

Part 1: Lessons learnt from international assessments

By Dave Tout

In this first of a two-part article, Dave Tout takes us behind the scenes of international assessment, and provides commentary on the implications and value of such assessment, for the wider community and for the LLN field. In the December edition of *Fine Print* we will publish Part 2 of the article: what international assessment can teach us about learning and teaching.

Introduction

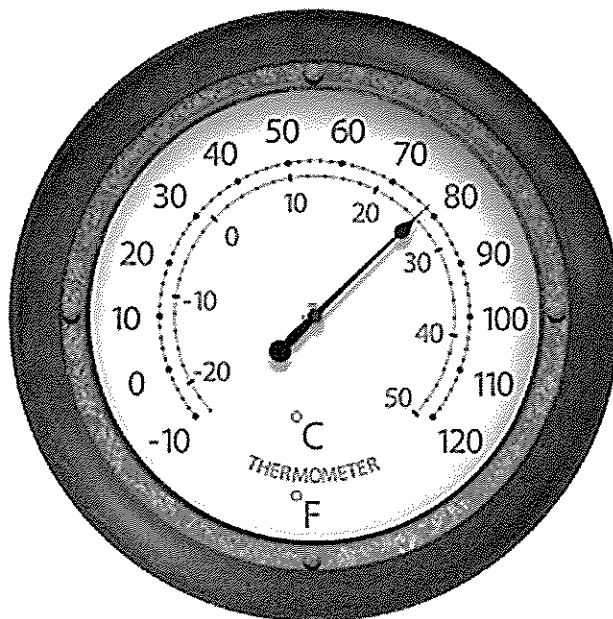
In Australia, 1989 saw the first national survey of adult literacy standards, *No Single Measure*, by Rosie Wickert; this was followed by the International Adult Literacy Survey (IALS) in 1996, the Adult Literacy and Life Skills (ALLS) survey in 2006 and now the Program for the International Assessment of Adult Competencies (PIAAC) undertaken in Australia in 2011. We are now in the situation where the preliminary results from PIAAC have been released by the Australian Bureau of Statistics (ABS, 2013). This article could just be about those results and the percentages at each level and look at whether Australia's performance has improved or not over the last few decades. But that's not what this article is about.

Most of my career has been in the adult language, literacy and numeracy (LLN) sector in Australia. I have also worked on some of the above international literacy and numeracy assessments: first as a member of the numeracy expert group for ALLS and then on the follow up study, PIAAC. Since 2009 I have also worked on and helped manage the mathematical literacy test development for the international Programme for International Student Assessment (PISA) of fifteen year-olds from sixty-seven countries across the world.

In this, and in a second article in the next edition of *Fine Print*, I will attempt to explain what I have learnt from my work on these international assessments, and why I believe there is value and crucial and important lessons for all educators, and others, from the work underpinning these empirically based research endeavours. While some criticism of the surveys is understandable and to be expected, this often blinds readers and the LLN sector from reading between and behind the lines and seeing what value can be generated, gained and learned from such investigations, data and research.

Complexities of international assessments

When I first began working on international assessments, I thought I knew a bit about developing, writing, using,



What is the temperature shown on the thermometer in degrees Celsius (°C)?

and interpreting assessments for adult numeracy learners. I soon realised how ignorant I was about the complex process and sophistication of international assessments. This not only related to the issue of writing assessments, but also about the methodological approaches used.

My ignorance led me to believe that when I was first approached to contribute back in 1998, I would most likely be asked to sit here in Australia, find or invent some appropriate stimuli and write some good questions covering different numeracy and maths skills, submit them via email and they would be fixed up (a bit) and then used in the international assessments. I soon learnt that this was not the case.

First there was the realisation that you cannot write the items without an understanding of the statistical theory sitting behind the assessments, second that you cannot proceed without an agreed theoretical framework and structure to describe what you are assessing, and thirdly,

that you needed a theoretical construct that will enable you to predict, in advance of an assessment actually taking place, how difficult each item is, *and* that this needs to be validated empirically. My journey and lessons began, and I am still learning.

