"social science research generally is not rocket science.
It is much more complex than that"

(Fulcher, 2015, p. 55)

The myth of measurement in social sciences

"Measurement, that is the objective representation of the attributes of objects and events of the real world by symbols on the basis of an objective empirical process, is a basic tool of modern human thought. It is the way in which we describe and reason about the world. Measurement has been developed through the physical sciences, which serve as a paradigm. From this basis its application has been extended to virtually all domains of human knowledge and discourse. However, the concepts and methods of measurement in this wider and more diverse range of disciplines offer significant conceptual problems, compared with measurement in the physical sciences that is the normative view of much metrological discourse." (Finkelstein, 2009, p. 1270)

"the concept of 'measurement', when used by psychological scientists, would be seen as a metaphor at best … and a conceptual error at worst." (Maul, Torres Irribarra & Wilson, 2016, p. 312)

"it is not only the *word* "measurement" that has been adopted by psychologists and educators, but an entire array of accompanying concepts such as quantity, units, networks of lawful relationships, causality, attributes, and so forth." (Maul, 2014, p. 40)

"Given that the intended interpretations of many (if not all) test scores involve claims about measurement, it would seem that clarity about the semantics of such claims is a necessary condition for clarity about test score interpretations. Furthermore, the concept of measurement commands significant social capital among laypersons and scientists alike, being associated with precision and trustworthiness; presenting test results as measurements in the absence of a coherent account of the semantics of such claims runs the risk of oversalesmanship, if not outright pathology (Michell, 1997, 1999, 2008)." (Maul, 2014, p. 40)

"So, perhaps we should entertain the hypothesis that the paradigm guiding psychometrics is after all not a scientific one but a *technological paradigm* (Dosi, 1982; Johnston, 1972). Like Kuhn's concept of *scientific paradigm*, that of a *technological paradigm* comprises a set of fundamental presuppositions (like the presupposition that psychological attributes are continuous quantities) and patterns of problem solving like [Item Response Theory] models, which guide the practices of the relevant discipline. However, there is a fundamental difference: while scientific paradigms guide researchers towards scientific discovery and thereby are chiefly responsive to relations between theories and evidence, technological paradigms guide disciplines in the construction of marketable products and thereby are chiefly responsive to the dynamics of the marketplace." (Michell, 2017, p. 420)

"Until the issue is satisfactorily tested, the claim to measure psychological attributes using psychological tests remains hypothetical. This fact, on its own, is not a defect. However, if there is no acknowledgment of the hypothetical character of the claim to be able to measure such attributes, then inquiry has become uncritical." (Michell, 2001, p. 213)

"Any science, as a social movement, serves a variety of interests, some not strictly scientific. When the processes of critical inquiry break down, it may be because these extra-scientific interests shape the discipline, thereby obtaining some advantage. Clues to these interests and advantages reside within the history of psychometrics." (Michell, 2001, p. 214)

Problem #1: The theoretical gap

How much progress have we made in developing theories of our constructs?

"If one attempts to sidestep the most important part of test behaviour, which is what happens between item administration and item response, then one will find no clarity in tables of correlation coefficients. No amount of empirical data can fill a theoretical gap." (Fulcher, 2015, p. 167)

“Within the domain of quantitative methodologies, mythologies flourish under a particular set of ingredients that include, but are not limited to, conceptual equivocation (or, in the worst cases) the total failure to pin down concepts with definitions), the misidentification of related concepts, the dogmatic adherence to favored statistical props, the failure to clarify relationships between statistical tools of representation and the components of empirical settings that they were designed to represent, and the projection onto empirical reality of mathematical necessities” (Maraun, Gabriel & Martin, 2011, p. 782)

“It may be useful here to consider the case of the measurement of intelligence as an example of the case where it is easier to devise test, than to establish what it is that they are measuring.” (Finkelstein, 2009, p. 1273)

“meaningful measurement is possible only if enough is known about the attribute so as to justify its logical operationalization into prescriptions from which a measurement instrument can be developed. An immense problem in psychology is that theories about attributes are often not precise enough to justify a logical operationalization.” (Sijtsma, 2012b, p. 787)

