MEDICAL COUNCIL OF CANADA
LE CONSEIL MÉDICAL DU CANADA

RECOMMENDATIONS FOR STANDARDIZATION OF THE MCC 360 SCALE

Marguerite Roy, Medical Council of Canada

Cindy Streefkerk, Medical Council of Canada

2017

Anticipated changes to the MCC 360 questionnaires

At the March 2016 meeting, the MCC 360 – Multisource Feedback (MSF) Committee expressed interest in making several changes to the MCC 360 questionnaires, including:

- Standardization of the scales across questionnaires
- Concentration on the assessment of CanMEDS intrinsic roles, particularly communicator, collaborator, and professional
- Greater generality and standardization with regard to the content of questions across specialties
- Greater triangulation of questions across respondent groups

Past PAR scales

The original set of PAR multisource feedback (MSF) questionnaires were developed for general practitioners and used a comparative scale. Over time, new questionnaires were developed in eight different specialty areas including anesthesiology, diagnostic laboratory, episodic care, medical specialties, pediatrics, psychiatry, surgery, and radiology. While all tools have employed 5-point Likert rating scales with an additional 6th option for “unable to assess”, there has been variation in the particular type of Likert scales used across respondent groups and specialty areas. Table 1 outlines the various scales currently in use.

Table 1. Likert scales used in the PAR instruments.

	Scale Value					
Type of Scale	**1**	**2**	**3**	**4**	**5**	**Unable to Assess**
Option 1 Comparative	Among the Worst	Bottom Half	Average	Top Half	Among the Best	Unable to Assess
Option 2 Agreement- 1	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Unable to Assess
Option 3 Agreement - 2	Strongly Disagree	(blank)	(blank)	(blank)	Strongly Agree	Unable to Assess
Option 4 Asymmetric - Evaluative	Poor	Fair	Good	Excellent	Outstanding	Unable to Assess

To standardize the scales across the questionnaires a few key questions were posed including:

- Which, if any, of the above scales should be selected for common use with the new MCC 360 tools?

- How many response options are appropriate?
- Should there be a neutral middle point?

MCC conducted a literature search on Medline and Psychinfo for research evidence to inform these decisions. The following is a high-level summary of research findings.

Desirable scale characteristics

What makes a good scale? The handbook on multisource feedback (Bracken, Timmreck, & Church, 2001) describes desirable scale characteristics for closed response items. Many of these have also been echoed for MSF questionnaires in the health services context (e.g., Wood, Hassell, Whitehouse, Bullock & Wall, 2006). Generally speaking, such scales or response options should:

- Fit appropriately with the item stem
- Cover the entire measurement continuum – scale options should completely cover the range of possible responses
- Be logically ordered and non-overlapping/mutually exclusive
- Have a precise and stable meaning
- Allow respondents to both sufficiently and accurately discriminate between scale options
- Be familiar language (low ambiguity) for respondents

Appropriate use of Likert scales

There are a range of scales and response styles that may be used when developing a questionnaire (Bowling, 1997). Many MSF tools employ Likert scales for this purpose both in healthcare and non-healthcare contexts (e.g., extended use in the business world). Likert scales use a fixed choice response format and are designed to measure attitudes or opinions. These ordinal scales are often used to measure levels of agreement/disagreement.

A Likert-type scale assumes that the strength/intensity of experience is linear, i.e. on a continuum from strongly agree to strongly disagree, and makes the assumption that attitudes can be measured. It is acceptable to treat scores from this type of response format as interval data to allow the use of common parametric tests (Cariffo & Peria, 2008; Norman, 2010).

Type of Likert scale

The type of Likert scale used should fit with the intended use of the tool. MCC 360 questionnaires are intended to measure observable behaviour from multiple perspectives (patient, non-physician co-worker, medical colleague, and self) to provide feedback to practicing physicians with regard to what they are doing well and behaviors that could be targeted for improvement. The purpose is not for physicians to rank themselves to other physicians, thus a comparative scale (e.g., among the best) is not appropriate for the intended purpose of MCC 360.

