Postscript

The following is short summary of the webinar in the "Firechat" series presented on 3 June 2014. The video is now available online at dotr.im/webinar030614. Please note that this is written as "notes" and not intended to be a peer review journal article. However, all references cited during the talk are given here.

Validity of test norms – Should there be an expiry date?

By Dr Ian Smythe

We take for granted many of the tests that are recommended to us. But how often is the content and validity questioned? Of course we all ask about those diverse types of validity (or subsets of validity as they are now called) of a test instrument including face, content and predictive validity (Wikipedia Validity, 2014). But how can the acceptance change over time?

There are many reasons why validity may change. These include (though not limited to:

a) Validity of a fad

Fads are temporary fashions, and there are many examples on the periphery of evidence-based psychology. For example, though arguably more enduring than most fads, Grant (2013) highlights the lack of validity in the Myers-Briggs Type Indicator. First published in 1962, it is apparently now used by 89 of the Fortune 100 companies. Yet 50% of those who try it will get a different outcome when they redo the task two weeks later. So how valid can this "test" be? Any use in the field of dyslexia should be treated with extreme caution.

b) Consensus of opinions change

There are a number of attempts to fit dyslexic individuals into categories that are based on clusters of results. For example, the ACID profile (a spiky WISC profiler with deficits in Arithmetic, Coding, Information and Digit Span) was widely used in the 80's and 90's, but has now fallen into disrepute. Why? Because people began to notice that there are many dyslexic individuals that do not conform to this profile, and quality research failed to replicate the predictability of earlier work. Thus opinion changed due to a lack of evidence of the validity of the ACID profile for dyslexia assessment.

Probably 95% of assessors would use reverse digit span as an indicator of working memory. However, since there is no clear definition of working memory, and it may be argued that the concept itself is still evolving, the choice of tests can be questioned. Few question the validity of the reverse digit span test, but does it measure much more than the ability to recount numbers in a different order? Most of the more recent proposed tests or working memory are equally valid. But what do they really measure? (See for example Kane et al 2007.)

c) Improved understanding

Ask most assessors what is executive functioning and they will give you a good response. Ask them how to measure it with valid instruments, and the answer becomes less clear. That is because the exact nature of it "still evolving." Indeed more than 50% of all research papers on the subject of a recent, apparently widespread, meta-analysis were from the years 2010-2012. Whilst it would appear to be an important concept in the field of dyslexia, until we have a valid theoretical model (and consensus) of executive function, we will not be able to have a valid measure.

Postscript

The following is short summary of the webinar in the "Firechat" series presented on 3 June 2014. The video is now available online at dotr.im/webinar030614. Please note that this is written as "notes" and not intended to be a peer review journal article. However, all references cited during the talk are given here.

d) The Flynn Effect – change over time

The Flynn effect was originally noted as a change in IQ over time (Flynn, 1987 – See Wikipedia Flynn Effect (2014) for details.) Since the original work, this change has been noted in many tests, including WISC, WAIS and Raven's matrices. It is this effect that will be the subject of the rest of this article.

Time to test validity

Since the original reports, many researchers have replicated change, but question the cause. Some suggest that the effect is due to a change in the construct of the tests, whilst others maintain it is a function of a changing society. Others have tried to look deeper into the subcomponents. For example Sailer (2013) provided details (adapted from Flynn, 2007) highlighting that the relative gain in IQ points in the period 1947-2002 was as low as 2 for Information and arithmetic, but was 22 for Picture Arrangement and 24 for Similarities. That is half a point per year. So if you are using the 1994 version of WISC….. you do the maths and work out the implications if you are using this test as a diagnostic tool. (See also Kanaya and Ceci 2010, and Wicherts et al 2004.)

However, this effect of change of the validity of the measures is not restricted to "intelligence tests" (N.B. To be clear, in my opinion these are of limited validity in the assessment of dyslexia, though some sub-tests do have some merit if interpreted appropriately.) Take for example reading tests.