“The greatest problem of psychological measurement is the frequent absence of well-developed attribute theories. Instead, items are often constructed guided by best guesses on the basis of whatever theory is available, but also based on intuition (what seems to be reasonable?), tradition (how were similar tests constructed?), and conformity (what do colleagues do or think?). Unfortunately, the role of attribute theory is often underestimated, and test and questionnaire construction seen as engineering and sets of items as useful measurement instruments.” (Sijtsma, 2012b, p. 790)

“test construction and test practice are plagued by bad habits. Construct validity is often ascertained by means of highly exploratory research strategies and is in need of more direction; reliability is often estimated using one of the worst methods possible and is given an incorrect interpretation … Not only are several validity and reliability issues unresolved or at least continue to be at the center of much debate, novel insights also seem to have trouble finding their way to the community of psychological researchers” (Sijtsma, 2009, p. 169)

Problem #2: Objectivity

How objective can our assessment processes really be?

“It is argued that in social systems the observer and analyst are not objective, but operates on the basis of ideologically motivated theories. The objects of observations are humans. They have their beliefs, desires and methods of reasoning and may not be amenable to description by simple models. The understanding of their behaviour must be based on empathy and the experience of life. There are thus philosophical challenges to the application of measurement to systems involving human actors.” (Finkelstein, 2005, p. 269)

“An important class of the descriptive assignment of numbers, the measurement status of which is problematic, arises in educational

testing. Marks in examinations may be objective, and are based on an empirical process, but it is problematic what they measure, other than the performance in a particular test. It is doubtful whether, when marks are treated on a ratio scale, they are not in fact measures on an ordinal scale. This affects the meaningfulness of statistics on marks, such as t[he] calculations of averages and the like. The conflation of marks, such as the calculation of weighted sums of marks, contains an element of subjectivity in the conflation scheme, which probably disqualifies such conflated marks from being considered measurements." (Finkelstein, 2003, p. 47)

Problem #3: Levels of measurement

What form of measurement are we using: nominal, ordinal, interval or ratio?

"The achievement tests I work with are generally aimed at assessing competence in broadly defined domains of knowledge, skills, and/or judgment, and in most cases, even a simple ordering [of students' test responses] … could be questionable. Taking achievement in chemistry as an example, different people, *a* and *b*, would typically have different patterns of competence. Person *a* might be good at solving numerical problems but perform badly in the lab, and person *b* might show the opposite pattern. Which person is higher in overall achievement in chemistry? Given an area of achievement that is broadly defined, we are likely to have, at best, a partial ordering, unless we arbitrarily decide that some patterns are better than others. … I think that most educational and psychological variables are not quantitative" (Kane, 2008, p. 104)

"the suggestion that psychological experiments could be done with the same degree of precision as physical experiments is far too optimistic given that the experimental manipulation of human behaviour is more prone to random and systematic error than that of physical phenomena. … It is more realistic to accept weaker forms of measurement, such as ordinal measurement or nominal measurement, as useful alternatives. Michell made this suggestion, which I think is a fruitful alternative to counting the number of points earned on a set of items, and pretend[ing] this is measurement, as psychologists often do." (Sijtsma, 2012b, p. 805)

Problem #4: Appropriate statistics

How well do we understand the statistical tools we are using?

"using statistics without sufficient experience is asking for trouble. … If one knows little about statistics but needs statistics on a near-daily basis, the intuitive heuristic reaction to a statistical problem is the natural reaction that unavoidably will produce errors." (Sijtsma, 2016, pp. 9-10)

"Reliability estimation is not so much difficult but plagued by strong habit, which has created a persistence in using old but inferior lower bounds, coefficient alpha in particular. The problem also is in the statistical complexity of alternatives, such as the glb, and estimation based on generalizability theory and structural equation modelling, which are not readily available to test constructors through a simple mouse click" (Sijtsma, 2009, p. 190)

Problem #5: Levels of analysis

Can we use the statistical tools to draw conclusions about individuals?

