PART II

Practical Measurement Issues: Lessons From 75 Years' work With *Item Response Theory*: Benefits, Problems, and Potential Solutions

In developing his *Progressive Matrices* (RPM) tests, J. C. Raven anticipated the development of *Item Response Theory* (IRT) in that he plotted what (technical disputes over terminology excepted) have since become known as *Item Characteristic Curves* (ICCS). These showed how the proportion of respondents getting each item right varied with total score. If the curves were irregular he tried to find the cause and correct (or, if necessary, reject) the item. He incorporated the curves for all the items into a single graph, so that he could see how closely the shapes of the curves corresponded to each other, whether they crossed over (implying that the order of difficulty varied with the ability of the respondents), and whether they were, as far as possible, equally spaced.

Although the logic for what he was trying to do was briefly explained in the test Manual (then known as the "Guide to the Use of" one or other of the tests) and elsewhere, the measurement model was not sufficiently differentiated from Classical Test Theory for most readers to appreciate just how distinctive it really was. This has only become clear to a significant number of people as a result of recent developments in Item Response Theory (IRT). Yet, although these developments have resulted in the logic of the approach being more widely understood, the fact that the construction of the RPM was based upon them still generally passes unnoticed. Failure to appreciate just how different the measurement model deployed in the construction of the RPM was from classical test theory unfortunately resulted in some fairly widespread criticism of the tests stemming from attempts to apply procedures associated with classical test theory to evaluate the internal consistency of the RPM and to endless erroneous conclusions being drawn from research.

Only recently, by, with considerable difficulty, replicating Raven's methods using modern computer programs has it become possible to appreciate how close Raven had come to placing the scientific status of "eductive" ability, and the RPM as a measure of it, beyond dispute.

The chapters in this Section belatedly rectify these oversights.

The chapter by Anca Dobrean (née Domuta) describes the sampling procedures employed in the Romanian standardisation of the SPM ***Plus*** that yielded the data base on which most of the later chapters in this Section are based.

Raven, Prieler, and Benesch compare computer-generated Raven-type "empirical" ICCs with those produced using modern IRT programs. It emerges that the most widely applied version of IRT – the 1 parameter model – can yield results which seriously mislead researchers. Serendipitously, the research ends up demonstrating that both eductive and reproductive abilities are every bit as "real" as – and measurable in the same way as – high jumping ability or life expectancy.

The author's chapter summarising research conducted whilst a Romanian version of the *Mill Hill Vocabulary Scale* was being developed again reveals – perhaps in an even more striking way – that widely promoted IRT programs do not deliver the expected benefits. On the other hand, *Distractor* Characteristic Curves – i.e. plots of how the choice of each *wrong* answer varies with total score – yield information which is very useful to test developers. Beyond that, the chapter illustrates just how difficult it is to create a genuinely parallel version of what must be almost the archetypical form of IRT test – a vocabulary test made up of words of increasing difficulty.

The chapter by Prieler and Raven discusses the enormous methodological problems which arise in the thousands of studies which claim to measure and compare change – whether in groups or individuals – using more or less any test developed on the basis of Classical Test Theory … or even IRT-based tests which do not yield linear Test Characteristic Curves. Such test may, collectively, be described as being grounded in “arbitrary *metrics*”. But equally, if not more, serious errors in evaluation studies said to provide the basis for “evidence based treatment” (for example, in psychotherapy or education) stem from the adoption of what are best described as “arbitrary *measures*” … ie evaluation studies in which the researchers have concentrated their attention on only one or two outcomes (perhaps measured with highly reliable tests) instead of trying to get a rough fix on all potentially important outcomes – ie on the *comprehensiveness* of the evaluation. Both deficits can be overcome by adopting IRT-based procedures developed by Fischer and outlined in this chapter.