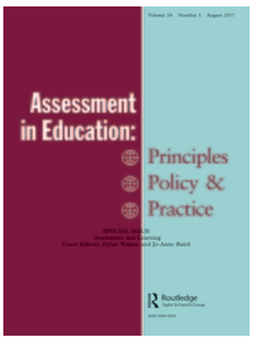

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: http://www.tandfonline.com/loi/caie20

Psychology's atomic bomb

Denny Borsboom & Lisa D. Wijsen

To cite this article: Denny Borsboom & Lisa D. Wijsen (2017) Psychology's atomic bomb, Assessment in Education: Principles, Policy & Practice, 24:3, 440-446, DOI: 10.1080/0969594X.2017.1333084

To link to this article: https://doi.org/10.1080/0969594X.2017.1333084

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE, 2017
VOL. 24, NO. 3, 440–446
https://doi.org/10.1080/0969594X.2017.1333084

Routledge
Taylor & Francis Group

COMMENTARY

Psychology's atomic bomb

Denny Borsboom and Lisa D. Wijsen

Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

ARTICLE HISTORY Received 7 February 2017; Accepted 17 May 2017

In their provocative and interesting article, Baird, Andrich, Hopfeneck, and Stobart (2017) suggest the possibility to articulate a post-modern account of psychometric testing. As a conceptual framework, postmodernism may not be every reader's cup of tea, but the themes that it addresses most certainly are of central importance if we want to better understand the role of psychometrics and educational testing in social reality.

The central role of educational testing practices in contemporary societies can hardly be overstated. It is furthermore evident that psychometric models regulate, justify and legitimise the processes through which educational testing organises our lives. For this reason, it is remarkable that few detailed conceptual investigations into these processes have been executed. There have been papers that focus on the philosophy of science behind psychometrics and its (lack of) substantive motivation (Borsboom, 2005; Maraun, 1998; Michell, 1997, 1999) but these rarely evaluate the relation between the psychometric models they analyse and the function of psychometric testing in the social sphere. On the other hand, there is a sizeable literature about the status of testing in society, as treated in the sociology of education, but this literature tends to take the psychometric models behind tests for granted. The power relations that are in play in education, such as the organisation and policies of educational institutions (Ballantine & Hammack, 2015), the regulation of knowledge by textbooks (Apple & Christian-Smith, 1991) and the influence of testing on learning (Frederiksen, 1984; Madaus, 1998), have been extensively studied. However, we find that the influence of psychometric properties themselves, such as measurement invariance or the use of sum scores, is largely absent from this research tradition.

Baird et al. (2017) invite us to go a level deeper and discuss the interaction between the *concepts* of psychometrics (e.g. unidimensionality) and societal issues. Perhaps closest to engaging in such interaction are Cronbach's (1988) and especially Messick's (1989) notions of validity, as cited by Baird et al. (2017), which called for an integration of facts and values that has a distinctly post-modern ring to it. However, these calls for integration are often philosophically limited, in the sense that they proceed from the idea that we should integrate the social consequences of testing in our validity judgement to honour our responsibility to 'do the right thing' (e.g. see especially Cronbach, 1988); sometimes, authors talk about 'adverse social consequences' of testing that need to be 'identified', as if such social consequences are akin to subatomic particles that we can discover, and as if educational tests and testing agencies can generally be expected to end up on the good side of history.

CONTACT Denny Borsboom dennyborsboom@gmail.com

A fair evaluation of the relation between psychometrics, testing and society cannot, of course, assume such things. It must leave open the possibility of reaching the conclusion that, e.g. psychometrics and educational testing have generally harmed rather than helped society, or that well-meant integrated validity judgments tend to lead us astray. Therefore, an open and historically even-minded survey of psychometric modelling and educational testing is called for, and may represent an important research topic. In this commentary, we aim to offer some observations that may be relevant for the analyses that Baird et al. (2017) undertake, and that could perhaps inspire further studies in that direction. We focus our comments on three topics: unidimensionality, measurement invariance and the psychometric legitimisation of power structures in western society.

The effect of unidimensionality

Baird et al. (2017) argue that the concept of unidimensionality is problematic in educational testing. This is indeed the case. Understood as a causal (rather than statistical; Markus & Borsboom, 2013) concept, unidimensionality means that a single dimension of individual differences determines the differences in every one of the items used in a test. This is plausible in, e.g. a psychological test for memory capacity, where the process and resources utilised in remembering the number series '8-476-3-26-0' are really the same as those in remembering the series '8938-47-354-3-356', so that individual differences in performance on such items can be plausibly considered to be determined by a common cause.

