MAIMON WORKING PAPERS www.maimonresearch.net

WORKING PAPER No. 11 17 MAY 2020

MANIFEST NONSENSE: COSMIN AND THE AXIOMS OF FUNDAMENTAL MEASUREMENT IN THE CHECKLIST FOR SYSTEMATIC REVIEWS

Paul C Langley Ph.D., Adjunct Professor, College of Pharmacy, University of Minnesota

Abstract

The COSMIN checklist for systematic reviews of patient-reported outcome measures (PROMs) represents an attempt to utilize classical test theory criteria as the basis for grading selected outcome measures. The endeavor, although widely reported over the past 10 years, fails on several counts. First, it fails completely to appreciate the need to meet the axioms of fundamental measurement in instrument development. Second, it fails to appreciate the contribution of Rasch Measurement Theory (RMT) to instrument development. Third, it fails to appreciate the need to follow the physical sciences in their focus on single attribute instrument development. The net result is that reviewers produce positive evaluations for PROMs that are of questionable quality.

Introduction

The purpose of this commentary is to consider why the *COnsensus-based Standards for the selection of health status Measurement INstruments* (COSMIN) checklist appears to have avoided any consideration of the axioms of fundamental measurement when evaluating patient reported outcome measures (PROMs) [1] [2]. This is a major oversight. Neglect of the axioms of fundamental measurement has been shown in previous commentaries in *INNOVATIONS in Pharmacy* to have led to the false acceptance of quality adjusted life years (QALYs) as fundamental to modeled claims for cost effectiveness [3]. The concern, therefore, is that a similar fate will befall the hundreds of PROMs developed over the past 40 years by academic groups and clinicians who fail to appreciate the limitations placed on instrument development by these axioms. The result being that they have limited quality and application, with invalid claims for responsiveness to therapy interventions as in the case, reviewed below, of the ASCQ – Me in cystic fibrosis [4] . To resolve this issue we must propose, indeed insist, on Rasch Measurement Theory (RMT) as the recommended framework for developing PROMs that meet the requirement of fundamental measurement [5]

Responsiveness

Responsiveness is defined under the COSMIN checklist as indicative of longitudinal validity; the ability of a PROM to detect change over time in the construct to be measured. Change can only be evaluated if we follow the axioms of fundamental measurement with an instrument that has interval properties and, if possible, to construct an instrument with ratio properties [6] [7] [8]. PROMs that fail the standards of fundamental measurement, i.e. those with ordinal or manifest scores, cannot measure change over time. They can only report change in response level without any knowledge of the distance between the responses. This may satisfy clinicians and other health professionals where rough changes in scores can be part of a clinical assessment, but not if we are to provide a measure of response to the same standards as the physical sciences.

In other words, unless we can demonstrate that the instrument has ratio or, less ambitiously, interval properties then any claims for measured responsiveness must be rejected. Instruments which fail these criteria must be put aside. Curiously, this is a requirement that COSMIN ignores. Indeed, the entire COSMIN checklist and accompanying materials neglect, either by intention or ignorance, any discussion of the importance of fundamental measurement in PROM development and acceptance. If this criterion is recognized the result would be the rejection of most PRO instruments that have been used to support claims for therapy interventions. Rather than the COSMIN checklist we could apply a far less cumbersome (and time consuming) series of checks that conform to the requirements of RMT. These checks are readily accessible and have been clearly articulated by Bond and Cox [5].

Four main types of measurement scale are recognized: nominal, ordinal, interval and ratio. Each satisfies one or more of the properties of: (i) identity, where each value has a unique meaning; (ii) magnitude, where each value has an ordered relationship to other values; (iii) interval, where scale units are equal to one another; and (iv) ratio, where there is a 'true zero' below which no value exists. Nominal scales are purely descriptive and have no inherent value in terms of magnitude. Ordinal scales have both identity and magnitude in an ordered relation but the unknown distances between the ranks means the scale is capable only of generating median and modes. The interval scale has identity, magnitude and equal intervals. It supports mathematical operations of addition and subtraction. A ratio scale satisfies all properties, supporting the additional mathematical operations of multiplication and division. To these would be added the simultaneous conjoint measurement scale that underpins RMT; the only approach to instrument development for latent unidimensional constructs that meets the required axioms of fundamental measurement. It is important to point out that the Rasch model makes it is possible to generate interval scales from ordinal observations. However, despite experience in applying Rasch measurement over the past 60 years, researchers in health technology assessment have put this to one side in their neglect of the axioms of fundamental measurement.