In the context of business, Dalessio (1998) suggested that agreement, satisfaction, and extent response scales are all adequate for 360-degree feedback questionnaires. Ultimately, the choice should fit the context and intended use.

Based on a preliminary review of the use of the asymmetric scale, there is no strong evidence to support that use of the scale leads to improved discrimination and enhanced score distribution.

Recommendation: Use of an agreement scale as a comparative scale is not appropriate for the intended use.

Number of response options

In a summary of the literature on questionnaire design Lietz (2010) reported that offering 5 to 7 response options is most commonly used and that enhanced psychometric qualities, including reliability and validity, are reported as the justification for including multiple response categories. But what number of response points is optimal?

Length of scale (i.e., number of options) can impact the process by which people map their attitudes onto the response alternatives (Krosnick & Presser, 2010). The value of increasing the number of scale points must be weighed against the potential confusion or variation in how respondents interpret and are able to clearly distinguish among response options. Ideally, the number of options should reflect and align with people’s mental representation of the construct of interest. Too many choices can lead to problems of accuracy (DeVellis, 1991). Bracken, Timmreck, & Church (2001) recommend 5 to 7 values/response options to allow distinctions that are fairly fine but not artificial.

Lissitz and Green (1975) explored the reliability of various numbers of scale points using simulations and found reliability increased from 3- to 5-point scales but then leveled off for 7-, 9- and 14-point scales. Similar results have been reported by others (Jenkins & Taber, 1977; Srinivasan & Basu, 1989). Hassell et al. (2012) investigated the effect of varying the number of rating scale points between 3, 4, 6, and 9 points on an MSF tool for medical residents. Ironically, they found that longer scales were associated with identifying fewer “concern” trainees as they identified more “above expectation” trainees. Furthermore, assessors reported having some difficulty interpreting longer scales.

Recommendation: Continue with a 5-point rating scale.

Including a neutral midpoint

There is some controversy in the literature about the value of including a neutral midpoint. For example, London, Wohlers, and Gallagher (1990) recommend against including a midpoint because it avoids receiving only neutral ratings. Others, however, have pointed out that respondents may become frustrated if not provided with a neutral option and increase the non-response bias (Burns & Grove, 1997).

Nadler, Weston, & Voyles (2015) found participants' interpretation of the midpoint could vary substantially. The authors suggest that when using a midpoint an explicit definition of what the midpoint indicates should be provided. Rohrmann (2007) tested the sociolinguistic and psychometric properties of 100 expressions for five quantifier dimensions including; intensity, frequency, probability, quality and agreement. He investigated use with 5-point to 9-point rating scales across 5 experiments involving both student and the general population. His findings indicate that the most familiar and preferred midpoint for an agreement scale was "neutral".

In addition to a providing a neutral midpoint, it is also recommended to use a "don't know" or "not applicable" option to prevent bias introduced when a respondent lacks adequate knowledge (Bracken, Timmreck, & Church, 2001).

Recommendation: Include a neutral midpoint, define what the midpoint means. Continue use of "don't know" or "not applicable".

Response labels

When scales are verbally labelled for each point the measurement of validity improves (Krosnick & Fabrigar, 1997) and extreme response style bias decreases (Moors, Kieruj, & Vermunt, 2014). Appropriate use of verbal labels remains a focal point for designing a survey that will yield more accurate data (Klockars & Yamagishi, 1988).

Ideally, the response scale labels should align with the purpose of the questionnaire (Crossley & Jolly, 2012). Rohrmann (2007) investigation into sociolinguistic and psychometric properties of a 5-point agreement scale provides evidence that the public is able to clearly distinguish and is familiar with the terms: strongly disagree, disagree, neutral, agree, and strongly agree.

Recommendation: Label each point of measurement

Overall recommendations

Recommend using Option 2.

Table 2. Recommended Likert scale.