Statistical theory behind assessments

Item response theory (IRT) is the statistical methodology that sits behind international comparative surveys such as ALLS and PIAAC. The first surprise to me was that the IRT approach turned assessment on its head—it was as much, if not more, about the performance of the items being tested and measured rather than the test-takers—test-takers are not given the traditional feedback with a score of x questions correct out of a total of y and if your x (or percentage correct) is higher than mine then you are better than me. Item response theory (IRT) allows a large number of items of varying difficulty to be developed to assess a wide range of skills across a domain (e.g. reading or numeracy) and test-takers answer different sets of questions.

Based on the performance of a large number of individuals, the test items are placed on a scale of difficulty relative to each other, independent of the ability of students taking the test. Once item difficulties are established, test-takers can be placed on a scale of ability relative to each other—the items and test-takers are placed together on the same scale. This then allows the construction of described scales and the ability of a population can be estimated more accurately against different levels or bands on the scale—what proportions of the population can respond correctly to the type and difficulty of items within that band.

The IRT methodology allows scores across different assessments to be compared via the inclusion of *anchor* or linking items. These common link items enable comparisons to be made across different tests; across different populations (e.g. different countries); and across time (e.g. different cycles—ALLS and PIAAC).

IRT is generally acknowledged as an improvement over classical test theory as it brings greater flexibility, provides more sophisticated information and provides improved validity and reliability of an assessment compared to traditional methods.

Test development process

My second lesson was to learn about the extensive process that sits behind international assessments, which help

to guarantee the quality of the stimulus materials and assessment items that make up any assessment developed.

Conceptual framework

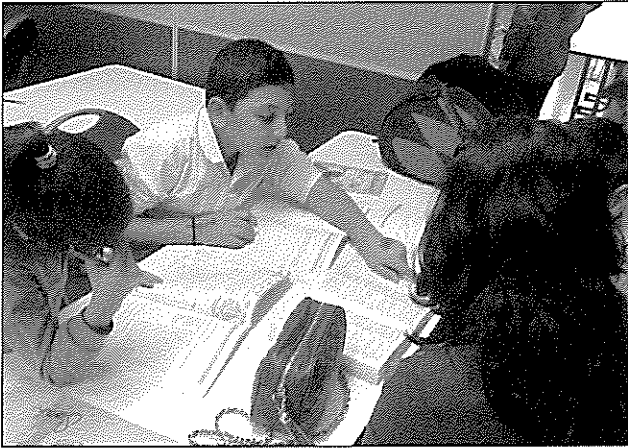
First, test development and writing does not proceed without an agreed conceptual framework which includes a description of what is being assessed and why and how. These frameworks are developed by teams of international experts from different countries with different cultural and language backgrounds. Extensive consultations take place to ensure that the assessment framework is relevant to the targeted cohort. A framework includes a key test construct which describes characteristics of the items, their type and style, their content and so on. Only once the framework is agreed, does writing of the items commence.

Quality assurance processes

Secondly, the development of the assessment items in these international assessments follows a comprehensive process to ensure the quality and validity of the items being developed. The assessments use a team approach to stimulus and question development, whereby a team of experienced test developers write the items and meet together as a panel to critique and review each other's items, and as such the items are revised and improved. Following that stage, the revised items go through a further comprehensive stage of reviews and revisions.

Next, in the item development process there is a review stage, which involves focus groups and pilots with potential test-takers to get their feedback and opinions on the draft items and the delivery mechanisms. This picks up issues related to interest and relevance, alongside issues related to the meaning and understanding or misunderstanding of the items. In international surveys this stage also includes feedback from participating countries and through a formal translation and review process with language experts. These processes pick up issues to do with the language structure and meaning of items, but also content and cultural issues.

The next crucial stage is trialling. The assessment items go through a final review and revision, design and final desktop publishing and are prepared for trialling in test booklets or for computer-based delivery. Trialling is done with a sample of the target population in each participating country before the final test. The trial data is collected and analysed psychometrically and from these detailed analyses, *misbehaving* items are rejected on a number of levels including for reliability, fairness and validity. Then for the remaining successful items, any fine-tuning is



undertaken, and a representative set of items are chosen and placed into final forms for each domain being assessed.

This final selection needs to meet the criteria established in the framework and the key test construct, and the range of quality assurance processes followed help to guarantee that the test as a whole is a valid, reliable and fair assessment of the constructs described in the assessment frameworks.

What did ALLS and PIAAC assess?