The Test of Word Reading Efficiency (TOWRE) is a measure of word-reading efficiency. However, as McKenna et al (2009) pointed out, a sight word list will only measure the sight words included. And if the individual has a huge sight word vocabulary, but not the words on that list, then the validity of the test becomes questionably. Irrespective of the specific test, the reason for doubting the validity of a test more than 17 years after norming could be that language has evolved (this was well before Facebook and the iPhone) and if used in the UK, one must remember that it is US based.

Comparison of two UK normed spelling tests also show considerable variation over time. The Schonell Spelling test was normed in 1932, and includes the word portmanteau, a word that is not exactly in daily use today. (It was probably included for its reference to two compartmented luggage rather than two morpheme words.) As Turner (1998, p205) highlighted, comparison of the Schonell (1932) with the Vernon (1977) spelling test showed that the tests "come adrift by a year at age 11 and 2.5 years at nearly 16." If you consider that there is 45 years between the two, and 37 years since the Vernon was normed, the validity of the Vernon norms may also be equally in doubt. Provocatively, if this was representative of the rate of change of a reading test, then it equates to almost a month per year at age 16. (Actually 30 months difference over 37 years, or 0.81 months per year.) So even a test produced in the year

Postscript

The following is short summary of the webinar in the "Firechat" series presented on 3 June 2014. The video is now available online at dotr.im/webinar030614. Please note that this is written as "notes" and not intended to be a peer review journal article. However, all references cited during the talk are given here.

2000 could be as much as 11 months out by now, a significant factor if you are looking for a discrepancy of two years in reading age. Put another way, it is possible that a person may be below a threshold on an older test, but above it on a more recently normed test.

In an attempt to highlight the change of language, a short analysis was carried out using 19 of the words from the Schonell reading test. The rank of the items was compared with their ranking in the British National Corpus (Wiktionary British National Corpus, 2014) which is based on books, and a contemporary word list based on tv and movie scripts (Wikipedia Word Frequency, 2014). Correlations were as follows:

Correlation between Schonell and Wiktionary List (compiled 2006) 0.68

Correlation between Schonell and British National Corpus (compiled 1993) 0.52

Correlation British National Corpus and Wiktionary List 0.64

Of course one can find plenty of reasons why the correlations are not so high, including that the Corpus is based on the written word whilst the Wiktionary is based on the spoken word. However, it may also be a reflection of the evolution of language. Whilst it is not intended as a piece of "research evidence" it does highlight the need to exert caution when reviewing the validity of tests. In particular, it may be argued that language diversity today is based more on television than reading books than it was even 20 years ago.

Finally, one should consider that each test may be affect over time by different factors. So whilst the Flynn Effect in "intelligence" tests may be due to a combination of "practice" effects and a change in face validity, improvements in tests such as Raven's Matrices have been shown to be strongly correlated to classroom pedagogy. Where the culture (classroom and home) encourages greater self-exploration and self-expression, the improvement in scores in reasoning tests have been greater than in educational environments where the teaching is more prescriptive.

And as for spelling, could there have been a fall in spelling standards that could be attributed to a greater reliance on the spell checker? That is, let's say that for a spelling test normed 20 years ago the average for 15 year olds was 30/40. But if you were to use the same spelling test, the average today may have dropped to 27/40. Should the test be renormed? Put another ways, as Galletta et al suggested in their article "Does Spell-Checking Software Need a Warning Label?"

Dynamic norming

Whilst it would seem impractical to overcome this problem, online testing offers the potential for a dynamic norming process. That is, as the data is collected, so the norm is developed. With the potential to time stamp each response, it is even possible to use a

Postscript

The following is short summary of the webinar in the "Firechat" series presented on 3 June 2014. The video is now available online at dotr.im/webinar030614. Please note that this is written as "notes" and not intended to be a peer review journal article. However, all references cited during the talk are given here.

process of rolling norms, whereby result gathered even a year ago are removed. This is successfully implemented in the Do-IT Profiler, though a manual override is employed to avoid practical issues such as short term demographic bias and outliers. An additional advantage is that one can also create the option of regional norms as a subset within the overall data capture.