Attitudes to, and awareness of, the need to meet the axioms of fundamental measurement in health technology assessment can be contrasted with the commitment to accurate and meaningful measurement in the physical sciences. The sheer irrationality of this position is seen in the history of measuring time. Alfred the Great pioneered (at least for Saxon England) the application of candles marked off in hourly segments. He made no claims that these were interval candle scales with invariance of comparisons. A strong draft and time flew by. He recognized these were ranked, approximate ordinal scales. Little seems to have changed from the eighth century for COSMIN. At least tradition supports them; the one saving grace was that Alfred's advisors had an excuse for being late to meetings.

With the exception of education and a lesser extent psychology, the commitment in the social sciences is to the accumulation of instruments that produce manifest scores. There are some exceptions, notably the needs-based outcome measures that have been produced over the last 20 years [9] [10] [11] . Otherwise there is a common belief that the use of ordinal scales constitutes measurement. In fact, unless response to therapy is measured accurately, we are unable to make any claims. Where data are derived from an interval-based measure, it becomes possible to show whether a therapy is efficacious.

Likert Scales

To illustrate the pitfalls in failing to achieve fundamental measurement, consider Likert scales. Most PROMs utilize Likert scales (typically with 5 or 7 response levels). The Emotional Impact Short Form scale of the ASCQ-Me instrument for adults with sickle cell disease (SCD) comprises five statements (e.g., in the past 7 days, how much did you worry about getting sick?') with five response levels ('not at all' to 'very much') for each statement [12]. The initial step in creating an aggregate response score is to add together the numerical values (1 – 5) assigned to the response levels for each statement, yielding a range of 5 to 25. This creates a score out of 20 with a minimum value of 5. These scores are then translated into a T-score with a standardized mean of 50 and a standard deviation of 10. The respondent is assessed on how their T-score deviates from that of the average respondent. A higher score represents a healthier status with respect to the average respondent in standard deviation units.

While this reporting as T-scores is unexceptional and is the common metric for the PROMIS system, care needs to be taken in interpreting the total score. Does this scale conform to the axioms of fundamental measurement? If the scale is to support the arithmetic operations to generate standardized T-scores then it must have ratio properties (i.e., a true zero). This is only possible if the scores for each of the five statements also have ratio properties as it not possible to add interval scores to generate a ratio score. As interval scores lack a true zero; they only support addition and subtraction within the individual scale.

Consider the emotional impact of SCD and the five questions that are presumed to elucidate the construct. The five levels are assigned an integer (5, 4, 3, 2, 1) for 'never/not at all' to 'always/often'. The usual method for analyzing Likert scale data is to disregard the implicit subjectivity of individual responses while making unwarranted assumptions about the meaning attached to the integer values. The scoring assumes that the scale is interval level with an integer value of 1 indicating a higher degree of agreement than 2, with integer 2 higher than integer 3 and so on. While this may seem a trivial point, it relies not on an assumption of a ranked response but of a ranked response where the distance between the integer responses is invariant. That is, an integer value of 2 means that the respondent is feeling twice as hopeless as a response with integer value 4; or someone with an integer value of 1 is feeling five times as hopeless as some who never feels hopeless and scores an integer value of 5. By assigning integer values the user falls into the trap of assuming that the responses are on an interval scale or even a ratio scale where the integer value of 1 is treated as a true zero.

But that is not all. It is also assumed that the responses for each of the 5 items are equivalent. Each item contributes the same amount to the total score. This assumes that the respondent finds it equally easy or difficult to respond to each item. This is at variance to the Rasch measurement system where, following the axioms of conjoint simultaneous measurement, it assumes that the probability of affirming an item depends on two factors: the ability of the respondent and the difficulty of the item.

