DOI: 10.1002/ppi.1423

WILEY

REVIEW

The trouble with numbers: Some fundamental flaws with using standardised outcome measures

Brian Rodgers

School of Public Health & Psychosocial Studies, Auckland University of Technology. Auckland, New Zealand

Correspondence
Brian Rodgers, Auckland University of Technology, South Campus, Private Bag 92006,Auckland 1142, Aotearoa New Zealand.
Email: brian.rodgers@aut.ac.nz

Abstract

The modern paradigm of evidence-based practice dominates the therapeutic world and influences all aspects of the profession. Yet this pervasive concept is based on surprisingly shaky ground. When looked at in detail, the source of the raw data used as the basis of much of this evidence, standardised outcome measures, can be seen to be fundamentally flawed. This article sets out the many methodological, sociopolitical, and technical flaws in standardised outcome measures, and asks what this means for the field of psychotherapy.

KEYWORDS

critique, methodology, outcome measures, psychotherapy, sociopolitical

1 | INTRODUCTION

Standardised outcome measures are at the heart of the evidence-based practice paradigm. In this paradigm, such measures provide a standardised "ruler" which allows the objective comparison of results of the effectiveness of interventions not just within a study, but across multiple studies. It is only via this standardised comparison across multiple controlled studies that empirically supported treatments are identified (Chambless & Hollon, 1998). Without this standardised ruler, no such comparisons can be made.

Within the fields of psychotherapy, counselling, psychology, and associated disciplines, such measures typically consist of a list of items in the form of questions, statements, or observations relating to a person's symptoms, behaviour, functioning, well-being, quality of life, etc. Each response to an item is assigned a numerical value, either a simple binary value (e.g., 1 = True, 0 = False) or using some sort of intensity scale (e.g., 0 = Never, through to 5 = Always). These values are then totalled according to a standardised schema to produce scores on one or more scales or dimensions (e.g., overall psychological distress, level of depression, functioning, etc.). Typically, a questionnaire is given to the client before therapy commences, then again some time later (usually at the end of therapy), and the change in scores is calculated to give a representation of the success or otherwise of the therapy.

When combined across multiple studies, this elegant and straightforward system provides compelling evidence for the efficacy of a therapeutic approach or particular intervention. But what if the data that these claims are based on is fundamentally flawed? What if this "standardised ruler" is not as straight and linear as claimed? This article argues that the claims of standardisation and objectivity of measurement within research on psychological therapies is grossly overstated, and that, when one looks in detail at the process of outcome measurement, significant flaws are

apparent. Specifically, the article will articulate a number of methodological, sociopolitical and technical issues that challenge the perceived elegance and simplicity of the predominant outcomes paradigm, and call for us to rethink how we utilise outcome measures in relation to practice.

2 | METHODOLOGICAL FLAWS

McLeod (2001) presented an interesting discussion of the history behind contemporary outcome measures. He pointed out that the self-report questionnaires used so routinely today to measure the outcomes of therapy were not actually designed for this purpose, but initially developed to improve the efficiency of existing methods of screening people for different roles (e.g., in the military or as employees, students, etc.). Only later did psychologists adopt these questionnaires as a way of incorporating greater quantification and objectivity into their method. This allowed the importation of the experimental method into psychology, making it resemble a "hard science." For this approach to work, measures had to demonstrate their validity and reliability, so great emphasis was placed on demonstrating the psychometric qualities of measures in controlled conditions. Once a measure was shown to be valid and reliable, it could then be used with confidence in other settings. What McLeod (2001) identified is that, while this approach may have a lot of appeal in terms of "fitting" with experimental method, it presents something of an "administratively created reality" whereby the only outcomes seen to be of importance are those that can be measured on an outcome questionnaire. Here McLeod (2001) referred to the critique of psychology by Danziger (1990, 1997), where it can be seen that intelligence becomes what can be measured by intelligence tests, with the result that school children are taught how to pass tests rather than how to actually learn. By extension, "successful" therapy becomes that which allows a client to respond to an outcome questionnaire, based on outcomes determined by the researcher and/or therapist, such that his or her "score" improves from before to after treatment.

