Patient-centered outcome measurement in psychiatry: How metrology can optimize health services and outcomes

Skye Pamela Barbic
skye.barbic@ubc.ca
University of British Columbia

Stefan J Cano
stefan.cano@modusoutcomes.com
Modus Outcomes

Karen Tee
ktee@foundrybc.ca
Foundry

Steve Mathias
smathias@foundrybc.ca
Foundry

Abstract:

In psychiatry, there is a call for clinicians to use patient centred outcome data routinely at the point of care to help tailor treatment plans to meet the patient's preference and needs. Given that many decisions in psychiatry are constructed from patients' narratives, it is critical that the conceptual, empirical, and measurement structure underlying patient reported outcome measures is robust and patient-centred. Here, we argue for the systematic accruement of patient-centred data in psychiatry to meaningfully enhance the treatment of mental disorders. Specifically, we suggest three crucial considerations for system transformation: (1) the engagement of international patient research partners to conceptualize and prioritize outcomes; (2) the application of modern test theory to develop and evaluate patient-centred outcome measures: and (3) funding allocation accountable to evidence-based services prioritized by patients.

Keywords: Measurement, Metrology, Psychiatry, Rasch

1. Introduction

The treatment of mental disorders is a global health challenge, costing one trillion dollars each year (Chisholm et al, 2016). One in four people will experience a mental disorder at some point in their lives and 450 million people currently suffer from mental health challenges (Peter at al, 2016). As a result, mental illness is the leading cause of ill-health and disability worldwide (WHO, 2013).

Despite established data supporting the impact and prevalence of mental disorders, there remains a colossal treatment gap globally (Peter at al, 2016). It is estimated that 75 % of individuals never receive care, let alone the appropriate care (Meffert et al, 2016). Common metrics of system and personal impact include emergency department utilization, suicide rates, hospital recidivism, and laboratory markers (Walket et al, 2015). Increasingly, it has been suggested that these outcomes alone may not capture the full potential and evolution of psychiatric services (Barbic et al, 2016).

As more people with mental disorders receive care in the community, reintegrating people into productive lives in society is a widespread overarching aim of psychiatric services. Here, we argue for the systematic accruement of patient-centered data in psychiatry to meaningfully enhance the treatment of mental disorders. Specifically, we suggest three crucial considerations for system transformation: (i) the engagement of international patient and family research partners to conceptualize and prioritize outcomes; (ii) the application of modern test theory to develop and evaluate patient centered outcome measures; and (iii) funding allocation that is accountable to evidence-based services prioritized by patients.

2. Engaging patient partners in outcome conceptualization and prioritization in mental health

Patient-oriented research and clinical practice focuses on patient-identified priorities and aligning clinical care accordingly (CIHR, 2016). This “movement” was develop to enhance researcher and clinician understanding of the effects of disease and treatment on patients’ daily lives. Rheumatology and oncology are examples of health fields that have emphasized patient-partners in research in the last decade to improve endpoint outcome measurement (Kirwan et al, 2017). Through their iterative, inclusive consensus processes, outcomes such as participation, function, and fatigue are endorsed targets for treatment and clinical trials in both fields. Systematic approaches to defining patient priorities in psychiatry have yet to be developed. However, psychiatry is not alone. In an editorial published in the British Medical Journal, Angela Coulter (2017) noted that the behavior of measuring what matters to patients is “surprisingly rare”, with a mere 11 % of patient reported outcome measures actually asking patients which outcomes are worth measuring (Coulter, 2017).

In response to this trend, national initiatives, such as Canada’s Strategy for Patient Oriented Research (SPOR) and the Patient Centered Outcome Research Institute (PCORI) in the United States, have been developed to fund research with the intention to help patients feel empowered to make informed decisions about their healthcare choices. As demand increases for accountability of mental health services to be person and family-centered, engagement of patients at all levels of research design are necessary. However, patient engagement in isolation may not be the solution. Unfortunately, many of the outcomes in psychiatry are not directly observable and present systematic challenges for capturing them and characterizing treatment effectiveness. If treatment selection, clinical trials, and/or government policies are to emphasize patient-centered outcomes, a robust approach to measuring them is also needed.