In short, the traditional summation of Likert scale data, the 'add 'em up' school, is based on the a priori assumption that all of the items are of equal difficulty for all respondents and that the threshold between steps are of equal distance or equal value. This is clearly nonsensical; but it is either not yet recognized or simply or ignored. As these assumptions cannot be justified, the result is an addition of five manifest scores which is invalid under the axioms of fundamental measurement. A progression to T-scores rests, therefore, on a series of assumptions regarding measurement that cannot be justified. In the physical sciences proposing such a measure would be rejected out of hand.

Incompatible Paradigms

RMT is not compatible with either classical test theory (CTT) or item response theory (IRT). They are, as Bond and Cox point out, competing paradigms [5]. RMT takes the perspective that if the instrument is to meet fundamental measurement standards then we should adopt the Rasch *data-to-model* paradigm. If we are not concerned with, or are happy to ignore, questions of fundamental measurement, then we can follow the CTT or IRT *model-to-data* paradigm. The key distinction is that *RMT uses the measurement procedures of the physical sciences as the reference point* [5]. We can aim for the standards in the physical sciences by, as Stevens pointed out in the 1940s, allocating numbers to events *according to certain rules.* It is these rules that comprise RMT [13]. To reiterate: RMT is designed to construct fundamental measures. CTT and IRT focus on the observed data, these data have primacy and the results describe those data. As Bond and Cox emphasize: In general, CTT and IRT are *exploratory* and *descriptive* models; the Rasch model is *confirmatory* and *predictive* [5]. In other words, exploratory models must account for all the data (e.g., EQ-5D-3L TTO tariff equation) [14] while the RMT as a confirmatory framework requires the data to fit the model. This establishes the case that where the principles of conjoint simultaneous measurement are realized, the instrument's application can be defended as an interval scale.

The point to emphasize (and reemphasize) is that there was never any intent in the development of the vast majority of PROMs, both generic and disease specific, to aim for an instrument that had the possibility of meeting the standards for instrument development that characterizes the physical sciences. The result is not unexpected: in this example, the ASCQ-Me fails to meet the standards required for fundamental measurement.

Nothing but Manifest Scores

It should, perhaps, not come as a surprise that the principal conclusion of this brief review is that unless an instrument has been designed to have interval properties it must, by default, have ordinal properties. It is just a contents list. Unless otherwise identified, the overwhelming majority of generic and disease specific instruments have ordinal properties. They are manifest scores that may be incorrectly used and interpreted as if they had interval properties. By extension, they fail to meet standards for ratio scales. There is no true zero. All we have are scores on, at best, a unidimensional number line where there is no fixed zero, rather a series of ranked scores (including, to cite an obvious example, the EQ-5D-3L zero score which is not a true zero) with variable but unknown (and unknowable) distances between them. Unidimensionality does not imply interval scoring. Rasch measurement, where items are selected to fit the Rash model, is not a perfect response. All too many items from an initial set considered as representative of the latent trait may have been eliminated on the grounds of disordered thresholds, poor fit, negative point-measure correlations and competing additional dimensions. Even so, in RMT the axioms of fundamental measurement have been recognized and applied. They are a challenge to be met. This is in contrast to CTT and IRT where the importance of fundamental measurement is simply ignored.

A Discipline in Trouble

If we accept the COSMIN notion of 'responsiveness' as the common denominator in accepting any PRO that meets CTT criteria for clinical applications, then we have to assume that those following the COSMIN checklist are either unaware of the axioms of fundamental measurement or are quite prepared to ignore them to avoid rocking the boat. COSMIN's rejection of fundamental measurement, as criteria for instrument assessment, whether intentional or otherwise, relegates health technology assessment to pseudoscience. This has been a constant theme of commentaries published in *INNOVATIONS in Pharmacy* over the past 4 years[15]. The failure to meet (or even understand) the demarcation test between intelligent design and natural selection has been formalized by COSMIN. The result is pseudoscience. It is unacceptable unless you believe that evidence is created not discovered and truth is a consensus among the audience prepared to reject the axioms of fundamental measurement.