Similarly, Slife (2004) critiqued the underlying naturalistic foundations upon which standardised self-report measures are constructed. By focusing on what is observable and measurable, naturalistic researchers are required to "operationalise" what they are investigating (i.e., to find a way to define something in terms of a measurable quantity). This approach inherently misses that which cannot be measured, such as the spiritual, cultural, transpersonal or existential, and even some relational elements of therapy (as only the things *having* the relationship can be observed, not the relationship itself). This can lead to a study of the manifestation of a phenomenon, not the phenomenon itself. Further, it is up to the researcher to choose which manifestation represents which phenomenon prior to undertaking a study, such that this choice becomes an implicit part of the method, and so is usually never questioned: "The upshot is that we cannot know for sure with traditional scientific methods what is actually being studied in our research investigations because we cannot know with certainty what a particular operationalization means" (Slife, 2004, p. 55).

Michell (1997, 2000) has taken this critique of operationalism further and argued that the current practice of quantification in psychology is fundamentally flawed. He presented the argument that the assumptions upon which psychometrics are based are inherently unscientific, and that they represent a "pathological" lack of acknowledgement of the implicit methodological limitations involved (Michell, 2000). Here, Michell (1997) referred to the adoption by psychology of Steven's (1946) theory that measurement is "*the assignment of numerals to objects or events according to rules* [emphasis added]" (p. 360). In contrast, Michell (1997) stated that measurement is properly defined as "*the estimation or discovery of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute* [emphasis added]" (p. 358). Put simply, this is stating that measurement should be the process of discovering or estimating the existing "quantitative structure" of what is being measured, not just putting numbers to things. For example, the concept of "length" lends itself to measurement because it has an existing "quantitative structure"; thus, there is an existing logical structure and relationship between different attributes of "length," which can be discovered.

Michell's (1997) assertion was that the underlying hypothesis of the quantitative structure of psychological constructs has remained untested. Consequently, all attempts to measure psychological attributes are fundamentally flawed. For example, no attempt has been made to demonstrate that "depression" is a quantitative attribute, and yet

measures of depression freely assume that it is. The Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), for example, assumes that responses to statements are interval data—that is that the numbers assigned to each item are of equal interval to one another so that they can be mathematically totalled and averaged, and the results statistically compared (Stevens, 1946). What Michell has challenged is the assumption that a response to an item such as "I am so sad or unhappy that I can't stand it" can uncomplicatedly be added to "I feel I may be punished" as if they were on the same linear dimension, and equivalent to another person's responses to two completely different questions such as "I feel irritated all the time" or "I am worried that I am looking old or unattractive." If the premise of the validity of operationalism is removed, then the above becomes nonsensical—along with all the results of research studies that utilise such an approach.

3 | SOCIOPOLITICAL FLAWS

In addition to his critique of operationalisation above, Slife (2004) has challenged the "hedonistic" tendency of most approaches to therapy research, namely the fact that most traditional therapy outcome measures tend to focus on the betterment of self. This places severe constraints on research into the meaning and purpose of suffering, self-sacrifice, and other aspects that do not fit with this assumed aim of therapy. There is, as Frank (1978) observed, no place for the study of "the redemptive power of suffering, acceptance of one's lot in life, adherence to tradition, self-restraint and moderation" (cited in Slife, 2004, p. 64) as these do not fit with our modern Western value system. The potential of finding meaning in suffering without necessarily eliminating it, or of helping people to experience greater purpose in life without necessarily "taking away the pain" or "making it better" is typically overlooked. Similarly, by focusing on the individual, traditional methods of research into psychotherapy outcomes miss the wider social context, and risk supporting and re-emphasising cultural biases inherent in Western psychotherapy theories. By assuming that problems are "contained" in the client, and that these can uncomplicatedly be brought to the research setting to be measured, contextual elements of the process and outcome of therapy are inevitably missed—and consistently so.

Similarly, Hughes (1995) has argued that the processes inherent in outcome measurement reinforce differences between the dominant group and "others." Differences are *named* by the dominant group as being worthy of investigation (e.g. depression, anxiety, etc.). Once something is named, it is then *quantified* so that it can be measured. This quantification creates the illusion that subjectivity and politics have been transcended, that the "numbers" equate to "objectivity," and that this objectivity is valued because it is perceived to be value-free. *Statistical analysis* further enhances claims to objectivity, whereby the researcher is seen to be a neutral observer and to employ a "neutral" method, so that what emerges from the analysis is considered an unbiased "fact." This process leads to the *reification* of the initial abstract concepts into concrete "knowledge," which can be used to segregate and classify, as well as forming the basis for social, political, and economic decisions and judgements. The culmination of this process is *objectification*, whereby something subjective is turned into an object, a "thing" which is "other."