3. The application of modern measurement methods to develop, test, and use patient centered outcome measures in clinical practice

Clinicians frequently use single items (i.e., the Global Assessment of Functioning - GAF) to assess concepts of interest (e.g., function, mood, quality of life) as they appear easy to use. However, single item 'scales' have been repeatedly been shown to be a poor substitute for comprehensive psychological and functional testing (Ramirez et al, 2008). One reason for this is the assumption that a single item 'scale' behaves like a perfect interval ruler, ambitiously claiming to represent a wide concept with one item. Single item 'scales' are actually scientifically limited because they have poor reliability, validity, and responsiveness (Hobart et al, 2007). The limited robustness of this type of scale is also vulnerable to the variability of patient and clinician interpretation. Thus, unlike a ruler that measures length, a single item 'scale' used in psychiatry may cover many different frames of reference that a patient may have. Culture, language, mood, age, gender can all challenge the measurement properties of single item 'scales' and can make comparisons between people, centers, or over time, difficult.

The limitations of single item 'scales' have led to the more frequent use of multiple item scales or questionnaires, in which ratings assigned to each item are combined to give a total score. Combining multiple items to provide a more comprehensive health profile of a patient may seem clinically intuitive. Statistically, this may also be sensible, as evidence supports that combining multiple items reduces random error and enhances scale reliability (Hobart et al, 2007). However, in the case of measurement, statistical significance does not necessarily translate into clinical meaning (and vice versa). Without a robust measurement model underpinning the scoring of multiple item scales, a total score becomes difficult to interpret. This is further complicated because there are many clinical tools that purport to measure the same construct. This makes operationalizing clinical practice and quality improvement impractical. As multiple-item Patient-Centered Outcome (PCO) instruments become increasingly in demand in psychiatry, clinicians and patients are encouraged to learn from advances in measurement to interpret these scales to improve disease management and health outcomes for people with mental disorders (Barbic et al, 2018&2016).

Another area of consideration for psychiatry is the amount of faith invested in the "reliability and validity" of a PCO instrument. Just as significant advances have been made in psychiatric practices (Kidd et al, 2014), notable advances have also been made in the field of measurement. In order to advance the health outcomes of people receiving psychiatric services, it is critical that psychiatry optimizes how measurement advances can improve care and the patient experience (Barbic et al, 2018). New methods in measurement not only provide an opportunity for increased precision in medicine, but an opportunity to use measurement to build a language for service design and provision (Cano et al, 2018). For example, what does a score of 15 (on a range of 0-60) mean in an instrument purporting to measure depression? How does that same score of 15 compare to the total scores generated by dozens of other existing instruments purporting to measure depression?

Interpretation of PCO data has been a longstanding challenge in psychiatry. Communicating results to patients, families, and other health professionals can often be difficult and abstract. This limits the potential to systematically communicate information about how to optimize the recovery and health of people with mental disorders (Barbic et al, 2018). As noted in a recent New England Journal of Medicine Blog, "One of the biggest challenges today is how to design a system that meaningfully engages patients and clinicians for better health" (NEJM, 2017). We argue that engagement cannot occur if key stakeholders do not speak the same language or share the same targets for care.

Fortunately, psychometric approaches, such as Rasch Measurement Theory (RMT) (Rasch, 1980), exist as potential solutions to this problem. RMT has unique properties that are invaluable to fields where outcomes are mostly latent and inferred from the patient (Andrix 2005&1988). These include references for traceability and means of evaluating measurement uncertainty for both patients and items (Pendrill 2014). As well, RMT methods embrace the opportunity for key stakeholder to work together to hypothesize the composition of concepts of interest as linear measures.

RMT methods allow researchers to study the extent to which items form adequately conformable sets to represent clinical concept of interest and map out a clinical hierarchy, to the extent to which specific items can be located to characterize specific areas in a measurement continuum. These then supports hypothesis testing and construct theory development of the clinical concept of interest under study.

RMT methods also allow researchers to study whether the distribution of item locations (or estimates) are independent of the sample distribution of respondents. This is critical for creating measurement tools that are robust for clinical centers, populations, and response shift (Pendrill 2014). The implications for this are profound for all areas of health, notably psychiatry—where most outcomes are subjective, and patients are required to report progress for a long-term treatment period. The application to metrological standards to psychiatry has the potential to foster patient-centered and evidence informed health, to ensure greater quality, accountability, and accessibility of meaningful care (Cano et al, 2018).

