respondent's performance. This includes demographic and attitudinal information in a Background Questionnaire (BQ), and self-report indicators on the respondent's use of, and need for, job-related skills at work; see OECD (2013b) for the BQ's conceptual framework.

In considering the validity of any social survey, especially in connection with the interpretation of results, I focus on three aspects of validity that relate to *relatively independent* aspects of the survey: *measurement*, *explanation*, and *sampling*. In the case of relating the PIAAC measures of numeracy to other characteristics of respondents (e.g the self-efficacy measures discussed by Pepper and Hodgen below), these three aspects lead to asking the following questions respectively:

A. Is the 'numeracy' (literacy, PSTRE) measure an appropriate indicator for the 'numeracy' referred to in research, policy and pedagogical debates? This can be called *Construct Validity*, and was treated in my talk as having several dimensions.

B. Many of the interesting findings are *correlations*, but have satisfactory *controls* been used to try to rule out alternative explanations? [*Internal Validity*]

C. All scores for a given country, and for any subgroups, are sample estimates. How do the sampling and estimation methods support the level of generalisation – to national levels, or to demographic groups - claimed in the report? [*External Validity*]. See Tsatsaroni & Evans (2014) and Evans (2014) for further discussion.

Using statistics in mathematics education: Confidence in PISA – David Pepper & Jeremy Hodgen

In this paper we outline an argument-based approach to the validation of assessments and argue that the OECD validation of PISA has not been sufficient for the proposed high stakes use of the assessment. We focus on the PISA assessment of 'characteristics and attitudes of students as learners in mathematics', specifically student confidence in mathematics. We argue that, by combining quantitative and qualitative sources of validity evidence, the OECD would be better-placed to identify threats to validity and potentially to avoid them in future surveys.

The PISA assessment of student competence in reading, mathematics and science receives much media coverage when results are published every four years (Luzon & Torres, 2011). The intended purpose of PISA is to serve '...the need of governments to draw policy lessons' through understanding the association between 'factors' and mathematical competence (OECD, 2004, p. 20). The survey therefore also assesses, using questionnaires, what PISA 2003 referred to as the characteristics and attitudes of students as learners in mathematics, or in PISA 2012 more recently as students' engagement, drive and self-beliefs in mathematics. Given the high stakes purpose of the survey, it is therefore important to validate not only the PISA tests but also the questionnaires for this particular high stakes purpose.

Here we focus on the PISA assessment of one type of self-belief, namely student *self-efficacy in mathematics*. There is much theoretical and empirical literature on self-efficacy (see Bandura, 1997), which refers to an individual's belief about whether they can successfully complete tasks in a particular area of performance, such as mathematics. Thus the PISA self-efficacy items asked students how confident they felt about eight mathematics tasks such as calculating the petrol consumption rate of a car. Whether a student believes they can successfully complete a mathematics task may affect not only how they approach that task but whether they attempt it at all.

There is, however, a lack of consensus about the nature of relations between mathematical self-efficacy, attainment and participation in mathematics, or how these relations might vary between groups or individuals. This is a consequence of the lack of longitudinal or experimental studies and, more fundamentally, the lack of validation for assessments of self-efficacy in mathematics (Pepper, 2014, forthcoming). The OECD's PISA survey does, however, offer the major international assessment of self-efficacy in mathematics.[2]

Assessment validity

With reference to educational and psychological measurement, Messick (1989) provided a widely-used definition of validity. In this definition, validity is not a property of the assessment but rather concerns its fitness for purpose:

> …an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.

Messick's conception of validity informed Kane's (2006) framework for an argument-based approach to validation of interpretations and uses of assessments. This approach requires, firstly, an interpretive argument specifying the interpretations and uses of an assessment and, secondly, a validity argument evaluating the assessment against these interpretations and uses. These arguments may comprise a chain of logical inferences. With reference to PISA, the most challenging would be an extrapolation inference from the assessment itself to the real-world situations emphasised in the OECD documentation for the survey. The arguments depend upon validity evidence, and the breadth of Messick's conception of validity means that it is important to seek out several strands of validity evidence for any validation. Indeed, the AERA/APA/NCME (1999) Standards for Educational and Psychological Testing identified five strands of validity evidence: instrument content; response processes; internal structure; relations to other variables; and, the consequences of the assessment. According to the Standards, validations should integrate evidence from each of the strands of validity evidence.

The OECD does not take an argument-based approach to validation and, in fact, does not have an explicit approach to the validation of PISA assessments. As concerns the interpretive argument, OECD (2004) does, however, indicate that the intended interpretation of the assessment of student self-efficacy in mathematics is across countries and the intended use is to inform education policies. As concerns, the validity argument, OECD (2004) also reports that the PISA validation of questionnaire items involved initial piloting of the questionnaire items in a limited number of countries and a field trial of the items in all participating countries. This resulted in the modification of questionnaire items but few details of the validation itself are reported. It is therefore unclear which strands of validity evidence were produced by the pilot and field trial. Furthermore, although the survey data are fitted to the Rasch model, the measurement properties of the PISA mathematical self-efficacy scale remain largely unreported. This raises the question: Is the PISA assessment of mathematical self-efficacy valid for its proposed interpretations / use?

