Patient Related Outcome Measures

Dovepress

open access to scientific and medical research

Open Access Full Text Article

COMMENTARY

How to Improve Interpretability of Patient-Reported Outcome Measures for Clinical Use: A Perspective on Measuring Abilities and Feelings

Jacek A Kopec[1,2]

[1]School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada; [2]Arthritis Research Canada, Vancouver, BC, Canada

Correspondence: Jacek A Kopec, School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada, Email jkopec@arthritisresearch.ca

Abstract: Two general classes of concepts measured by patient-reported outcome measures (PROMs) are abilities and feelings. Over the past several decades, there has been a significant progress in measuring both. Nevertheless, current multi-item scales are subject to criticism related to scale length, score dimensionality, interpretability, cultural bias, and insufficient detail in measuring specific domains. To address some of these issues, the author offers an alternative perspective on how questions about abilities and feelings could be formulated. Abilities can be defined in terms of a relationship between the level of performance and the associated perception of difficulty, and represented graphically by an ability curve. For feelings, it may be useful to measure frequency and intensity jointly to determine the proportion of time in each level of intensity. The resultant frequency × intensity matrix can be presented as a bar graph. Empirical data to support the feasibility and validity of these approaches to PROM design are provided, potential advantages and limitations are discussed, and some future research avenues are suggested.

Keywords: patient-reported outcome measures, PROMs, abilities, feelings, measurement, quality of life

Introduction

Measuring health outcomes by self-report is now widely accepted by the medical research community and patient-reported outcome measures (PROMs) play an increasingly important role in healthcare.[1] PROMs are useful for evaluating the effects of medical interventions and measuring disease burden in populations. PROMs can also be used to monitor progress in individual patients, improve doctor-patient communication, and help clinicians assess the needs of their patients.[1,2]

Methods for PROM development and validation have evolved over the past several decades. The dominant methodology originated in psychometric theory as applied in education and psychology.[3,4] The vast majority of modern PROMs use multi-item scales to measure selected aspects of health, referred to as domains or constructs and regarded as latent variables. Qualitative methods are routinely applied to develop an item pool and statistical methods are used to select the final items and show that the scales measuring each domain are valid, reliable, and responsive to change.[4–6] Validation often involves assessing if the scales are unidimensional and whether the scores correlate with other variables as expected. Increasingly, item response theory (IRT) is applied to select the items and calculate the score.[6,7]

Despite the progress in health measurement methodology, a potential problem with current PROMs that may pose a barrier to a wider use of such measures in a clinical setting is score interpretability.[8–11] There are several reasons why scores from PROMs may not be easily interpretable. Scores are often based on many items, pertaining to a range of different activities or experiences, in order to ensure content validity for scales designed to measure relatively broad concepts. Domain names, such as mobility, physical function, usual activity, vitality, self-efficacy, psychological distress,

Received: 24 December 2021
Accepted: 17 March 2022
Published: 25 March 2022

or emotional role limitations, may reflect the underlying latent constructs as defined by the PROM developers but have less meaning to the clinical user. For a clinician, interpretation of scores for such domains may be difficult without the knowledge of the actual items. In a clinical context, responses to individual questions can be more meaningful than numerical scores derived from multi-item scales and converted to standard deviations above/below a population mean or percentage of the maximum possible score.[4] In computer adaptive testing, individual items are considered exchangeable (different patients respond to different items), which makes interpretation even more difficult. Furthermore, the cognitive processes involved in formulating a response to PROM questions are poorly understood.[11] It seems likely that many questions are interpreted differently by different people, leading to incomparability of responses.[11,12] In addition, most PROMs are developed in high-income countries, often in English. Translation and cultural adaptation of PROMs is a challenging and time-consuming process. Despite best efforts, bias resulting from differences in meaning is hard to avoid, especially for items that involve culture-related concepts or activities.[13] Finally, the key assumptions of measurement models used in developing and validating PROMs, such as scale unidimensionality and item (local) independence, may be violated even in the presence of apparently acceptable results of standard statistical tests.[14]

Some of the aforementioned issues can be addressed by applying modern statistical methods. For example, multidimensional IRT models can be used to analyze pools of items that assess multiple constructs.[15] Latent variable mixture models may increase measurement invariance between individuals by improving item selection and score estimation methods.[16] Advances in understanding response processes may suggest better approaches to item development.[17] A review of these and other improvements in psychometric methods would be a useful exercise, but is beyond the scope of this commentary.

