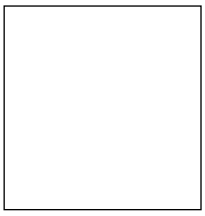
So wrong for so long

Changing our approach to change



JOERG A. PRIELER *on the importance of item response theory.*

I COMPLETED my PhD while serving as an officer in the Austrian army. Finding myself in those military circumstances was uncomfortable, but it turned out to be serendipitous. I needed a research topic related to selection testing, and the army needed to validate a selection procedure to predict the likelihood of officer candidates successfully completing the rigorous training (Prieler, 2000). My supervisor was Gerhard Fischer, a pioneer in the area of applying item response theory (IRT) methodology to the evaluation of change as measured by psychometric tests.

The selection procedure involved the evaluation of change in performance on a battery of reasoning, memory, concentration and personality tests across two conditions: in a rested state and 'under load' (after a stressful 12-hour march carrying a heavy backpack). This was a logical and well-thought-out selection design, since a key factor influencing the likelihood of candidates completing officer training is the extent to which they can maintain level of performance under stress. Little did I know that what I was to learn from this study would become the greatest source of frustration for me in my whole career. It showed me the flaws in how we, as a profession, approach the evaluation of change as measured by tests before and after intervention.

This is important because the measurement of change is central to psychology in all its applications. Psychologists must assess the efficacy of their interventions, whether with groups or at an individual level. At the group level we need to analyse accurately whether people who differ with respect to particular characteristics are more or less helped by particular interventions. At an individual level we must show that an intervention was effective for a particular person. If we get it wrong the implications can be great. Imagine the consequences of drawing erroneous conclusions from drug trials, or being wrong about the effectiveness of interventions relating to the education of children or development initiatives at work.

Fortunately I had my supervisor to warn me against using classical methods to evaluate the differences in test performance across the two conditions, but even a cursory review of the literature supports his position (e.g. Bereiter, 1963; Lord, 1963).

'IRT is less well understood and more complex to apply than traditional approaches'

It has been recognised for years that analytic methods based on classical test theory (CTT) are likely to undermine the prospect of arriving at any meaningful and accurate conclusions about the real effect of interventions (Williams & Zimmermann, 1996). Yet as a profession we still, more often than not, employ this flawed methodology.

Methodology now exists to evaluate the nature of gains and losses after intervention more effectively. In my study, the use of IRT allowed me to identify specific items within the test battery that were predictive of success in training. I was able to eliminate those items that showed no predictive validity, shortening the assessment procedure and increasing the efficiency of the process. The Austrian army adopted my approach. Had I used traditional methods (based on CTT) I would have rejected the battery out of hand for its low predictive validity, and would have lost important information about how aspects of the procedure showed real predictive power.

The IRT methodology I used is tried

WEBLINKS

Item response theory: *www.edres.org/irt*

'A visual guide to item response theory': *www.metheval.uni-jena.de/irt/VisualIRT.pdf*

Measuring change is central to psychology – from officer selection to taking the high jump

TABLE 1 Key advantages of IRT over CTT for the analysis of change

CTT	IRT
Relationship of score to ability level is based on overall score across items	Direct relationship is established between ability level and parameters of individual items (such as difficulty of item and discriminative power at different points in the distribution)
Factors emerging are seen as 'primary' influences on test performance with individual items being variously affected by other factors	Factors emerging are less contaminated by secondary factors because attention has been given to homogeneity of items
'Bad' items reduce predictive power	'Bad' items are eliminated
Level of ability is defined in relation to a particular sample	Level of ability can be defined independently of any sample
Correlation is used to compare performance on two test occasions and this further clouds the analysis	No need to use correlation so disadvantages are removed
It is not possible to measure the significance of change at the individual level	The significance of change at the individual level can be objectively measured using IRT tests

and tested, and yet we as a profession still are not embracing it. The problem seems to be that IRT is less well understood and more complex to apply than traditional approaches. But this is not a good reason to avoid embracing it. To do so is analogous to a medical doctor's diagnosis of cancer being dependent on whether the scanner uses the methodology designed by company X or company Y. No medical doctor would ever say 'Yes I know I'm using an inferior method to predict whether you will die or not, but this method is so much easier for me'.

Here I make an attempt to persuade you of the importance of changing your approach to one of psychology's central practices: the way the efficacy of interventions is measured. I use the example of ability tests because the relative merits of CTT and IRT are easier to illustrate in this domain; but the points made hold true also for measuring behavioural change over time.

Key advantages

The key advantages of IRT methodology over CTT analytical methods are shown in Table 1. The problematic issues in traditional CTT approaches arise from two key factors.

Firstly, the analysis of change from baseline to post-intervention is based on scores that are derived by summarising performance on the test overall. The scores on the test items are added up to give an overall 'raw score'. As a result the information we get is necessarily compressed; we cannot distinguish between good and bad items in terms of predictive power, so the 'bad' items cloud the predictive power of the 'good' items.

Secondly, the 'raw score' on a traditionally constructed test must be compared to a reference group in order to understand what it tells us about the individual's level of ability. According to classical test theory a person's 'true' level of the attribute being measured is defined as the overall score on a test plus random error. The level of the attribute being tested can be defined only in relation to the particular test that has been used and the particular norm group selected as the reference point.

How these limitations distort the measurement of change

The difficulty of the test will limit the extent to which it is sensitive to differences at the extreme score ends of scales. This aspect of insensitivity is referred to as a 'ceiling' effect at the high end and a 'floor' effect at the low end. Ceiling effects occur when respondents cannot demonstrate their higher levels of an attribute because there are insufficient advanced items in the test, or because there is insufficient time for test-takers to complete the more advanced items. Floor effects similarly limit our understanding of what a person can and cannot do. When a person scores zero on a test we know this means they have a low level of the attribute, but we don't know how low because there are insufficient low level items to assess their current level of

functioning. When tests with too low ceilings or too high floors are used to measure change, researchers will necessarily draw mistaken conclusions about the degree of change or the relative merits of different interventions.

Another problem is that gain scores can have very different significance at different points in the scale. Consider an example from sport: the high jump (Prieler & Raven, 2002). For a 20-year-old athlete who is 2 metres tall to improve performance from 180 to 185cm would not be a great challenge. But the change from 220cm to 225cm is a very big increase in performance, because it is near the top of what an athlete can achieve. So too with psychological attributes: 'before and after' scores designed to assess responsiveness to learning programmes or behavioural therapy or drug treatment signify different things at different points in the distribution.

How IRT can help

Although classically constructed tests can measure change over time they can only assess the relative amount of change in relation to a particular norm group. IRT allows us to index change directly in relation to the individual (or group in the case of averaged scores), because ability level is quantified according to characteristics of individual items rather than the overall test score and is independent of the performance of a specific norm group. This allows us to predict more directly the performance on particular items with respect to level of the trait. Graphs called 'item characteristic curves' (ICCs) can be plotted showing directly the relationship between level of the trait being measured and probability of solving the item.

The direct relationship between trait level and item difficulty level means that IRT-based analysis is more appropriate for measuring changes in the trait over time and in response to interventions. More recent IRT models have been developed to allow us to take account of additional parameters; for example, we can index how well different items discriminate at differing levels of the attribute by considering the slope of the ICC – the steeper the curve the more discriminating the item is overall. Moreover the steepness of any one curve will vary at different points allowing us to identify at what level of ability the item discriminates most sensitively. This is important for measuring change since, as we have noted, gain scores can have very different significance at different points in the scale (Embretson, 1991).

'Practitioners should not be put off by the apparent complexity'

Using IRT in practice

It is true that the IRT-based methodology is more complex to apply than traditional methods, but software to ease the application of it has been available for nearly a decade now (Fischer & Ponocny-Seliger, 1998). That said, it is disappointing to note that IRT methodology has not yet been incorporated into SPSS.

Moreover, the IRT-based methodology can be applied to tests used for evaluation of change that are not themselves based on item response models.

By adopting this methodology to the evaluation of gain scores, we will avoid erroneous conclusions about the effects of interventions; we will be able to document the impact of interventions for individuals more accurately and specify more precisely the differential impact of interventions across groups of people with different characteristics. Readers who wish to learn more are referred to the excellent texts by Fischer (1991, 1995a, 1995b, 2001; Fischer & Ponocny-Seliger, 1997).

Conclusions

We cannot continue to use inferior methods to evaluate change; newer and more robust methodology is now available.

Practitioners should not be put off by the apparent complexity of the methodology; software is available to smooth its application. The theory is more accessible to practice than it has ever been. It is time to change our approach to change.

■ *Joerg Prieler is with Hogrefe Ltd, Oxford. E-mail: prieler@hogrefe.at.*

Contribute to a new Methods section in *The Psychologist*: e-mail your suggestions to jonsut@bps.org.uk.

DISCUSS AND DEBATE

How can we as a profession facilitate and encourage learning and use of IRT methodology?

To which current political, economic and social issues might psychologists contribute IRT solutions?

In terms of active promotion and use of IRT methodology, how does British psychology compare with psychology in other countries?

E-mail 'Letters' on psychologist@bps.org.uk or contribute (members only) via www.psychforum.org.uk.

References

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C.W. Harris (Ed.) *Problems in measuring change*. Milwaukee and London. University of Wisconsin Press.

Embretson, S.E (1991). Implications of a multidimensional latent trait model for measuring change. In L.M. Collins & J.L. Horn (Eds.) *Best methods for the analysis of change*. Washington, DC: American Psychological Association.

Fischer, G.H. (1991). A new methodology for the assessment of treatment effects. *Evaluación Psicológica, 7*(2),117–147.

Fischer, G.H. (1995a). Linear logistic models for change. In G.H. Fischer & I.W. Molenaar (Eds.) *Rasch models, recent developments and applications* (pp.158–180). New York: Springer-Verlag.

Fischer, G.H. (1995b). Some neglected problems in IRT. *Psychometrika, 60*(4), 459–487.

Fischer, G.H. (2001). Gain scores revisited under an IRT perspective. In A. Boomsma, J. Van Dujin & T.A.B. Snijders (Eds.) *Essays on item response theory* (pp.43–68). New York Springer.

Fischer, G.H. & Ponocny-Seliger, E. (1997). Multidimensional linear logistic models for change. In W.J. van der Linden & R.K. Hambleton (Eds.) *Handbook of modern item response theory* (pp,323–346). New York: Springer-Verlag.

Fischer, G.H & Ponocny-Seliger, E. (1998). *Structural Rasch modelling. Handbook of the usage of LPCM-Win 1.0,* Groningen: ProGAMMA

Lord, F.M. (1963). Elementary models for measuring change. In C.W. Harris (Ed) *Problems in measuring change*. Madison: University of Wisconsin Press.

Prieler, J.A. (2000). Evaluation eines Ausleseverfahrens für Unteroffiziere beim Österreichischen Bundesheer (Validation of personal selection of officers in the Austrian army). Unpublished doctoral dissertation, University of Vienna.

Prieler, J.A. & Raven, J. (2002). *The measurement of change in groups and individuals, with particular reference to the value of gain scores*. Retrieved 25 April 2007 via www.wpe.info/papers_table.html

Williams, R.H & Zimmermann, D.W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20*, 55–69.