In order to provide a more comprehensive validation of the PISA assessment of student self-efficacy in mathematics, Pepper (2014, forthcoming) conducted a

[2] While PISA assesses self-efficacy in mathematics and the similar but distinct construct of self-concept in mathematics, TIMSS assesses only self-concept in mathematics.

study drawing validity evidence from the PISA 2003 data set (for 41 countries, 10,274 schools and 276,165 students) and from my cognitive interviewing (with 41 students in 5 schools across England, Estonia, Hong Kong and the Netherlands). This cognitive interviewing involved concurrent verbalisation (asking students to 'think aloud' as they completed the PISA mathematical self-efficacy instrument) and retrospective prompts for clarification and elaboration. This represents a blend of techniques used in existing studies (Ericsson & Simon, 1993; Willis, 2005). These sources provided validity evidence of instrument content, response processes, internal structure, and relations to other variables. Since the validation focused on intended rather than actual uses, the consequences of the assessment were beyond the scope of the study.

Although the independent validation found that the items fitted the Rasch model and therefore represented a coherent scale, the effective range of items was narrow, patchy and low in comparison with the self-efficacy in mathematics latent in the population of students assessed in PISA 2003. In fact, the self-efficacy in mathematics of nearly 20% of students - almost the entire top quartile - exceeded the effective range of the items. This means that the assessment may not provide an accurate estimate of the construct, particularly for students with higher levels of self-efficacy. This is an issue for the PISA assessment of student self-efficacy in mathematics since the OECD emphasises overcoming difficulties. In addition to these issues across the sample, there was evidence of moderate to large differential item functioning (DIF) for six of the eight items in England, Estonia, the Netherlands or Hong Kong. This means that, given their responses to the items as a whole, some students' responses to individual items were unexpectedly high or low. To investigate whether this DIF reflected differences that are relevant or irrelevant to the construct, it was important to interpret the PISA 2003 data using validity evidence from the cognitive interviewing of student response processes for the items.

There was evidence of two major construct-irrelevant response processes. The first was a 'localising' response process. This involved students inferring varied task demands that were representative of local curricula or situations. This was problematic because local curricula are not necessarily representative of PISA's specified domain of mathematics in real-world situations, and local challenges are at odds with the very notion of a unitary PISA target domain. The localising response process appeared to be facilitated by the generic formulation of the tasks, which gave students scope to infer a wide range of demands. The second response process involved students responding with reference to their 'familiarity' with the task context rather than with reference to their confidence in inferred task demands. This response was therefore unrepresentative of the theorised self-efficacy trait. Students' lack of familiarity with some tasks, compounded by the absence of details inherent to generic tasks, made it difficult for them to infer any demands. These response processes generally weaken any extrapolation from the assessment to self-efficacy in real-world mathematics tasks, but they were particularly evident in Hong Kong and appeared to result in a substantial underestimate of Hong Kong students' self-efficacy in mathematics. The manifestations of the response processes across the items in England, Estonia and the Netherlands were more mixed, so it was unclear whether, overall, there was under- or over- estimation of the construct.

Given that there were threats to the validity of the PISA assessment of student self-efficacy in mathematics, the scope of the validity argument for the assessment is necessarily less ambitious than the interpretive argument. In fact, a validity argument for the proposed high stakes use of the assessment is untenable. A more transparent validation of the assessment by the OECD would better inform actual use of the

assessment. The actual use of the assessment should be limited to informing the development of the items for future assessments, including PISA 2021. Since this could include large-scale international assessments with the potential to inform understanding of relations between self-efficacy, participation and attainment, and their generalisability between education systems, this is nonetheless an important and potentially significant use for the present assessment. This independent validation indicates that the development of such assessments should be informed by various strands of validity evidence. The validation was conducted with PISA 2003 data then available but could, in due course, be replicated with the PISA 2012 data and, with more resources, cognitive interviewing in a larger number of countries.

References

AERA/APA/NCME (1999). *Standards for Educational and Psychological Testing*. Washington: American Psychological Association.

Bandura, A. (1997). *Self-efficacy: the exercise of self-control*. New York: W.H. Freeman and Company.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* Cambridge, MA, US: The MIT Press.

Evans, J. (2014). New PIAAC results: Care needed in reading reports of international surveys. *Adults Learning Mathematics International Journal, 9*(1), 37-52.

Evans, J., Wedege, T., & Yasukawa, K. (2013). Critical Perspectives on Adults' Mathematics Education. In M. A. Clements, A. Bishop, C. Keitel, J. Kilpatrick & F. Leung (Eds.), *Third International Handbook of Mathematics Education.* New York: Springer.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17-64). : Praeger Publishers.

Luzon, A. & Torres, M. (2011). Visualising PISA scientific literature versus PISA public image. In A. Luzon & M. Torres (Eds.), *PISA Under Examination* (pp. 269-302).

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan Publishing Co.

OECD. (2004). *Learning for tomorrow's world: first results from PISA 2003*. Paris: OECD.

OECD (2013a). *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills.* Paris: OECD. Online: http://www.oecd.org/site/piaac/#d.en.221854 [Accessed 22.5.14]

OECD (2013b). *The Survey of Adult Skills: Reader's Companion*. Paris: OECD. Online: http://www.oecd.org/site/piaac/publications.htm [Accessed 22.5.14]

Pepper, D. (2014, forthcoming). *Confidence in PISA: Validating an international assessment of student self-efficacy in mathematics.* King's College London.

Schleicher, A. (2008). PIAAC: A New Strategy for Assessing Adult Competencies. *International Review of Education, 54*(5-6), 627-650. Online: http://www.oecd.org/dataoecd/48/5/41529787.pdf [Accessed 22.5.14]

Tsatsaroni, A. & Evans, J. (2014). Adult Numeracy and the Totally Pedagogised Society: PIAAC and other international surveys in the context of global educational policy. *Educational Studies in Mathematics* (in press).

Willis, G. (2005). *Cognitive interviewing: a tool for improving questionnaire design*: Sage Publications.