As described above, sitting behind international surveys are theoretical frameworks. In ALLS and PIAAC, the frameworks attempt to describe what literacy and numeracy in the 21st century incorporates, and how this might be assessed in an international assessment. In ALLS (ABS, 2007), there were assessments that were to provide information on knowledge and skills in prose literacy, document literacy, numeracy and problem solving. There was also a health literacy scale reported.

In PIAAC (Jones et al, 2009; Gal et al, 2009), literacy was described as understanding, evaluating, using, and engaging with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential. Numeracy was described 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. This was elaborated to describe more detailed components of numeracy. More details will be included in the second of my two articles.

The items themselves

The tasks are, as much as is possible in a large-scale international testing situation, based on adult contexts and *real-life* scenarios and texts. Even though these are

large-scale international assessments, they are not simply very basic, school like, literacy and numeracy multiple choice type questions devoid of any relevant context or purpose. The assessment requires test-takers to respond in different ways to a range of accessible and not uncommon or unusual situations and documents. They are assessments of the types of interactions adults may need to undertake in the 21st century on a daily basis whether at work, at home, in the community or related to study or training.

Test developers research and find real texts and situations that the target group—in this context adult—may encounter and these form the initial impetus for the writing of the assessment items. The items may then be simulated and often simplified as the real world texts may be too complicated to be accessible to the majority of adults. Texts include artefacts such as advertisements, webpages, labels, newspaper articles, instructions and signs, maps, diagrams and plans, photos, etc.

The assessments are undertaken in people's homes, and there is no time limit and test-takers are urged to try each question. To support this, tasks are not ordered in the test forms in terms of increasing difficulty and test-takers who cannot answer a particular question may be capable of answering the following questions.

In both ALLS and PIAAC, a ruler and calculator were provided to test-takers for use in the numeracy items (or they could use their own if they preferred). Yes, adults were asked to measure the dimensions of a given item.

Feedback on the items

As mentioned above, part of the quality assurance process for these surveys involves focus groups with potential test-takers to gain their feedback and opinions on the draft items and the delivery mechanisms. Feedback and information is also obtained from the participants in pilots, which provide information about how adults respond to the test items and the test situation. Throughout all these processes, and in feedback from people who administer such assessments to participating adults, invariably the feedback from the test-takers has been positive.

Background questionnaire and screening questions

In both ALLS and PIAAC, test-takers are first asked a series of about three hundred background questions that are then used in the research and analysis based on the results. The test-takers are then presented with a screening

test containing a small number of simple, low level tasks. If the test-taker fails to complete more than half the items correctly, the interview is concluded. This is to identify and take into account test-takers who, for example, have very low levels of English language and/or literacy ability.

Limitations

Both the ALLS and PIAAC surveys are designed to provide an empirical *snapshot* of the performance and abilities of the adult population in relation to a test of literacy (reading in this instance), numeracy and the other skills being assessed. These international surveys do have limitations. Survey assessment items can only imitate real life literacy and numeracy tasks and cannot be genuinely socially or contextually situated. Dependence on reading and on information processing via reading rather than allowing for oral transactions and responses, or for accessing support, places limitations on the interpretation of the results. There is no assessment of writing skills per se and no writing scale has been developed. As such, ALLS and PIAAC as described in their frameworks and reports, are surveys about *aspects* of literacy and numeracy, not the whole range of ways literacy and numeracy are part of today's modern society.

The reality with the numeracy (and similarly with the literacy) tasks is that the stimuli are shorter and simplified versions of real texts (documents or webpages) adults meet in their life, such as when browsing the internet, when reading and interpreting information at the doctors or infant welfare centre, at Centrelink, in OH&S manuals or standard operating procedures at work, or in an education or training course. As such it is reasonable to expect that a broad set of questions based on such simplified stimuli given to a representative sample of the adult population would provide valid and valuable insights and information about the general literacy and numeracy abilities and skills of the Australian population, especially when it is known the exhaustive quality assurance process the questions and stimuli go through prior to their use.

But what are the results saying?

At the start I said this article would not focus on the results of ALLS and PIAAC and the percentages or estimated numbers of adults performing at each level, as those results are available in the ABS reports and have been and will be documented in a range of other papers. But I am in the privileged position of knowing what every numeracy item was in both ALLS and PIAAC, and all I know is that as a long-time LLN educator, those results, no matter how you



read them, demonstrate unequivocally that a significant number of people aged from fifteen to seventy-four years old in Australia do not have access to sufficient numeracy and maths skills to be able to cope equitably with life in the 21st century. The data and results are *not* about rates of illiteracy or innumeracy, and don't claim to be.