As an example, in South Africa a spelling test is used with Johannesburg Further Education and Training colleges, deployed using the Do-IT Profiler. At the time of writing, over 15,000 students have been tested out of a population of 50,000, or 30% of the entire population. This is considerably more than most test norm protocols. However, there are considerable variations in socio-economics status, and in some instances it would be inappropriate to compare one area with another. Use of dynamic online norming makes that feasible. Thus one could compare classroom averages within a suburb, to help identify where additional support is needed at the local level. Furthermore, as local, Province and national interventions are implemented, one would expect, say, literacy levels to improve. In that case it would be illogical and inappropriate to maintain an out of date norm. The system implemented in South Africa allows removal of older data, thus ensuring results reflect the current state of health of the local education system.

Conclusion

By using such dynamic, computer assisted methods, validity concerns due to the time lag between the norming process and use of the test is removed. Thus no "Use by" date is required, since old stock is automatically removed and norms are always fresh.

Conflict of interest: The author acknowledges a conflict of interest as he is co-owner of the company that delivers this type of service, not only in South Africa but also UK and Ireland.

References

Flynn J. R. (1987). "Massive IQ gains in 14 nations: What IQ tests really measure". Psychological Bulletin 101: 171–191. Web access: http://www.jugendsozialarbeit.de/media/raw/flynn1987_What_IQ_tests_really_measure.pdf

Grant A (2014) http://www.huffingtonpost.com/adam-grant/goodbye-to-mbti-the-fad-t_b_3947014.html

Galletta DF, Durcikova A, Everard A, and Jones BM (2005) Does Spell-Checking Software Need a Warning Label?

http://isites.harvard.edu/fs/docs/icb.topic761456.files/p82-galletta%20durcikova%20everard%20jones%20cacm.pdf

Postscript

The following is short summary of the webinar in the "Firechat" series presented on 3 June 2014. The video is now available online at dotr.im/webinar030614. Please note that this is written as "notes" and not intended to be a peer review journal article. However, all references cited during the talk are given here.

Kanaya T and Ceci SJ (2010) The Flyne Efect in the WISC subtests among school children tested for special education services. Journal of Psychoeducational Assessment. https://lesacreduprintemps19.files.wordpress.com/2012/05/the-flynn-effect-in1.pdf

Kane MJ, Conway ARA, Miura TK and Colflesh GJH (2007) Working Memory, Attention Control, and the N-Back Task: A Question of Construct Validity. Journal of Experimental Psychology: Learning, Memory, and Cognition. Vol. 33, p615– 622. Web access: https://www.princeton.edu/~aconway/pdf/Kane2007nback.pdf

McKenna MC, and Dougherty Stahl KA (2009) Assessment for Reading Instruction, Second Edition. Guilford Press. New York

Mencap (2012) Mencap survey highlights Britain's poor spelling http://www.mencap.org.uk/news/article/mencap-survey-highlights-britains-poor-spelling

Sailer S (2013) Has a 15-year-old explained the Flynn Effect? http://isteve.blogspot.com/2013/11/has-15-year-old-explained-flynn-effect.html

Turner M (1997) Psychological Assessment of Dyslexia, Whurr, London

Wicherts JM, Dolan CV, Hessen DJ, Oosterveld P, van Baal GCM, Boomsma DI and Span MM (2004) Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. Intelligence 32. P509-537. http://wicherts.socsci.uva.nl/wicherts2004.pdf

Wikipedia British National Corpus: http://en.wikipedia.org/wiki/British_National_Corpus

Wikipedia Flynn Effect (2014) http://en.wikipedia.org/wiki/Flynn_effect

Wikipedia Validity (2014) http://en.wikipedia.org/wiki/Category:Validity_(statistics)

Wiktionary Word Frequency (2014) http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists