Questions, comments & responses

Questions came via the chat so are verbatim (but anonymised, each from a different participant). Answers are based on what we remember of our verbal answers but expanded and may have tweaked the answer a bit from that on the day.

Comment 1*:* If outcome measures and practice-based evidence are to be taken seriously at policy-making level there will need to be protocols around defining the quality of data gathered, how it's gathered and what sort of standards can be considered on some sort of parity with what has become the standard, the RCT. Is this really why practice based evidence is still considered 2nd rate?

Response*: On the day we mostly picked up the first sentence. We pretty much agree with that but RCTs perhaps have dominance because they have a great design simplicity: in the overwhelming proportion of RCTs one of two interventions is randomly allocated to each participant and the paradigm is the double blind RCT (DBRCT) in which it should be impossible for the participant or the practitioner (or, at the point of evaluations, any researcher involved in participant contact) to know which intervention was received. That's a wonderful paradigm for pharmaceutical research where double blindness is possible and it creates a very simple fit between design and statistical analysis and allows causal attribution. Doing an RCT is a huge amount of hard work but this simplicity means that it's fairly easy to agree general protocols and standards for a good RCT. This isn't so easy for PBE, first of all PBE doesn't have to be quantitative, in principle the term covers all research emerging from routine practice. However, let's assume we are talking about Routine Outcome Measurement (ROM) PBE: even here the sheer diversity of services and the absence of randomisation mean that there are many many different designs each of which have pros and cons for particular questions and services. Certainly, we need standards and protocols (as long as they don't try to turn PBE back into "laboratory model" research). For really influential PBE we standards about how to describe the service, how to describe client complexity (often excluded in RCTs) and above all standards for description of refusals and attrition.*

However, we suspect that the current absence of such standards is not why PBE "is still considered 2nd rate" but rather it is because PBE exposes two huge problems with 99% of therapy RCTs:

- *They can't be double blind so the entire logic of causal attribution which is sound with DBRCTs, is just not applicable for therapies. Sadly, NICE and the research funding bodies which overwhelmingly fund therapy RCTs over PBE want to keep the pretence of causal attribution and don't want it shown up by other approaches that don't pretend to strong causal attribution.*
- *The greater variety of PBE design and the issues that creates not only show up this logical deficits of therapy RCTs that cannot be double blind but also highlight their generalisability problems. This makes the entire business of grant allocation and the "hierarchy of evidence" uncomfortably complex.*

Sadly our field seems to lack the courage to say that the current focus in the EBP paradigm for psycho-social interventions is illogical and lacks generalisability.

We pick up these issues particularly in Chapter 10 but all that goes before that tries to expand on the complexities without becoming depressing: we encourage people to set up their data collection systems in intelligent ways that address the questions they are interested in and the audience they are speaking to. The introductory chapter 1 tries to set the epistemological frame. All four chapters of part II put flesh on the issues behind good PBE ROM data handling; chapters 6, 7 and 8 go into practicalities. Hm, perhaps only chapter 9 has relatively little to say about these issues!

***Comment 2**:* Something I'm digesting is the idea of "when not to measure". And somewhere in there might be an ontology of measurements. Could you maybe speak a bit on the stance of "when not to measure"?

***Response**: Great question! We confess to being a bit wary of philosophy: hugely attracted to its elegance and willingness to question while also very aware that we're not philosophers and a word like "ontology" is dangerous territory. In the book there are a number of points at which we note that the dominant self-report questionnaire ROMs designed for nomothetic comparisons across people really aren't like physical science measures. We recognise that no one physical science measure works for all things it might measure: you don't weigh minute quantities by putting them on lorry weighing bay nor would you put a lorry on a laboratory weighing scales, however most lorries can be weighed in a weighing bay. The inescapable reality of human uniqueness and the subjectivity of what we are trying to measure is that there will be a individuals who won't answer a general ROM as it might by another person, and so much so that comparing scores may give misleading information if fed into procedures assuming nomothetic comparability. Sometimes a measure might be fine for tracking change within one client but not for comparing that client's scores with those of other clients.*

There are also situations in which a measure might cause shame e.g. to clients with low literacy. This is never an issue with physical science and engineering measures: lorries don't have feelings, clients do. If a check on a lorry shows it is overloaded that may cause pain to the driver and owner (though it might save the lorry, driver and other vehicle occupants from danger) but neither the driver nor the lorry can alter what the scales says the weight is. Someone feeling shamed by their struggle with a measure might answer randomly or just refuse to complete the measure. Humans are not simple and can change their answers as they speculate about how the scores might affect them: not something lorries can do.