Through such processes, outcome measures effectively become political devices that control the flow of resources, finance, and influence. Research studies that utilize standardised measures are seen to have greater credibility and are more likely to be published in higher ranked journals or attract more prestigious grant funding. Hence, employers will typically reward researchers who utilise such measures in their studies more than those who do not. Similarly, treatment approaches that align themselves with the values implicit in these standardised measures are likely to demonstrate greater efficacy, and hence be seen as "empirically supported." These approaches will then be reckoned more worthy of funding by health services, insurance agencies etc., creating more opportunities for paid employment. This will in turn influence practitioners to align their practice to such treatment approaches, as well as promoting demand for training in them. Even clients are not immune to such influence, with those who are able to respond to questionnaires in such a way as to register at the "clinical" level becoming eligible for funded treatment, whilst those who do not are potentially deprived of treatment or required to fund their therapy themselves.

As can be seen, the values embedded in standardised outcome measures pervade all aspects of the profession. For minority and/or indigenous cultures, such mechanisms become yet another instrument of colonisation, with Western values and attitudes being unilaterally imposed (Tudor, 2012). Here standardised outcome measures become the arbiter of what is considered "normal" and healthy, whilst anything that does not fit with this is judged to be abnormal or unhealthy, or is silenced completely. Behaviours and experiences that potentially have cultural appropriateness become judged as unacceptable or indicators of distress. For example, whilst hearing voices, having visions or experiencing premonitions is culturally acceptable and even prized in some cultures, such experiences are usually considered pathological through a Western lens (Luhrmann, Padmavati, Tharoor, & Osei, 2014). Questionnaire items such as "Unwanted images have been distressing me," "Disturbing thoughts come into my mind that I cannot get rid of" and "I feel that something bad is going to happen," each and all of which could be seen as appropriate, valuable and meaningful experiences from a cultural perspective, are, within standardised outcome measures, uniformly categorised as problematic and indicators of ill-health.

4 | TECHNICAL FLAWS

Even if the above issues are ignored, numerous technical flaws still exist in the use of outcome measures within psychotherapy research. For example, "response error" is an acknowledged issue in questionnaire design, whereby there is (seen to be) a discrepancy between a theoretical "true score" that a respondent would report in a perfect situation and the score that is actually recorded (Willis, 2005). Deviations from this "true score" are seen to arise from people not really understanding the questions being asked of them, not being able to recall relevant information accurately, using flawed judgement or estimation strategies, having trouble mapping their internal judgements onto one of the response items, or editing their answers in a misleading way before reporting them (Tourangeau, 2003). What is being identified here is that, when a questionnaire is put in front of someone, their responses are not a simple download of information, but are more likely to be "generative" rather than "true."

Meier (1994, 2008) has argued that this discrepancy arises out of a mismatch between a given questionnaire and the respondent's current or ongoing cognitive, affective and behavioural state. For example, if a client's reading level does not match that of the questionnaire, they may guess at the meaning and "generate" a response, or a client may "fake good" in order to convince a researcher that the therapy is working. Even when clients are appropriately motivated and capable of responding honestly and accurately, their response behaviour may be influenced by external factors. For example, the presence of a researcher may distract a client, or they may feel they do not have time to think about the questions fully, or the questionnaire may not allow them to respond in a way they feel is appropriate. What Meier (1994) has highlighted is that when there is a significant mismatch between a questionnaire and a respondent, then supposedly "objective" measures become more akin to projective tests containing ambiguous stimuli which result in idiosyncratic associative responses.

Similarly Schwarz, Grayson, and Knäuper (1998) have identified the potentially problematic process for clients of determining the quantitative meaning of generally ambiguous response options. Most standardised self-report questionnaires ask respondents to rate some aspect of their experience or feelings using a form of Likert scale, such as "Not at all, Only occasionally, Sometimes, Often, and Most of or all the time." This seemingly simple request is, however, extremely complex. Not only are respondents asked to determine the quantitative meaning of an ambiguous label (e.g., how often is "Often"), but also to quantify their own feeling or experience and map this on to it. For example, the question "I have been happy with the things I have done" asks a person to rate how often they have felt this way over the last week. First, this requires the person to conceptualise the meaning of "happiness" for themselves, which, from a therapeutic, let alone a political perspective, is no trivial task. Next, the person has to connect with their experience over the last week and to identify moments that more or less approach this concept—again, assessing how close a feeling is to "happiness" may be no easy matter. Then, the person is expected to collate these variable experiences to form some sort of aggregate—but does "a little happy" count as half of "very happy" in rating an experience?