4. Patient-centered outcome service delivery and funding allocation

The routine use of PCOs in psychiatry provides an opportunity to help drive how health care is organized, delivered, and funded (Ahmed et al, 2012; Kirwan et al., 2014). As noted above, metrology provides a framework for health sciences to develop a common language between patients, clinicians, researchers and key stakeholders. This form of communication in psychiatry can be used to compare performance metrics of clinical services to evidence-based guidelines or standards. Through benchmarking, there is the potential to learn how well psychiatric interventions perform and align to outcomes that are meaningful to the patient.

Benchmark assessments have been used rigorously by other areas of social science such as Education to monitor progress of stakeholders. Under the assumption that outcomes are derived from rigorous patient-centred conceptual and measurement models, benchmarking can inform the allocation of funding towards the areas of greatest need. This ensures decision makers and policy makers are equipped with the evidence needed to develop plans on how to make improvements or adapt best practices to optimize the health and wellbeing of patients. In this paper, we provide a clinical example of how one Canadian province is applying modern test theory for benchmarking to reconceptualise and fund how services for youth and young adults with mental illness are delivered and evaluated.

5. Example of designing mental health systems with the principles of metrology: Foundry

In 2015, an initiative was approved in Western Canada to develop a new model of care for youth and young adults with mental illness. The purpose of the initiative was to provide "one-stop" services for young people to help them receive the timely and youth-centred support that they need to thrive (Mathias S, 2018). Foundry currently has 120 partnerships across the province of BC and eleven centres that are open or scheduled to open soon. Foundry represents community agencies, government, donors, youth and young adults, and families coming together to improve the wellness of young people in British Columbia. All Foundry centres are branded and present a consistent feel, regardless of geographical location. An example of the two centres and branding is shown in Figures 1a and 1b.

One key ingredient to the success of these centres is consistency of services. It is expected that any young person can walk into any Foundry and expect to find a core set of services that can be offered and a reliable approach to assessment, treatment, and follow-up. To maximize this consistency, our team identified that a youth-centred systematic method of collecting youth-centred data was needed. In response, "Toolbox" was introduced to Foundry.

Figures 1a and 1b. Examples of waiting rooms of two Foundry centers (Kelowna Left, North Vancouver Right), where health data is collected.

5.1 Toolbox

Toolbox© is a data collection platform used at all Foundry centres. The purpose of Toolbox© is to collect youth-centred data to inform service delivery and quality improvement. Toolbox© is an electronic data capture and management system that can capture data from any device (tablet, cell phone, computer, etc...). Toolbox© allows for rapid collection of data to study the extent to which the patient-reported outcome measures are fit for purpose to measure the needs of youth. Toolbox© allows for control over data collection and query management for patient-reported outcomes. Foundry has a dedicated team focused on Toolbox© and its application to each Foundry centre. The team is responsible for ensuring that the data system collects information that is clinically relevant, youth-centred, and robust to measure the needs and priorities of young people across British Columbia. One area of focus for the team is ensuring real-time testing of patient-reported outcome measures (PROMs).

5.1.1. Hypothesis-driven, patient-centered outcome measurement

Well before a PCO instrument is used in Toolbox©, extensive expert (i.e., clinician, youth, family, policy-maker, and psychometrician) feedback is obtained about the candidate measure. First, relevant concepts of interest are identified. Next, a literature review is conducted to identify existing measures to capture these concepts. Once a bank of measures is identified, the content is reviewed by the expert team. The measures endorsed by the team are then subjected to pilot testing at one Foundry centre.

For example, in 2017, Kessler Psychological Distress Scale (K10) (Kessler et al, 2002) was identified as a candidate measure for Foundry to capture youth distress. The desired purpose of the K10 for the team was to evaluate change in distress for all youth accessing Foundry centres. To inform this decision, the K10 was given to 350 young people receiving Foundry services and tested for its psychometric properties. Before we initiated testing, we asked the expert group to hypothesize the ordering of the items (from lowest distress to highest distress) and suggest any anomalies in the data they would expect. Methods guided by RMT were used to test these hypotheses and test the scale properties. Specifically, we tested the ordering of response option thresholds, fit, spread of the item locations, residual correlations, person separation index (PSI), and stability across time.

Tables 1 and 2 and Figure 2 show a summary of the psychometric properties of the K10 for the test population.

Table 1: Traditional psychometric methods-data quality, scaling assumptions, targeting, reliability (n=350).

Psychometric property	Total
Data Quality	
Missing data (%)	0.0
Computable scale scores	350
Scale assumptions	
Item mean scores: mean (range)	2.94 (2.59-3.20)
Item SD: range	1.16-1.36
Targeting	
Mean score (SD)	25.62 (9.67)
Possible score range*	10-50
Observed score range	10-45
Floor/ceiling effect**	<1/<1
Reliability	
Cronbach's alpha	0.90
Mean inter-item correlation	0.01-0.45

* Higher scores indicate greater distress

** Floor effect= % scoring 10 (lowest distress); ceiling effect= % scoring 50 (highest distress).

Table 2: Measures of fit and location (SE) of K10 items. Items are located in order of low (item 1) to higher distress (item 3).

Item	Descriptor	Location [b]	SE	Fit [c]	ChiSq [a d]	Prob
1	feel tired out	-0.505	0.078	3.540	7.98	0.157
8	everything was an effort	-0.439	0.074	-0.397	5.77	0.329
7	feel depressed	-0.357	0.076	-0.735	4.68	0.456
2	feel nervous	-0.355	0.081	0.297	7.43	0.191
5	feel restless or fidgety	-0.228	0.076	1.746	11.34	0.045
4	feel hopeless	0.039	0.077	-2.085	5.30	0.380
10	feel worthless	0.141	0.071	-0.702	4.44	0.488
6	could not sit still	0.491	0.077	1.072	8.82	0.116
9	nothing could cheer up	0.527	0.078	-0.846	13.79	0.017
3	nothing could calm you	0.687	0.079	-0.770	3.71	0.592

[a] Degrees of Freedom (5,248)

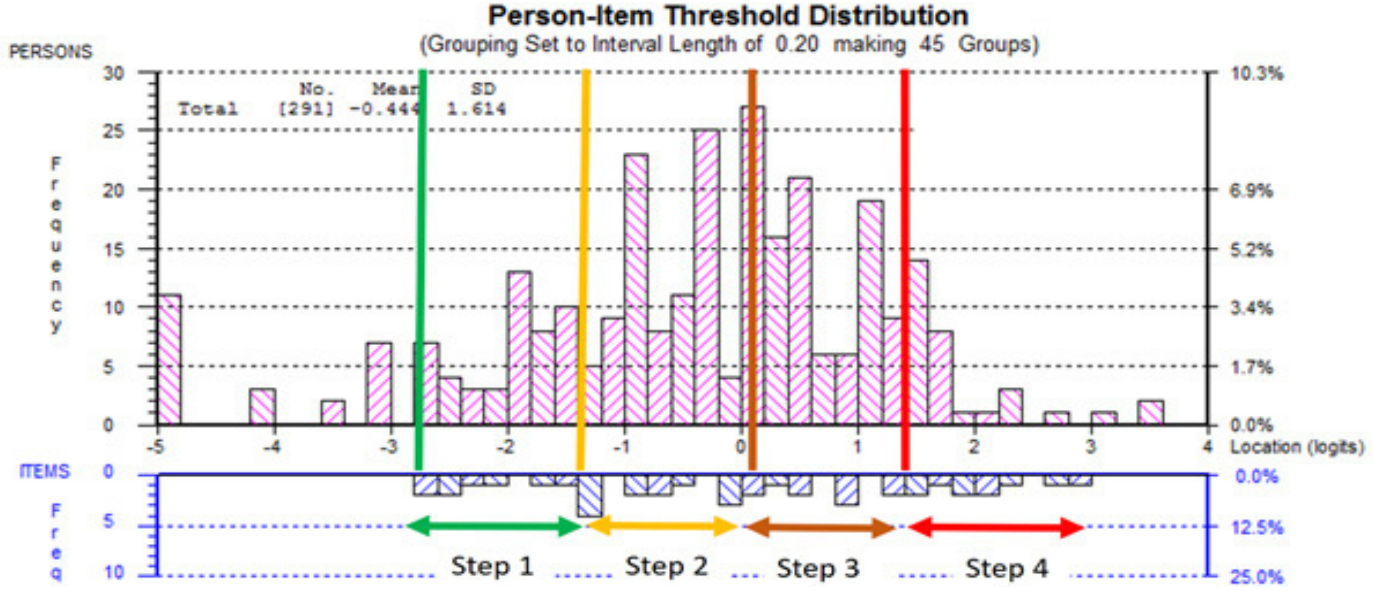
[b] Mean location score obtained for items. In RUMM 2030, the scale is centred on zero logits, representing the item of average difficulty for the scale.

[c] Residual statistics are the standardized sum of all differences between observed and expected values scored over all persons. An item with a residual statistic less or greater than 2.5 warrants further investigation.