Another point relates to where reliability and validity were established for the questionnaire. In many papers, the fact that they were established in a particular group appears to be seen as generalisable to all groups. This is not only illogical it is bad science/research. A questionnaire on wellbeing, validated in a group of University students is unlikely to transfer to a group of older adults in supported accommodation. A questionnaire on interpersonal functioning validated in a "general" population will not transfer to a forensic one. Part of the reason this is ignored, is that it requires a whole lot of work to establish appropriate validity in different groups, so it is easier to turn a blind eye to the problems.

Another related issue here is that of translations of measures. We do go into this in the book, but good translations require a good protocol with local focus groups picking up cultural issues as well as local testing. There is still a colonialist mentality that simply putting a measure through a forward and backward translation is sufficient for it to be used in translation. Again, this is bad science. So, if there is not a translation that has been developed with a thorough protocol, then it would be a time not to measure (Evans et al., 1997; Evans et al., 2021; Paz et al., 2021; Rogers et al., 2013; Yassin & Evans, 2021 – ouch: I have been at this issue for a long time!).

There are also some very real issues that don't seem to have been discussed much, about how to use measures in group therapies (and family/systemic therapies), see comments 3 and 4.

These are practicalities and are not arguments against using ROMs generally, just about recognising that although most are designed for nomothetic comparability this doesn't mean they have it for everyone. This is exacerbated because our currently prevalent psychometric tools aren't very good for detecting these issues precisely because reasonable measures do give fair comparability for a majority, but not all, people completing them.

That is all picking up "when not to measure" more at a pragmatic level than a fully philosophical one. The ontology, or perhaps ontologies, of ROMs could fill a much larger book and we'd love to see philosophy departments, and anthropology and sociology departments picking this up. For us John McLeod's paper "An administratively created reality: Some problems with the use of self-report questionnaire measures of adjustment in counselling/psychotherapy outcome research" (McLeod, 2001) remains vital reading for anyone interested in this. The large literature, expanding rapidly, led by figures like Michel or Trendler pick up questions that were certainly recognised by Cronbach, Meehl and even Cattell well back in the last century but generally seem to us to throw out the proverbial ROM baby with the bathwater.

***Comment 3**:* In the "We need new models!" group, we were discussing about the need for epistemological changes. Psychological science, most psychological treatments and outcome measures have been historically individual-based. The development of more social-based approaches and outcome measures should be encouraged. Otherwise, many contextual, economical, social factors may be overlooked. Besides it seems that it is convenient for many people in power to approach psychological treatments and research in an individualistic, individualising utilitarian manner.

***Response**: Crikey. We didn't get to this on the day and it must have languished in the chat. Apologies.*

Where to start. It's extremely hard to create a nomothetic self-report questionnaire measure without being forced you into an individualistic frame of mind. It's hard enough to write intrapersonally focused questions that might work for most people. Moving to interpersonal focus questions is far harder: you have to try to write the question so it might work for people who have (none, one or more) intimate personal relationships, for those with a formal occupation and for those who may not have such an occupation. This doesn't preclude analysing such data to explore contextual, economical, social factors but sadly this is rare in our field. One theme through the book is to encourage data analyses, where the data permit it, to look at such factors and this is explicit in Chapter 8 on "Service-level change and outcome measurement". To a worrying extent current ROM work is trapped in a very individualistic model as the breakout group clearly recognised and we would agree that this is worryingly aligned with political and social trends to managerialism, commoditisation and dehumanisation.

One running theme for Jo-anne and I, and our friend and colleague in Ecuador, Clara Paz, is about how we might have measures that appreciate individual identity and experience as only ever nodal within interpersonal, social, sociological and multi-generational historical contexts. That never really makes it to the face of the narrative in the book but we are very clear that there is qualitative change measurement no matter how much "qualitative measurement" currently sounds like an oxymoron.

There is some work, particularly in Latin America, in a "community psychology" tradition, drawing to varying degrees on drama therapy, psychodrama and perhaps the creative therapies more generally. That work starts with the notion of the individual as merely, though potentially wonderfully, nodal within networks, not as a monadic, autistic (not in the disease/diagnosis sense) individual. Some work has looked at using group story making, image making and construction of theatre, enactment as "measuring" the state of systems and the individuals within those systems and that seems to us vital work, and links with Jo-anne's increasing interest in applications and extensions of group relations work and humanitarian rather than individualist interventions. Sadly, these are not areas we know well from a research perspective and it's currently very hard to bridge from that to our Global North paradigms (hence, of course, part of the reason this is not a dominant discourse in the book). Chris's "rigorous idiography" ideas, e.g. "Significance testing the validity of ideographic methods: a little derangement goes a long way" (Evans, Hughes & Houston, 2002) start to explore some options to stop treating the quantitative/qualitative distinction as if it were a simple dimension and rather than a topological space. But that's probably for the next generation to start to open up!