The intention of this article is to offer a different perspective on how some problems with score interpretability might be approached in the development and application of PROMs. The approach suggested here involves measuring more specific, explicitly defined domains, and modifying the way questions pertaining to these domains are asked and responses are presented. The proposed approach is not meant to replace the dominant methodology or advance psychometric theory. Rather, the author hopes that the viewpoint presented here may encourage more discussion about ways to improve PROM's interpretation and increase their use by clinicians.

Classification of Health Outcomes

PROMs are often classified into measures of physical, mental, and social health,[18] although many other classifications have been proposed.[19] However, from a measurement point of view, it may be useful to classify health domains measured by PROMs into two types, abilities and feelings. This simple classification is useful because abilities and feelings differ in some key characteristics and may, therefore, require different approaches to the development and validation of measures for each. In this framework, abilities refer to things people do (activities), for example, ability to walk, think, or work. Feelings include emotions, such as being sad or anxious, and sensations (symptoms), such as pain or fatigue.

An important difference between abilities and feelings is that feelings are inherently subjective (internal), whereas abilities can be (and often are) measured with the help of an external observer and/or a physical measurement device. Another difference is that abilities tend to be relatively stable, whereas feelings fluctuate more frequently and rapidly over time.[20,21] It should be clear that abilities can change, eg, due to an accident or disease, and some feelings can be quite persistent. Nonetheless, questions about frequency or duration are commonly employed in PROMs when measuring feelings but are considered less relevant for assessing abilities. The distinction between abilities and feelings is not always obvious because there are strong and reciprocal causal relations between them, a fact that is often utilized for measuring both. A subjective feeling of difficulty associated with activity is important for measuring abilities, for example, in questions asking about difficulty walking a specified distance, and interference with activities is critical for measuring feelings, for instance, in questions asking how much pain interferes with daily activities.

A Model for Measuring Abilities

In self-report measures of physical abilities, the typical approach is to select a sample of activities, ask about difficulty or limitation in performing these activities, and combine the responses into a summary score using standard psychometric

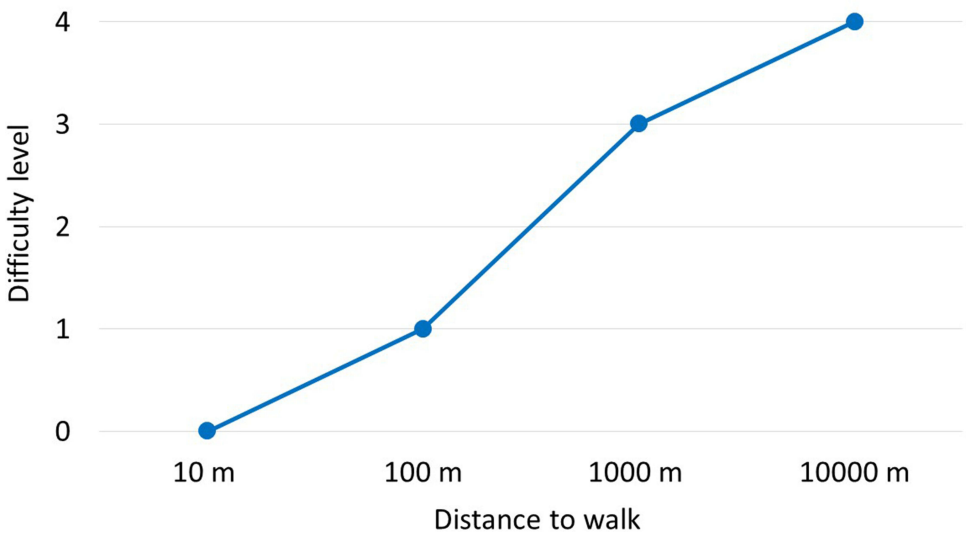

Figure 1 A hypothetical ability curve for walking.

methods. Examples are measures of mobility, upper extremity function, or overall physical function.[22] Such measures combine many different abilities, for example, walking, running, standing, climbing stairs, dressing, doing chores, or playing sports. For each ability, difficulty is usually assessed for a single, selected level of performance, eg, walking 1 km or climbing 2 flights of stairs. Self-report measures of mental abilities are developed using similar principles and may include items about ability to think, remember, and concentrate.[22] These types of scales may be susceptible to the problems with score interpretability discussed previously, such as ambiguity about the concept being measured, multi-dimensionality, or a lack of measurement invariance.