In many cases in educational testing, this assumption is not only unlikely to be met, but in fact undesirable. For instance, the test used in PISA, as discussed by Baird et al. (2017) is supposed to target the construct 'preparedness for life' (OECD, 2016). Now, it seems to us that the individual differences in representative items (how many bookshelves can a carpenter make with 4 wooden panels 12 small clips and 14 screws?) are not likely to be caused by individual differences in preparedness for life, and we also do not believe that the people constructing PISA believe such things. The items used in PISA and similar tests are a heterogeneous set taken from a variety of substantive domains, each of which recruits multiple neuro-cognitive processes and mechanisms, which are together deemed relevant for a person's ability to function in modern society.

In fact, one should neither expect nor desire unidimensionality in this kind of test, as one can easily see by considering the implication, which directly follows from any unidimensional IRT model for dichotomous items, that the expected score on one of the items exhaustively characterises the ordering of persons on the latent variable. This means that a unidimensional model implies that, if you knew people's expected score on the item 'how many bookshelves can a carpenter make with 4 wooden panels 12 small clips, and 14 screws?' this would tell you everything there is to know about the measured latent variable (here, people's preparedness for life). That is sufficiently absurd to justify the conclusion that unidimensionality, as causally understood, is generally a mirage in educational testing (of course, there are exceptions in the form of highly specified item formats that target well-defined psychological functions, but these are rarely used in educational testing).

Unidimensionality, thus, can hardly be causally motivated in broad educational tests, which are explicitly designed to assess individual differences in tasks of different content. However, it still plays an important role in the discourse surrounding tests, as is also evident in the controversies surrounding PISA (Kreiner & Christensen, 2014; Zwitser, Glaser, &

Maris, 2016), with some arguing that, unless a test is unidimensional, one cannot meaningfully compare total scores (literally taken, this does not follow: that one cannot interpret total scores in terms of differences on a single latent variable does not mean one cannot interpret them at all). In addition, PISA, like many testing projects, assembles items into tests partly on the basis of unidimensionality checks. Thus, even though in these applications unidimensionality is probably best seen as a concept we force upon reality, rather than a realistic hypothesis on the structure of the empirical world, it still deeply influences the way we operate with and reason about tests.

A concrete example of this power exerted by the notion of undimensionality is how the organisation of reality into tests that are supposed to measure a single attribute directly influences the actions of test constructors and policy-makers. As Baird et al. (2017) highlight, countries have actually been reported to reform their curriculum based on PISA results. Now, the fact that PISA is based on the notion of unidimensionality, and thus, selects items which tend to overlap, may lead countries to move their curriculum in the direction of the same quasi-unidimensional space. This will effectively mean that the curriculum is narrowed, so that it aligns with the same dimension formed in the PISA data. This is a very concrete way in which a concept, which is invented by psychometricians to think about reality (unidimensionality) changes the way tests are constructed, and ultimately feeds back to change the educational system itself.

Fairness and measurement invariance

Also in a broader sense, the notion of unidimensionality has strongly influenced how human beings, and specifically their capacities, are viewed. The rise of the test and thereby the notion of undimensionality has led to a more objective judgement about the capabilities of a person. Since its conceptualisation in the models of Spearman (1904) and his followers, unidimensionality has seemingly implied that the mental capabilities of human beings could now be reduced to a single score. These scores gave an objective indication of a person's abilities, but that exact same process has also led to the over appreciation of certain qualities such as verbal intelligence, and has taken away attention to other factors that, for instance, might explain the relation between intelligence and for example, job performance (McClelland, 1973). Thus, the focus on unidimensionality almost by its nature narrows the scope of theories about test scores.

This has had the concomitant effect that race, gender or religion were no longer factors that were to be taken into account when judging a person's cognitive abilities – a modern novelty, compared to most of history. In psychometrics, this notion became mathematically entrenched as the requirement of measurement invariance: the idea that a test should function in the same way across gender or across different demographical groups (Mellenbergh, 1989).