Willful or Immaterial?

COSMIN is widely accepted; this acceptance enshrines a commitment to PROM development that fails to meet the axioms of fundamental measurement. Whether this acceptance of the COSMIN checklist is through ignorance or the willful disregard of fundamental measurement, is an open question. There is a lot invested in the COSMIN checklist model of PROM development and interpretation, notably its robust endorsement of the EQ-5D instruments with their ordinal or manifest properties. It is difficult to see a willingness to admit that the emperor has no clothes. Would the COSMIN authors be willing to structure their checklist in the framework of RMT? A situation where the application of CTT is appropriately considered once the instrument has been completed and the items selected to meet the Rasch model. At this juncture the application of CTT is confirmatory rather than exploratory. However, a more cynical observer might conclude that the COSMIN standards, where the axioms of fundamental measurement are ignored, is just a belated attempt to rescue the hundreds of PROMs, their publication in prestigious (and less prestigious) journals and the myriad theses and degrees that have been awarded, from the ignominy of having to admit error and the attendant concerns from those who have taken false claims at face value.

Despite numerous attempts to argue for the application of the axioms of fundamental measurement, leaders in technology assessment have held firm [16]. One result is the commitment to incremental cost-per-QALY models where the QALY, as the utility is a manifest score, is an impossible mathematical construct. A commitment to these lifetime models implies evidence is created not discovered. Truth lies within, not without. Truth is consensus [17]. Welcome to the standards of health technology assessment.

ACKNOWLEDGEMENTS

I would like to thank Dr S P McKenna of Galen Research and the University of Manchester for his comments and suggestions on an earlier draft.

REFERENCES

[1] Mokkink L, Terwee C, Knol D et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMJ Medical Research Methodology*. 2010;10:22

[2] COSMIN website https://www.cosmin.nl/

[3] Langley PC. Nonsense on Stilts – Part 1: The ICER 2020-2023 Value Assessment Framework for Constructing Imaginary Worlds. *Inov Pharm*. 2020;11(1): No 12 https://pubs.lib.umn.edu/index.php/innovations/article/view/2402

[4] Langley PC. Value Assessment in Cystic Fibrosis: ICER's Rejection of the Axioms of Fundamental Measurement. Inov Pharm. 2020;11(2):No. 8 https://pubs.lib.umn.edu/index.php/innovations/article/view/3248

[5] Bond T, Cox C. Applying the Rasch Model: Fundamental measurement in the human sciences. New York: Routledge, 2015

[6] Merbitz C, Morris J, Grip J. Ordinal scales and foundations of misinference. *Arch Phys Med Rehabil*. 1989;70:308-7

[7] Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice (Editorial) J Rehab Med. 2012;l44:97-8

[8] Tennant A, McKenna S, Hagell P. Application of Rasch Analysis in the development and application of quality of life instruments. *Value Health*, 2004;7(1 Suppl 1):S22-26

[9] McKenna S, Wilburn J. Patient value: its nature, measurement, and role in real world evidence studies and outcomes-based reimbursement. *J Med Econ*. 2018;21)5):474-80

[10] McKenna S, Heaney A, Wilburn J et al. Measurement of patient reported outcomes 1: The search for the Holy Grail. *J Med Econ*. 2019;22(6):516-22

[11] McKenna S, Heaney A, Wilburn J. et al. Measurement of patient-reported outcomes. 2: Are current measures failing us? *J Med Econ.* 2019;22(6):523-30

[12] ASCQ-Me http://www.healthmeasures.net/explore-measurement-systems/ascq-me

[13] Stevens S. On the theory of scales of measurement. *Science*. 1946; 103:677-680

[14] Drummond M, Sculpher M, Torrance G et al. Methods for the Economic Evaluation of Heath Care Programmes (3d Ed.). New York: Oxford University Press, 2005

[15] Piglucci M. Nonsense on Stilts: How to tell science from bunk. Chicago: University of Chicago Press, 2010)

[16] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes. (4th Ed.) New York: Oxford University Press. 2015

[17] Wootton D. The Invention of Science: A new history of the scientific revolution. New York: Harper Collins, 2015.