Finally, the person is required to compare their internal aggregate against the vague labels on the questionnaire, and, thus, have to decide whether their "2.5" experiences of happiness amount to "Only occasionally" or "Sometimes"? Due to the complexity of this process, it is very difficult to know if the client's understanding of a questionnaire item taps the same facet of an issue and the same "evaluative dimension" as that intended by the researcher (Schwarz, 1999).

These issues become even more complex when using a questionnaire to determine the amount of change a client has undergone from before to after therapy. Meier (1994) has indicated that some aspect of the respondent is bound to be different from one occasion to the next, thereby changing *how* a person responds. With respect to counselling and psychotherapy outcome research, this is particularly problematic given the likely effects of being in therapy. Using Meier's model of potential sources of "response error" above, it is very likely that a client may well have experienced a change in their cognitive and/or affective characteristics as a result of being in therapy. This would then introduce a completely different set of "response errors." Any difference registered may be attributable more to the way the questionnaire is responded to than to a meaningful change in the person's life.

Similarly, McLeod (2001) has highlighted that the process of therapy potentially introduces clients to new ways of defining and making sense of their situation. He has identified that therapy outcome measures are full of terms and concepts inherent in the "discourse" of therapy. A client new to therapy may well initially answer a questionnaire with a very different understanding of what is being asked than when they have finished their therapy. For example, a newcomer's response to the item "I have thought I am to blame for my problems and difficulties" might be "Not at all" because they believe everyone else to be at fault. At the end of therapy they may have come to see how their actions have affected others and respond with "Often" as a sign of taking greater responsibility for things. Here the therapeutic values of "responsibility" and "accountability" can be seen to have influenced the way the question is interpreted, fundamentally changing the meaning of the questionnaire item. This shift does not match at all with the intended scoring scheme, which would interpret this change as a negative outcome of the therapy. Like Meier (1994), McLeod makes the point that the potential for changes in the way questionnaire items are interpreted creates a difficulty when considering the results of outcome studies. One can never be certain that reported outcome scores are representative of actual changes in behaviour, life functioning, well-being, etc, rather than changes in the meanings attributed to questionnaire items.

The above can be seen as indicating that, rather than measuring along a single, linear dimension, responses to self-report questionnaires can be considered indicative of different types of change. Golembiewski, Billingsley, and Yeager (1976) suggest that change can be conceptualised in terms of at least three distinctly different classifications: alpha, beta, and gamma change. *Alpha* change is considered to occur when the meaning of the construct being measured and the psychological interpretation of the units of measurement in the outcome measure stay the same; that is, the change measured is a "true" representation for the respondent of the actual change that has occurred. Here the measure is acting as a standard "ruler." On the other hand, *beta* change occurs when the meaning of the construct being measured stays the same, but the psychological interpretation of the units of measurement changes; for example, the respondent uses a different set of criteria to rate the intensity of distress after therapy than before. Here the change measured is either an under- or overrepresentation of the actual change that has taken place for the respondent. In this instance, the measure is now acting like a "rubber ruler," which has stretched or compressed from one reading to the next. Finally, *gamma* change involves a "quantum shift"—a redefinition or reconceptualisation of the "psychological space" for the respondent. This is a major change in the perspective or frame of reference within which the phenomena are perceived, such that the previous meaning of the "measurement" becomes irrelevant. The above example of the change in the way a client understands the item "I have thought I am to blame for my problems and difficulties" would be representative of gamma change. Here the measure is no longer acting as a "ruler" at all, but more like a projective test, where responses need to be "interpreted" rather than "calculated" (Semeonoff, 1976).