"Articles reporting construction of short tests often discuss test-score reliability but Mellenbergh (1996) correctly noted that the group characteristic of reliability is not very informative about the precision of individual measurement. Hence, reliability values of 0.8 or even 0.9 do not guarantee accurate individual measurement." (Sijtsma, 2012a, p. 10)

Problem #6: Faulty assumptions

Classical Test Theory

"Applying classical test theory is easy, and a commonly accepted escape route to avoid notorious problems in psychological testing, such as constructing unidimensional tests. The model is, however, so enormously detached from common interpretations of psychological constructs, that the statistics based on it appear to have very little relevance for psychological measurement. Coupled with the unfortunate misinterpretation of the true score as the construct score, of random error as irrelevant variation, and of reliability as some kind of fixed characteristic of tests, instead of as a population dependent property of scores, it would seem that large parts of the psychological community are involved in self-deception." (Borsboom, 2005, p. 47)

Item Response Theory

"*Educational Measurement* is largely written from a highly specific psychometric perspective. Apart from a handful of authors taking their lead from generalizability theory, most authors either explicitly or implicitly reason from a two- or three-parameter logistic [Item Response Theory] model; that is, they assume unidimensionality, continuity, local independence, smooth item response curves, normal latent distributions, and so on. It would surprise me if such assumptions were indeed satisfied in typical applications of educational testing." (Borsboom, 2009, p. 709)

The Rasch model

"several assumptions have already been made. The first, and most obvious one, is that only one term or quantity ... is necessary to characterise an individual or, to put it another way, an individual's ability is 'unidimensional'. Likewise, every item has only one characteristic, its difficulty. Although various methods have been suggested for testing the assumption of 'unidimensionality' of difficulty, there has been little work on the problem of adequately testing the 'unidimensionality' of ability. Indeed, the usual methods of testing unidimensionality of difficulty are based essentially *on the assumption* of the unidimensionality of ability." (Goldstein, 1979, p. 214)

The alternative

Hermeneutics, 'the art of interpretation'

"There is a crisis mentality accompanied by a flurry of activity to design assessment and accountability systems that both document and promote desired educational change. Current conceptions of reliability and validity in educational measurement constrain the kinds of assessment practices that are likely to find favour, and these in turn constrain educational opportunities for teachers and students." (Moss, 1994, p. 10)

"From a psychometric perspective, the call for "detached and impartial" … assessment reflects a profound concern for fairness to individual students and protection of stakeholders' interests by providing accurate information. From a hermeneutic perspective, however, it can be criticized as arbitrarily authoritarian and counterproductive, because it silences the voices of those who are most knowledgeable about the context and most directly affected by the results [i.e., students and teachers]. Quantitative definitions of reliability locate the authority for determining meaning with the assessment developers. In contrast, Gadamer (cited in Bernstein, 1983) argues that the point of philosophical hermeneutics is to correct "the peculiar falsehood of modern consciousness: the idolatry of scientific method and of the anonymous authority of the sciences" (p. 40) and to vindicate "the noblest task of the citizen – decision-making according to one's own responsibility – instead of conceding that task to the expert" (p. 40)." (Moss, 1994, p. 10)

"Regardless of whether one is using a hermeneutic or psychometric approach to drawing and evaluating interpretations and decisions, the activity involves inference from observable parts to an unobservable

whole that is implicit in the purpose and intent of the assessment. The question is whether those generalizations are best made by limiting human judgment to single performances, the results of which are then aggregated and compared with performance standards, or by expanding the role of human judgment to develop integrative interpretations based on all the relevant evidence." (Moss, 1994, p. 8)

"The descriptions of students' activities and the products of their work represent the 'data' that the teacher as researcher uses for ongoing assessment of students' learning. Such data is multifaceted, when compared with either a multiple choice test of writing or a holistically graded essay. These highly varied and inconsistent sets of data psychometrically speaking … [constitute], psychometrically speaking, a highly unreliable picture that provides a highly valid representation of what the [student] knows." (Williamson, 1994, p. 168)

"If reliability is put on the table for discussion, if it become an option rather than a requirement, then the possibilities for designing assessment and accountability systems that reflect a full range of valued educational goals become greatly expanded." (Moss, 1994, p. 10)