I believe that the evidence about this is real and valid. This is independent of the issue as to which level of the ALLS or PIAAC scale is used as a baseline or minimum measure or standard. What actual LLN skill levels people need for what job, for which qualification, for which life purpose is arguable and debatable (and is a good argument to have) but, on my reading, there *is* an issue with the lack of literacy and numeracy knowledge and skills of far too many Australians. And my recent work on PISA and the abilities of fifteen year-olds is, I believe, consistent with the messages from the adult surveys. The ability to be empowered to have the capacity to make considered decisions, whether they be on the spot decisions at a workplace or when out shopping, or following written instructions about a medical or health matter, or making decisions about financial matters, or understanding the implications of gambling, require good foundational LLN skills. The results of these surveys show that millions of Australian teenagers and adults do not have such foundational LLN skills and they are, potentially, disempowered.

One impact of surveys—support for LLN

As an adult numeracy educator working in the adult LLN field for many decades, I have always been passionate about the crucial, and often under-recognised, role that LLN skills play in empowering and enabling adults and young people to continue to learn and develop, to participate fully in society and to be productive in their work. We, as practitioners, are aware of how low LLN skills can be a barrier and can disenfranchise people from accessing and enjoying the many benefits of the world as we know it—my ability to understand, use and critically apply my mathematical (and literacy) skills gives me, and my family, a range of choices that provide various pathways and options for success in life. And we believe all people should have that access and choice, and this is the reason why many of us work in the sector.

However, in most of the eras I have worked, there has been a frustration that the adult LLN field was marginalised, under-resourced and under-supported, with a highly casualised and underqualified workforce. I could not understand why everyone else—other teachers and trainers, government, industry and employers, unions—weren't as equally passionate about the vital role that LLN skills play in empowering and enabling adults to succeed. Through my work on the international surveys I have worked with and shared conversations about the surveys and their results with a wide range of personnel including researchers, bureaucrats, economists and statisticians, journalists, and more. It is the results and research from the international surveys that make them sit up and notice about adult literacy and numeracy.

There is now extensive evidence of the value of having higher literacy and numeracy skills—our society values these skills and rewards them. This extensive research and conclusions are based on the data from the background questionnaire and the skill performance in the surveys. A good example of related research was that a health literacy scale was created as a by-product of ALLS. Based on such international data considerable research related to health literacy has been undertaken overseas:

The international research on *health literacy* is considerable. Studies have found links between lower literacy and a higher risk of hospitalisation, higher rates of depression and an inability to understand and comply with the use of prescription drugs. (Hartley & Horne, 2006)

There has also been considerable research about the impact of low literacy and numeracy skills on the economy and productivity and on a country's GDP. No matter the stories about the unique (and extremely rare) successful individuals who say they have made it without being able to read, write or do arithmetic, for the vast majority of individuals low levels of literacy and numeracy skills will have a negative impact on both their social and economic future. It is therefore of value to support and enhance an individual's literacy and numeracy skills—from all viewpoints—individual, societal or economic. Evidence of this has resulted directly from research from the adult literacy and numeracy international assessments, and which supports our own personal experiences and anecdotal evidence from our teaching as LLN practitioners.

It is quite clear that the results and research resulting from the ALLS survey (ABS, 2007) has seen the beginning of a set of significant responses and investments in adult LLN. Both government and industry have begun to argue for and acknowledge that the core skills of LLN are important and need to be addressed and supported on a national basis. In the last few years a considerable number of reports and initiatives have been published and introduced which contribute to meeting these needs. These include the Industry Skills Councils report, *No more excuses—an industry response to the language, literacy and numeracy challenge*; the research by the Australian Industry Group, the *National workforce literacy project: report on employers' views on workplace literacy and numeracy skills*; the Language, Literacy and Numeracy Practitioner Scholarships Programme; the National Foundation Skills Strategy for Adults released by the federal government; the development of new higher level LLN qualifications for VET trainers and the development of a new Foundation Skills Training Package.