The problem for outcome measurement from Golembiewski et al.'s (1976) perspective is that standardised quantitative questionnaires are typically designed purely to report alpha change, that is, a change in the mean score of a measure from before to after therapy. There is an assumption of linearity between responses made before therapy and those made after therapy. If responses show a significant change, then it is assumed that "real" change

has occurred, rather than that there has simply been a recalibration of the measurement dimension as in beta change. Conversely, if no change is measured in responses then it is assumed that no "real" change has occurred, rather than that there is the possibility of a radical reconceptualisation of the problem as with gamma change. For example, at the beginning of therapy an individual might rate the item "I find my work/school satisfying" as "Rarely" because they feel depressed. At the end of therapy, however, they may use the same rating because they now realise that the work they are doing is not for them and are looking for a new job. Hence this "no change" score hides a fundamental, positive outcome for the individual that does not get reflected in their response to the item. Especially in counselling and psychotherapy where gamma change may well be a desirable outcome, this would seem to be a fundamentally limiting aspect in terms of gaining meaningful outcome data.

5 | CONCLUSION

So where does all this leave us? If we fully concur with Michell's (1997, 2000), McLeod's (2001), and Slife's (2004) methodological critiques, then the whole endeavour is essentially flawed and meaningless beyond an "administrative reality." However, this administrative reality has a lot of "reality" for a lot of people and, as discussed in the sociopolitical critique, it has a lot of power. Do we then, as Elliot (2002) suggested, "render unto Caesar", and pay our tribute to the administration, so that we are able to continue to practice, or is rendering to Caesar a reduction, an act of collusion that merely reinforces the unhealthy power dynamics endemic in our society. Even when we have navigated this philosophical minefield, how can we possibly collect anything approaching meaningful outcome data given the numerous technical issues, problems, and flaws inherent in the endeavour? I believe that these questions are fundamental to the contemporary practice of psychotherapy. We can no longer choose to "opt out" and pretend that evidencing our practice is not necessary. Equally, we cannot ignore the many methodological, sociopolitical, and technical complexities and issues with this endeavour, and just "give them what they want" whilst pretending it has no impact on what we do.

Personally, I see two potential strands that may provide a way forward.

First, we can begin to claim the outcomes paradigm more as our own, and influence it so that it aligns more with our values, rather than "selling out" wholesale to the existing dominant paradigm, or having it forcibly imposed through financial pressures. To an extent, I see this shift as currently being attempted by the "practice-based evidence" movement which repositions the evidence base of therapy away from clinical trials and into regular practice (Barkham, Hardy, & Mellor-Clark, 2010). Here the integration of routine outcome measurement into everyday practice provides an opportunity for practitioners to evidence the efficacy of their work. Not only this, but the systematic collection and reporting of this data enables the results of everyday practice to be seen and heard by the "administration" in a way that we, as a profession, are more in control of. Here, online systems such as CORE-Net (http://www.coreims.co.uk) have the ability to coordinate the collection and reporting of therapeutic outcomes on a scale never before possible. Beyond this, there is potential to reposition the collection of outcomes data as another therapeutic tool, one which actually contributes to a client's process of recovery (Rodgers, 2015). Going down this path, outcome measures could take on a number of roles, such as enhancing the communication with their practitioner, providing opportunities to "check in" on progress, and facilitating client reflexivity. The role of practitioners and researchers in turn could shift from being detached "data collectors," to collaborative facilitators in the client's process of using such measures.

The other potential strand is to investigate and promote alternatives to the usual "standardised" measures. Again, some researchers and practitioners are already moving in this direction by utilising client-generated outcome measures such as PSYCHLOPS (Ashworth et al., 2004) and the Personal Questionnaire (Elliott et al., 2015). Rather than using a predefined list of items, these measures typically ask clients to identify what they want to work on in therapy in their own words. The value of this approach is that these questionnaires are still able to provide a normalised quantitative result but are more nuanced and responsive to the individual client change processes.

Extending this, the use of more qualitative methods for evaluating the outcomes of therapy would allow greater insight into the richness and diversity of these individual client change processes—and is both more personal and more political. Methods such as change interviews, diaries, timelines, and life space mapping all have the potential to be utilised as part of a routine data collection protocol (see Rodgers & Elliott, 2015, for a fuller discussion of these methods). Similarly, drawing on the rich tradition of using single cases to articulate theory, new approaches to analysing, reporting, and disseminating single cases aim to reposition systematic case studies as a legitimate evidence base for psychotherapy (see McLeod, 2010, for a fuller discussion of this). These approaches have the potential to mitigate some of the technical flaws inherent in standardised outcome measurement, as well as to influence the dominant paradigm to become more inclusive in what is seen as "evidence" in the evidence-based paradigm.

More broadly, I see the challenge here as one of how psychotherapy and aligned disciplines can avoid becoming marginalised and outcast in an industry dominated by the medical paradigm, yet not lose their soul in the process of staying in relationship within the dominante paradigm. My question and challenge for the reader is: Where do you stand?