	Scale Value					
Type of Scale	**1**	**2**	**3**	**4**	**5**	**Unable to Assess**
Option 2 Agreement - 1	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Unable to Assess

Based on this summary of the literature:

- Maintain use of 5-point Likert scale, as more than 5 points probably leads to confusion over meaning and distinctness of options
- Maintain use of a 6th “Unable to Assess” option
- Apply verbal labels to anchor each response option, and options need to be explicit
- Move to common agreement scale as the purpose is not to rank physicians but to provide feedback about what they are doing well and where there is room for improvement. It is too soon to move to an asymmetric scale (need more evidence)
- Maintain the neutral middle point as it avoids non-response bias, and evidence supports that this is the most preferred and familiar term

References

Bowling, A. (1997). *Research Methods in Health*. Buckingham: Open University Press.

Bracken, D.W., Timmreck, C.W., & Church, A.H. (Eds.) (2001). *The Handbook of multisource feedback: The comprehensive resource for designing and implementing MSF Processes*. San Francisco: Jossey-Bass.

Burns & Grove (1997)

Carlifo, J., & Perla, R.J., (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, *42*, 1150-1152.

Coleman, A.M., Norris, C.E., & Preston, C.C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-poing and 7-point scales. *Psychological Reports*, *89*, 355-362.

Crossley, J., & Jolly, B., (2012). Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. *Medical Education, 46,*28-37.

Dalessio, A. (1998). Using multisource feedback for employee development and personnel decisions. In J. W. Smither (Ed.), *Performance appraisal: State of the art in practice* (pp. 278–330). San Francisco: Jossey-Bass.

DeVellis, R. (1991). *Scale development: Theory and applications* (2nd Ed.). Sage Publications.

Friedman, H.H., & Amoo, T., (1999). Rating the rating scales. *Journal of Marketing Management*, *9*, 114-123.

Hassell, A., Bullock, A., Whitehouse, A., Wood, L., Jones, P., & Wall, D. (2012). Effect of rating scales on scores given to junior doctors in multi-source feedback. *Postgrad Med J*, *88*, 10-14.

Jenkins, G. D., & Taber, T. D. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62*, 392-398.

Klockars A., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement*, *25*, 85-96.

Krosnick, J.A., & Presser, S. (2010). Question and questionnaire design. In P.V. Marsden & J.D. Wright (Eds.), *Handbook of survey research* (pp. 263-314). Bingley, UK: Emerald Group Publishing Limited.

Krosnick, J.A., & Fabrigar, L.R.. (1997). Designing rating scales for effective measurement in surveys, In L. Lyberg, P. Biemer, M. Collins, L. Decker, E. DeLeeuw, C. Dippa, N.

Schwarz, & D. Trewin (Eds.). *Survey measurement and process quality* (pp. 141-164). NY: John Wiley & Sons.

Lietz P. *(2010). Research into questionnaire design: A summary of the literature. International Journal of Market Research, 52, 249-272.*

Lissitz, R.W., & Green, S. B. (1975). Effect of the number of scale points on reliablility: A Monte Carlo approach. Journal of Applied Psychology, 60, 10-13.

Moors, G., Kieruj, N.D., & Vermunt, J.K. (2014). The effect of labeling and number of response scales on the likelihood of response bias. *Sociological Methodology*, *44*, 369-399.

Nadler, J.T., Weston, R., & Voyles, E.C. (2015). Stuck in the middle: The use and interpretation of mid-points in items on questionnaires. *The Journal of General Psychology*, *142*, 71 -89.

Norman, G. (2010). Likert scales, levels of measurement and "laws" of statistics. *Advances in Health Sciences Education*, 15, 625-632.

Rohrmann, B., (2007). Verbal qualifiers for rating scales: *Sociolinguistic considerations and psychometric data*. Technical Report, *rohrmannresearch.net/pdfs/rohrmann-vqs-report.pdf*.

Srinivasan, V., & Basu, A., (1989). The metric quality of ordered categorical data. Marketing Science, 8, 205-230.

Wood, L., Hassell, A., Whitehouse, A., Bullock, A., & Wall, D. (2006). A literature review of multisource feedback within and without health services, leading to 10 tips for their successful design. *Medical Teacher*, *28*, e185-191.