[d] Chi-square statistic compares the difference of observed values with expected values across groups with different ability levels (from less to high distress) across the latent trait being studied (distress).

Table 1 specifically outlines the traditional psychometric properties of the tool. Using this method, our team examines the quality of the data, scaling assumptions, targeting (do the items target the population), and reliability (as measure by Cronbach's alpha and inter-item correlations). The information is used to understand the performance of the individual items and how they work collectively to cover the full range of the possible raw scoring options.

Figure 2: Person-item threshold distribution. The distribution of K10 scores for youth are shown by the pink bars- with a range from low (left side) to high distress (right side). An ideal theoretical distribution is from -3 to +3 logits, with a mean of 0 and SD of 1.0. Here we have a mean of -0.44 and SD of 1.6. The blue bars represent the distribution of the item thresholds (-2.7 to +3.0 logits). In an ideal measure the item distribution targets the person distribution. In this example, the targeting of the items to the people is excellent. The cut-off scores of the K10 have been used to partially inform the level of care a person receives based on the cut-off score (Step 1-Step 4). The cut-off scores can be tested regularly to ensure that the level of services is appropriate for the level of distress reported by the patient.



However, we require scales that has interval properties and can be added up to make a total score and one limitation of using traditional psychometric methods is that they are unable to provide the information we need to evaluate this property. Therefore, we used RMT methods to inform the extent to which the items are invariant [a property critical for comparing (1a) a patient over time, (b) patients over time, and (3) different Foundry centres] and were operationalize in the testing context as described by the Expert group (were ordered as expected). As outlined by Rasch himself, "the comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison, and it should be independent of which other stimuli within the considered class were or might also have been compared"(Rasch, 1961: 332). For distress, and other constructs under investigation at Foundry, it is critical that all tools have psychometric properties that are not sample dependent. With the expectation that the number of Foundry centres will grow in number over time, we anticipate that this approach will be critical for regular program evaluation and future PROM development and testing.

5.1.2. Clinically meaningful application of patient-centered outcome measurement

Another area of testing is the clinical utility of the PCO instrument itself. Specifically, what is the extent to which the total score can inform clinical care? As with the physical sciences, our team feels that a total score should produce a meaningful value that is understood by the patient, the clinician, family members and other stakeholders. For example, using methods guided by RMT to understand the underlying measurement model of the K10, we can understand the item hierarchy of the K10 and how items line up consecutively to map a story of the construct. As shown in Table 2 and Figure 2, transferring the location score into a raw score can allow a patient to understand how to interpret the total score. The total score can also become a tangible decision-making tool to map clinical services.

Figures 2 depicts an example for how cut off scores from the K10 are used at Foundry to inform the intensity of evidence-based services. This provides a clear communication tool for patients and stakeholders. A patient can understand, based on their total score, what level of service to expect. This is independent of the centre. This is particularly helpful as well for communicating treatment approaches to other stakeholders involved in the youth's health treatment plan including family members and teachers. The care pathways ensure that a full menu of services is clear to patients and can constantly be evaluated in real time. Foundry is in the early phases of service delivery. However, we anticipate that this approach to program evaluation will allow for benchmarking and the allocation of funding and resources towards innovation and evidence-based services that show clinically meaningful change in a patient.

The rigorous approach to measurement allows ongoing opportunity to identify and understand practices, methods, and processes at Foundry. As an organization, this allows the potential to customize practices to individual youth and each Foundry centre. We also anticipate that benchmarking, based on meaningful measurement, will allow for more efficient use of resources through the identification and implementation of best practices modified to local Foundry centres and communities. Benchmarking will allow our team to focus on comparing internal processes and innovation and compare those from other initiatives in Canada and beyond. As such, data are used in a clinically meaningful way as a communication mechanism for continued improvement and learning, as well as networking with like-minded organizations to optimize the health and well-being of young adults with psychiatry disorders.

6. Conclusion

Patient-centred outcome measurement in psychiatry is essential for advancing outcomes and delivering high quality services to patients and their families. The capacity to utilize metrological standards in psychiatry should be a fundamental aspect of health service design, delivery, and evaluation in psychiatry and mental health. To do so, partnerships between clinical leaders, patients, and measurement experts are essential. Patients receiving psychiatric services have the right to the highest quality of care and to understand how their health outcomes translate to evidence-based services. Entry points to action include a formal linkage with the metrology and mental health communities. Strengthening the measurement-health linkage will improve international quality mental health services, interventions, and patient experiences.