REFERENCES

Ashworth, M., Shepherd, M., Christey, J., Matthews, V., Wright, K., Parmentier, H., … Godfrey, E. (2004). A client-centred psychometric instrument: The development of PSYCHLOPS. *Counselling and Psychotherapy Research*, *4*, 27–33.

Barkham, M., Hardy, G. E., & Mellor-Clark, J. (2010). *Developing and delivering practice-based evidence: A guide for the psychological therapies*. Chichester, UK: Wiley-Blackwell. https://doi.org/10.1002/9780470687994

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*(1), 7–18.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511524059

Danziger, K. (1997). *Naming the mind: How psychology found its language*. Thousand Oaks, CA: Sage.

Elliott, R. (2002). Render unto Caesar: Quantitative and qualitative knowing in research on humanistic therapies. *Person-Centered & Experiential Psychotherapies*, *1*(1), 102–117. https://doi.org/10.1080/14779757.2002.9688281

Elliott, R., Wagner, J., Sales, C., Rodgers, B., Alves, P., & Café, M. J. (2015). Psychometrics of the personal questionnaire: A client-generated outcome measure. *Psychological Assessment*, *28*(3), 263–278. https://doi.org/10.1037/pas0000174

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science*, *12*(2), 133–157.

Hughes, D. M. (1995). Significant differences: The construction of knowledge, objectivity, and dominance. *Women's Studies International Forum*, *18*(4), 395–406.

Luhrmann, T. M., Padmavati, R., Tharoor, H., & Osei, A. (2014). Differences in voice-hearing experiences of people with psychosis in the USA, India and Ghana: Interview-based study. *The British Journal of Psychiatry*, *206*(1), 41–44. https://doi.org/10.1192/bjp.bp.113.139048

McLeod, J. (2001). An administratively created reality: Some problems with the use of self-report questionnaire measures of adjustment in counselling/psychotherapy outcome research. *Counselling and Psychotherapy Research*, *1*(3), 215–226.

McLeod, J. (2010). *Case study research in counselling and psychotherapy*. London, UK: Sage. https://doi.org/10.4135/9781446287897

Meier, S. T. (1994). *The chronic crisis in psychological measurement*. New York, NY: Academic Press.

Meier, S. T. (2008). *Measuring change in counseling and psychotherapy*. New York, NY: Guilford Press.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355–383.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology*, *10*(5), 639–667.

Rodgers, B. (2015). *Outcomes informed practice—using outcome measures to promote client agency*. Presentation given at the Department of Psychotherapy and Counselling Forum, AUT, Auckland, New Zealand.

Rodgers, B., & Elliott, R. (2015). Qualitative methods in psychotherapy outcome research. In O. Gelo, A. Pritz, & B. Rieken (Eds.), *Psychotherapy research: Foundations, process and outcome*. Vienna, Austria: Springer-Verlag.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105.

Schwarz, N., Grayson, C. E., & Knäuper, B. (1998). Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research*, *10*(2), 177–183.

Semeonoff, B. (1976). *Projective techniques*. London: John Wiley & Sons.

Slife, B. D. (2004). Theoretical challenges to therapy practice and research: The constraint of naturalism. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed.) (pp. 44–83). New York, NY: John Wiley & Sons.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Tourangeau, R. (2003). Cognitive aspects of survey measurement and mismeasurement. *International Journal of Public Opinion Research*, *15*(1), 3–7.

Tudor, K. (2012). Southern psychotherapies. *Psychotherapy and Politics International*, *10*(2), 116–129. https://doi.org/10.1002/ppi.1265

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Brian Rodgers is a senior lecturer in the counselling programme at Auckland University of Technology. His primary research interest focuses on the outcomes of mental health interventions, in particular counselling and psychotherapy. Brian has produced numerous publications and presentations on this topic, including alternative approaches to outcome evaluation that allow us to discover what "comes out" of therapy from a client's perspective. He has a particular interest in how we can integrate authentic enquiry into both research and practice, such that both endeavours collaboratively combine to improve therapeutic outcomes and produce meaningful research results.

How to cite this article: Rodgers B. The trouble with numbers: Some fundamental flaws with using standardised outcome measures. *Psychother Politics Int*. 2017;15:e1423. https://doi.org/10.1002/ppi.1423