There's lots that needs to be done, especially in relation to supporting and up-skilling the LLN skills and knowledge of the VET and LLN workforce. But at least the above is evidence that the adult LLN sector may look forward to getting some of the recognition and support it deserves and needs. And this has arisen out of the ALLS results and related research.

Teaching and learning lessons

There are other lessons that can be learnt from the work and research resulting from these international assess-

ments, and these have impacts in terms of teaching and learning of LLN skills. For example, their conceptual frameworks often include a meta-analysis of existing research and related LLN issues and as such can be a useful source of relevant international research and theory. There are a number of clear messages for teaching, that result from the empirical data, about how and what adults can and cannot do. The expert groups, in conjunction with related research, have identified factors affecting task, text and item complexity and difficulty. Further description of these aspects and benefits of the assessments will follow in the second of my articles in the next edition of *Fine Print* journal.

Conclusion

Combined, I believe that the results, the research and the underpinning conceptual frameworks for international assessments such as ALLS, PIAAC and PISA add to the expertise and knowledge of not only the education and research communities but also to the awareness, understanding and valuing of LLN skills by government, industry and the workforce. As well, the empirically based research emanating from such assessments, alongside associated theoretical works such as around text and task complexity, can and do contribute to understanding about the teaching and learning of literacy and numeracy.

There are particular issues relating to numeracy. In *No Single Measure* back in 1990, Rosie Wickert stated:

... it is clear from the results that when people have poor literacy skills, they have even worse numeracy skills. The need to upgrade numeracy skills in the context of literacy must be taken into account of in all decisions to raise the level of adult literacy in Australia. (Wickert & Kevin, 1995. p. x)

I don't believe the situation has improved significantly since then in relation to numeracy in LLN, and numeracy has been, and still is, the poor (and lost) cousin in LLN; this is despite research that indicates that numeracy may have a stronger impact than literacy. The ALLS and PIAAC data also indicate to me that it is possible that the bar is set lower for numeracy. We should not lower our standards or expectations: we need to counter the community and cultural attitude that it's OK to not be good at maths.

And to conclude, I'd like to again say that the data and results from ALLS, PIAAC and PISA should be taken seri-

ously. My involvement in this international work has made me even more committed and passionate about the need to empower *all* learners, no matter their age and background, to have access to the wide world of numeracy and mathematics (and literacy). ALLS, PIAAC and PISA have been instrumental in providing the evidence to argue to invest in and improve LLN skills across the whole population.

Dave is an experienced numeracy educator who is particularly interested in making mathematics relevant, interesting and fun for all students especially those students who are disengaged from mathematics. He has worked in a range of programmes in schools, TAFEs, ACE providers, universities, AMES and industry. Over the last fifteen years he has also been involved in the development and writing of the numeracy components of the international Adult Literacy and Lifeskills Survey (ALLS) and the Programme in Assessment of Adult Competencies (PIAAC). Dave joined the Australian Council for Educational Research (ACER) in 2008 where he is now the manager of Vocational, Adult and Workplace Education Services.

References

- Australian Bureau of Statistics. (2007). *Adult literacy and life skills survey: summary results Australia*. (cat. no. 4228.0), Australian Bureau of Statistics, Canberra.
- Australian Bureau of Statistics. (2013). *Programme for the international assessment of adult competencies, Australia, 2011–2012* (cat. no. 4228.0). Australian Bureau of Statistics, Canberra.
- Australian Industry Group & Department of Education, Employment and Workplace Relations. (2012). *When words fail, national workforce literacy project final report*. Australian Industry Group, North Sydney.
- Gal, I., Alatorre, S., Close, S., Evans, J., Johansen, L., Maguire, T., Manly, M. & Tout, D. (2009). *PIAAC Numeracy: a conceptual framework*. OECD Education Working Papers, No. 35, OECD Publishing. Retrieved from <http://dx.doi.org/10.1787/220337421165>.
- Hartley, Robyn & Horne, Jackie. (2006). *Social and economic benefits of improved adult literacy: towards a better understanding*. National Centre for Vocational Education Research (NCVER), Adelaide.
- Industry Skills Councils (Australia). (2011). *No more excuses: an industry response to the language, literacy and numeracy challenge*. Industry Skills Councils, Australia.
- Jones, S., Gabrielsen, E., Hagston, J., Linnakylä, P., Megherbi, H., Sabatini, J., Deutsches, Tröster, M. & Continued on page 39 ...