Measurement invariance is generally taken to be a very important concept in the comparison of groups (Kreiner & Christensen, 2014). However, it is worth noting that the concept itself is deeply intertwined with the causal interpretation of unidimensionality that, as Baird et al. (2017) have correctly noted, is highly problematic. However, despite the nuanced description of how tests relate to measurement in Baird et al. (2017), the fact remains that psychometric models like the Rasch model (the basis of PISA) are directly modelled after standard approaches to measurement in physics (Rasch, 1960), because the idea underlying

these models is simply that each of the item responses results from a trade-off between a single measured ability and an item-specific difficulty, and this idea is also deeply ingrained in the notion of measurement invariance.

If the common cause interpretation of the latent variable model is correct, then it is natural to see violations of measurement invariance across countries as undesirable; after all, if the measurement model underlying the item responses differs across groups, then the scores are no longer exchangeable, which means that it matters which items, precisely, are selected to compare countries (Kreiner & Christensen, 2014). As a result, items with differential item functioning should be removed, as they bias the estimation of the underlying ability, conceptualised as a real, causally efficient attribute on which a definite order in fact exists. After all, one cannot bias one group relative to another, if there is no standard of truth with respect to which 'bias' can be defined.

However, what if the assumption that the items all measure the same attribute is false, and the trait estimates construct a convenient dimensional index rather than measure an ability, as suggested in Baird et al. (2017)? In that case, there is no clear imperative as to what to do with items that violate measurement invariance. In fact, if items do not measure the same ability, then the question which items exhibit differential item functioning in which direction is actually much more interesting than the rank order of countries that politicians and newspapers tend to focus on; for these results can inform policy-makers on educational domains on which their country performs less well or better than one would expect given the overall performance (Zwitser et al., 2016).

Thus, shifting from a realist position on measurement ('these items measure the same ability') to a constructivist one ('these items are used to construct a unidimensional index of item scores;' Borsboom, Mellenbergh, & van Heerden, 2003) has direct implications for how one views concepts collateral to the psychometric model, like measurement invariance. Again, apparently philosophical choices regarding seemingly technical concepts in psychometric theory are more directly tied to societal concerns than most would suspect, and a deeper analysis of the relations between psychometric concepts and social reality is called for.

Psychometric power structures

Although some may consider their paper provocative in its explicit embracement of postmodernism, we think that in certain respects Baird et al. (2017) actually do not go far enough in penetrating the psychometric fabric of social reality. For instance, they approvingly quote Foucault's (1975) work, but scarcely exploit one of the most important concepts in that work when it comes to psychometric testing: that of *power*. One of the central tenets of postmodernism is that, by shaping and moulding the concepts and terminology we use to represent the world, we actually change our realities; and it is central to Foucault in particular that these changes allow certain groups to exert power over other individuals.

Now, there are few disciplines for which this description is more apt than psychometrics and educational testing. In the same way that the concept of psychiatric diagnosis (Foucault, 1963) is used to regulate society and, occasionally, to incarcerate individuals, so psychometric tests are wielded as instruments of power over almost all people in many contemporary societies: we use tests to determine who goes to which school, to regulate university admission, and to select people for jobs. In this respect, the rise of the notion

that certain abilities could be measured in the first place and that the psychological test, rather than the subjective judgement of a teacher or employer, is the appropriate method to do so, has changed society deeply.

As a result, the psychological test could be argued to be the most important invention in the history of psychology. In fact, the psychological test is to psychology what the atomic bomb is to physics: a scientific application that has radically changed global power structures. And if the psychological test is analogous to the atomic bomb, then psychometrics is analogous to nuclear physics. Not, of course, in the sense that it is a similarly advanced scientific theory (Michell, 1999; Sijtsma, 2012), but in the sense that psychometric theory regulates the assembly, use, and evaluation of psychometric tests, such as those used in educational testing.

What is the ideology supported by the psychometrics of educational testing, taken at large? It would seem to us that there can be little doubt that educational testing stands in service of the societal doctrine of meritocracy – i.e. the idea that one's position in society should be selected on the basis of one's abilities (possibly extended to assessment of character, as in personality testing). Of course, nothing about the concept of meritocracy is self-evident, as is clear from the fact that the great majority of societies that have existed over time scarcely employed. However, this idea is so ingrained in the educational testing literature that it is rarely if ever questioned, and one even gets the impression that much of the literature considers meritocracy intrinsically fair.