Portfolio assessment

"Often, the need to focus on competing textual needs at the same time overwhelms uncertain writers, nonnative or native speakers, in timed situations. In the portfolio assessment context, ESL writers can be convinced that concentrating on ideas, on content, support, text structure, and so on, are worthwhile because they need not fear the cost of such attention to achieving technically correct language - which most of them have been conditioned to believe teachers value first and foremost." (Hamp-Lyons & Condon, 2000, p. 61)

Further reading

Eiko Fried & Jessica Flake's 'Measurement Matters':
https://docs.google.com/document/d/11jyoXtO0m2lUywpC04KjLvI5QcBUY4YtwEvw6cg2cMs

Contact me

kyle.smith@hdr.qut.edu.au

References

Borsboom, D. (2005). *Measuring the Mind*. Cambridge: Cambridge University Press.

Borsboom, D. (2009). Book review: Educational Measurement (4th ed.). *Structural Equation Modeling: A Multidisciplinary Journal, 16*(4), 702-711. doi:https://doi.org/10.1080/10705510903206097

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement, 34*(1), 39-48. doi:https://doi.org/10.1016/S0263-2241(03)00018-6

Finkelstein, L. (2005). Problems of measurement in soft systems. *Measurement, 38*(4), 267-274. doi:https://doi.org/10.1016/j.measurement.2005.09.002

Finkelstein, L. (2009). Widely-defined measurement - An analysis of challenges. *Measurement, 42*(9), 1270-1277. doi:https://doi.org/10.1016/j.measurement.2009.03.009

Fulcher, G. (2015). *Re-examining Language Testing: A philosophical and social enquiry*. London: Routledge.

Goldstein, H. (1979). Consequences of using the Rasch model for educational measurement. *British Educational Research Journal, 5*(2), 211-220. doi:10.1080/0141192790050207

Hamp-Lyons, L., & Condon, W. (2000). *Assessing the Portfolio: Principles for Practice, Theory, and Research*: Hampton Press.

Kane, M. (2008). The benefits and limitations of formality. *Measurement: Interdisciplinary Research and Perspectives, 6*(1-2), 101-108. doi:https://doi.org/10.1080/15366360802035562

Maraun, M. D., Gabriel, S., & Martin, J. (2011). The mythologization of regression towards the mean. *Theory & Psychology, 21*(6), 762-784. doi:https://doi.org/10.1177/0959354310384910

Maul, A. (2014). Justification is not truth, and testing is not measurement: Understanding the purpose and limitations of the Standards. *Educational Measurement: Issues and Practice, 33*(4), 39-41. doi:https://doi.org/10.1111/emip.12055

Maul, A., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement, 79*, 311-320. doi:https://doi.org/10.1016/j.measurement.2015.11.001

Michell, J. (2001). Teaching and misteaching measurement in psychology. *Australian Psychologist, 36*(3), 211-217.

Michell, J. (2017). On substandard substantive theory and axing axioms of measurement: A response to Humphry. *Theory & Psychology, 27*(3), 419-425. doi:https://doi.org/10.1177/0959354317706746

Moss, P. A. (1994). Can there be validity without reliability. *Educational Researcher, 23*(2), 5-12. doi:https://doi.org/10.3102/0013189X023002005

Sijtsma, K. (2009). Correcting fallacies in validity, reliability, and classification'. *International Journal of Testing, 9*(3), 167-194. doi:https://doi.org/10.1080/15305050903106883

Sijtsma, K. (2012a). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika, 77*(1), 4-20. doi:https://doi.org/10.1007/s11336-011-9242-4

Sijtsma, K. (2012b). Psychological measurement between physics and statistics. *Theory & Psychology, 22*(6), 786-809. doi:https://doi.org/10.1177/0959354312454353

Sijtsma, K. (2016). Playing with data – or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika, 81*(1), 1-15. doi:https://doi.org/10.1007/s11336-015-9446-0

Williamson, M. (1994). The worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing, 1*(2), 147-173.