References

Chisholm, D et al., (2016), Scaling-up treatment of depression and anxiety: a global return on investment analysis. *Lancet Psychiatry* **3**, 415-24.

WHO-World Health Organization, (2013), *Mental Health Action Plan 2013-2030* .Geneva, Switzerland.

Patel, V et al., (2016), Addressing the burden of mental, neurological, and substance use disorders: key messages from Disease Control Priorities. *Lancet* **387**, 1672-1685.

Meffert, SM et al., (2016), Novel implementation research designs for scaling up global mental health care: overcoming translational challenges to address the world's leading cause of disability .*Int J Ment Health Syst* 10 19.

Walker, ER et al., (2015), Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. *JAMA Psych* **72**, 334-341.

Barbic, SP et al., (2016), The application of Rasch measurement theory to psychiatric clinical outcomes research: Commentary on Screening for depression in primary care. *BJPsych Bull* **40**, 243-244.

CIHR-Canadian Institutes for Health Research, (2016). *Strategy for Patient-Oriented Research.* Canada.

Kirwan, J R et al., (2017). Emerging Guidelines for Patient Engagement in Research. *Value in Health* **20**, 481-486.

Coulter, A (2017), Measuring what matters to patients. BMJ 356, j816.

Ramirez, A et al., (2008). Global Assessment of Functioning scale (GAF), further evaluation of the self-report version. *Eur Psych* **23**, 575-579.

Hobart, J et al., (2007), Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neuro* **6**, 1094-1105.

Barbic, SP et al., (2018), Patient-centred outcome metrology for healthcare decision-making. *J. Phys. Conf. Series* **1044**, 1-6.

Kidd, SA et al., (2014), Mental Health Reform at a Systems Level: Widening the Lens on Recovery-Oriented Care. Can *J Psych* **59**, 243-249.

Cano, SJ et al., (2018), Patient-centred outcome metrology for healthcare decision making. *J. Phys. Conf. Series* **1044**.

Catalyst, NEJM (2017), Hardwiring Patient Engagement to Deliver Better Health (https://join.catalyst.nejm.org/events/archived-hardwiring-pe-2017/register/on-demand).

Rasch, G (1980), *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Demark.

Andrich, D (1988), *Rasch models for measurement.* Beverley Hills, United States.

Andrich, D (2005), *Stat Behav Sci* **4** 1698-1707.

Pendrill, L (2014), Man as a measurement instrument.t *J Meas Sci* **9**, 24-35.

Ahmed, S et al., (2012). The Use of Patient-reported Outcomes (PRO) Within Comparative Effectiveness Research: Implications for Clinical Practice and Health Care Policy. *Med Care* **50**, 1060-1070.

Kirwan, J et al., (2014), Updating the OMERACT filter: implications for patient-reported outcomes. *J Rheumatol* **41**, 1011-5.

Mathias, S (2018), Foundry: Where wellness matters. URL: www.foundrybc.ca.

Kessler, RC et al., (2002), Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychol Med* **32**, 959-976.

Rasch, G (1961), *On general laws and the meaning of measurement in psychology.* Berkeley, USA.

Authors Profiles:

Skye Barbic has received a Ph.D. from McGill University- Montreal, Canada in 2013. She is now an Assistant Professor in the Department of Occupational Therapy at the University of British Columbia. – Vancouver, Canada and the Research Lead for Foundry. Her research interests are in the areas of Health Measurement, Youth Mental Health, and Rasch Measurement Theory.

Stefan J. Cano is a Chartered Psychologist and Associate Fellow of the British Psychological Society. He 24 years' experience in patient centered outcome measure development and psychometric research. Having previously worked extensively in qualitative and quantitative research projects, Stefan's main area of interest is in developing, applying, and improving mixed methods psychometric research in clinical studies and therapeutic trials.

Karen Tee received a Ph.D. from Simon Fraser University, Burnaby, Canada. She is a Clinical Psychologist and the Director of Service Innovation for Foundry in Vancouver, Canada, and has experience in youth mental health both in clinical operations and direct service delivery for the past 25 years.

Steve Mathias received an MD from the University of British Columbia – Vancouver, Canada. He is the Executive Director for Foundry (foundrybc.ca) in Vancouver, Canada and the Head of the Department of Psychiatry at St. Paul's Hospital. His research interests are Integrated Health, Youth Mental Health, and At-Risk Youth.