That, in our view, is a mistake. Of course, there are professions for which a certain level of functioning would be deemed necessary by all (e.g. one wants a bus driver who can drive a bus), but in current meritocratic societies like our own, the wish to select the 'best' candidates extends to virtually all capacities deemed measurable and relevant for any thinkable profession. In many cases, educational tests as regulated by psychometric theory are not in any way 'fair' in their treatment of individuals who have lower levels of cognitive ability: they generally have less access to educational possibilities, as a result will likely be employed in less desirable jobs with a higher liability for health problems, and in spite of that will often have less easy access to health care and lower life expectancies (Gottfredson & Deary, 2004). And, ironically, to compensate for these inconveniences, our societies 'reward' individuals who demonstrate lower cognitive ability as operationalised in psychometric tests by employing them in jobs that, on average, pay them lower salaries (Zagorsky, 2007).

And where do these differences in test scores come from? A recent study into the genetic architecture of CITO-scores, used for placement in the Dutch educational system, suggested that almost 60% of the variance these scores is of genetic origin (Bartels, Rietveld, Van Baal, & Boomsma, 2002). If this is correct, then educational tests serve as instruments in a practice that places individuals at different societal positions based on variance in test scores which largely reflect genetics. Ignoring for the moment whether this is a sensible thing to do or not, our psychometric operationalisation of meritocracy may, in this way, support the gradual formation of an *intellectual aristocracy*: a division of societal positions according to a stochastic function of genetic resemblance, indirectly measured through educational tests. And it does this by exerting power over individuals through the application of educational tests; a power which, in turn, is legitimised by the mathematical formulae of psychometric theories. That, we think, is the kind of psychometric food for thought that the Foucauldian scholar should digest.

It is an interesting question how learning theories, as discussed in Baird et al. (2017) should fit into this story. As Baird et al. (2017) acknowledge, current learning theories are insufficiently detailed and lack the scientific rigour needed to profitably inform psychometric models. As such, the literature is dominated by *theoretical reflections* on learning rather than an actual *science* of learning. If one wants to really accelerate the influence of learning theories on psychometric practice, the only viable option is to come up with a scientific theory of learning that is sufficiently formalised to dictate the structure of psychometric models and educational tests. Until that time, our educational system will have to live with tests that are slaved to psychometric concepts like unidimensionality, however, dubious their relevance to the subject matter. If only for this reason, the operating principles of standing psychometric theory require a much more critical assessment than they currently enjoy.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the H2020 European Research Council [ERC Consolidator grant number 64720].

Notes on contributors

Denny Borsboom is a professor of Psychological Methods at the University of Amsterdam. He has published widely on conceptual issues in psychometrics, validity theory and developed network modeling approaches to psychometrics.

Lisa D. Wijsen is a PhD student at the University of Amsterdam. She studies the history of psychometrics in relation to other sciences as well as societal developments.

References

Apple, M. W., & Christian-Smith, L. (Eds.). (1991). *The politics of the textbook*. New York, NY: Routledge.

Baird, J., Andrich, D., Hopfeneck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice 24*, 317–350.

Ballantine, J., & Hammack, F. M. (2015). *The sociology of education: A systematic analysis*. London: Routledge.

Bartels, M., Rietveld, M. J., Van Baal, G. C., & Boomsma, D. I. (2002). Heritability of educational achievement in 12-year-olds and the overlap with cognitive ability. *Twin Research, 5*, 544–553.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203–219.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Foucault, M. (1963). *The birth of the clinic: An archaeology of medical perception*. London: Routledge.

Foucault, M. (1975). *Discipline and punish: The birth of the prison*. New York, NY: Random House.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193–202.

Gottfredson, L. S., & Deary, I. J. (2004). Intelligence predicts health and longevity, but why? *Current Directions in Psychological Science, 13*, 1–4.
Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*, 210–231.
Madaus, G. E. (1998). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum: 87th yearbook of the national society for the study of education, Part I* (pp. 83–121). Chicago, IL: University of Chicago Press.
Maraun, M. D. (1998). Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory and Psychology, 8*, 435–461.
Markus, K., & Borsboom, D. (2013). *Frontiers of validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.
McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist, 28*, 1–14.
Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127–143.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American.
Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355–383.
Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge: Cambridge University Press.
OECD. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. Paris: OECD Publishing.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.
Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology, 22*, 786–809.
Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology, 15*, 201–293.
Zagorsky, J. (2007). Do you have to be smart to be rich? The impact of IQ on wealth, income and financial distress. *Intelligence, 35*, 489–501.
Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2016). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika*. doi:10.1007/s11336-016-9543-8