***Comment 4**: (oral in the session)*: One challenge for SCORE-15 (Stratton et al., 2013) has been that (family/systemic) therapists and teams get families involved in discussing the measure in the session, weaving it into the session discourse and so surely invalidating it as an overall change outcome measure.

***Response**: Yes! This is a theme in the book: you shouldn't ask measures to do two clearly psychosocially distinct tasks. By all means weave scores into the session and use them to shape the conversations. This is what we call ECM: embedded change measurement, also called FIT: Feedback Informed Therapy. We prefer "ECM" as all therapy, always, has been feedback informed: one's responses to a client are shaped, feedback informed, not only by overt verbal content but non-verbal feedback, hearing a new quiver in the voice, seeing new facial expressions, seeing sweat, perhaps worsening or diminishing of bodily odour even: qualitative change measurement is a foundation of all therapies, not something new. Using the same measure for ECM and for overall therapy change assessment is extremely unwise because of the response pressures put on the embedded measure. One recent evaluation of an ECM intervention that used the ORS (Outcome Rating Scale) within sessions broke with this methodologically dubious tradition of using the same measure as an ECM and for evaluation of the ECM intervention and used the OQ-45 for the group change comparisons (Bovendeerd et al., 2021). We pick this up in various ways in chapters 2, 3, 4 and 9.*

References

Bovendeerd, B., de Jong, K., de Groot, E., Moerbeek, M., & de Keijser, J. (2021). Enhancing the effect of psychotherapy through systematic client feedback in outpatient mental healthcare: A cluster randomized trial. Psychotherapy Research, 1–13. https://doi.org/10.1080/10503307.2021.2015637.

Evans, C., Hughes, J., & Houston, J. (2002). Significance testing the validity of ideographic methods: A little derangement goes a long way. British Journal of Mathematical and Statistical Psychology, 55(2), 385–390. https://doi.org/10.1348/000711002760554525

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. https://doi.org/10.1080/14733140112331385100.

Stratton, P., Lask, J., Bland, J., Nowotny, E., Evans, C., Singh, R., Janes, E., & Peppiatt, A. (2013). Detecting therapeutic improvement early in therapy: Validation of the SCORE-15 index of family functioning and change: Validation of the SCORE-15 index. Journal of Family Therapy, 36(1), 3–19. https://doi.org/10.1111/1467-6427.12022.

Evans, C., Dolan, B., & Toriola, A. (1997). Detection of intra- and cross-cultural non-equivalence by simple methods in cross-cultural research: Evidence from a study of eating attitudes in Nigeria and Britain. *Eating and Weight Disorders*, *2*, 67–78. https://doi.org/10.1007/BF03397154.

Evans, C., Paz, C., & Mascialino, G. (2021). "Infeliz" or "Triste": A Paradigm for Mixed Methods Exploration of Outcome Measures Adaptation Across Language Variants. *Frontiers in Psychology*, *12*, 695893. https://doi.org/10.3389/fpsyg.2021.695893.

Paz, C., Hermosa-Bosano, C., & Evans, C. (2021). What Happens When Individuals Answer Questionnaires in Two Different Languages. *Frontiers in Psychology*, *12*, 688397. https://doi.org/10.3389/fpsyg.2021.688397.

Rogers, K. D., Young, A., Lovell, K., & Evans, C. (2013). The challenges of translating the clinical outcomes in routine evaluation-outcome measure (CORE-OM) into British Sign Language. *Journal of Deaf Studies and Deaf Education*, *18*(3), 287–298. Scopus. https://doi.org/10.1093/deafed/ent002.

Yassin, S., & Evans, C. (2021). A journey to improve Arabic-speaking young peoples' access to psychological assessment tools: It's not just Google translate! *Counselling and Psychotherapy Research*, capr.12431. https://doi.org/10.1002/capr.12431.

Useful links/resources

Glossary: https://ombook.psyctc.org/glossary/
General, developing, materials supporting the book: https://ombook.psyctc.org/book/
CORE site: https://www.coresystemtrust.org.uk/
Chris's non-CORE work (mostly but not all change measurement): https://www.psyctc.org/psyctc/
Chris's "R blog" about using R, mostly for ROM): https://www.psyctc.org/Rblog/
Chris & Clara's CECPfuns R package (more geeky still): https://cecpfuns.psyctc.org/
SCORE: https://www.aft.org.uk/page/score
PSYCHLOPS (example of a hybrid measure): http://www.psychlops.org.uk/
SPR UK Leeds conference, April 8-9 2022. https://spr-uk.wixsite.com/conference2022
Goal based outcomes tool: https://goalsintherapycom.files.wordpress.com/2018/03/gbo-version-2-march-2018-final.pdf