How can interpretability of ability measures be improved? It may be noted that the level of difficulty with a given activity depends on the level of performance that can often be defined in physical units of distance, weight, speed, duration, etc. Rather than defining ability as a latent variable reflected by a score on a multi-item scale (where each item is an independent measure of the underlying construct), ability could be thought of as a relationship between the level of performance and the associated perception of difficulty.[23] This relationship can be elicited by asking a series of ordered, linked questions and presented as a (monotonic) *ability curve*. For example, to measure ability to walk, we can ask about difficulty walking different distances ranging from, say, 10 m to 10 km (Figure 1). This seems straightforward for simple physical activities, but the general approach might be applicable to other types of activities, including mental activities. A summary score for a given ability, corresponding to the area under the ability curve (integral of the ability function), can be estimated as a simple sum of difficulty scores for each level of performance. A summary measure combining several abilities into an overall score can be obtained as well, if desired, using standard statistical methods such as factor analysis. To this end, each ability curve (or area under the curve) can be treated as an independent "item", scored on a numerical scale. Properties of individual questions can also be analyzed, although this would require statistical models that do not consider such questions as independent.

A Model for Measuring Feelings

Current PROMs include questions about a large number of specific feelings (emotions or symptoms) combined into broader constructs such as anxiety, depression, anger, fatigue, or pain.[24] Other measures ask multiple questions about interference of a particular feeling, for example, pain, with different activities. Alternatively, multiple symptoms are included in a symptom checklist whereby each symptom is measured by a single question.[25]

The problems with scale interpretability, including multidimensionality, concept ambiguity, or a lack of cultural equivalence in measures of feelings are similar to those we have seen with abilities. In particular, integrating different feelings into a single score may not always be warranted. For example, it is common to define depression or anxiety as

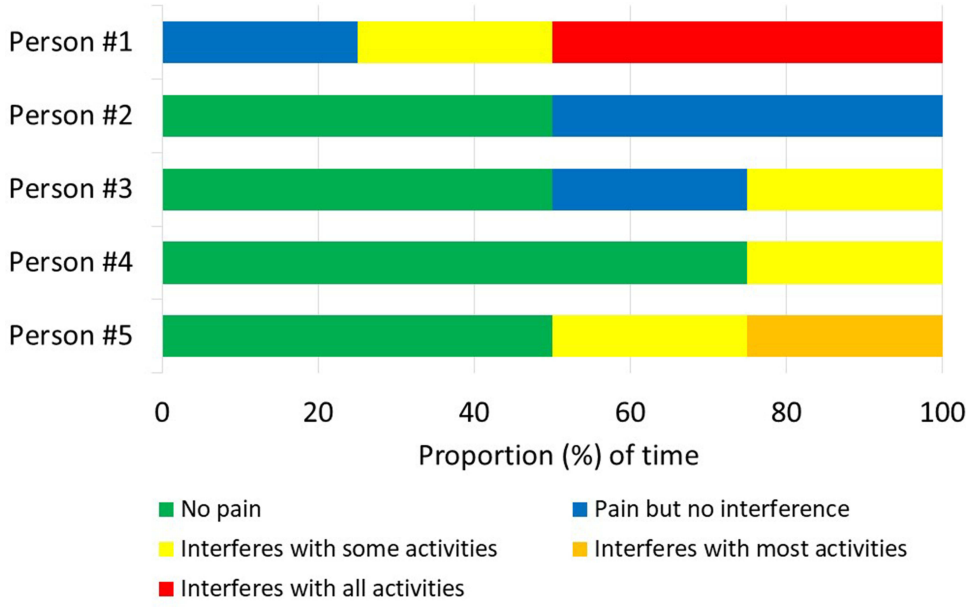

Figure 2 Examples of individual time-with-pain patterns among members of the Canadian Association of Retired Persons.

a condition that encompasses a range of different feelings. Items measuring depression may ask about being depressed as well as being unhappy, sad, disappointed in oneself, pessimistic, discouraged about the future, emotionally exhausted, a failure, helpless, hopeless, worthless, and so on.[24] It is possible that these feelings are caused by a single biological mechanism or that the adjectives describing them are semantically equivalent and, therefore, measure the same latent construct of depression. However, in some contexts, it may be beneficial to avoid the assumption of a latent construct and instead assume that each item measures a slightly different and potentially useful concept. For example, a question about being hopeless measures hopelessness and a question about being helpless measures helplessness.[26]

Another problem is that current PROMs routinely ask about either frequency (or duration) of each feeling and ignore intensity (severity) or ask about intensity and ignore frequency.[22] It should be clear that both intensity/severity and frequency/duration are important to measure for all feelings. For example, being a little anxious sometimes is clearly different from being very anxious sometimes. Similarly, knowing the average intensity of pain, fatigue, or anxiety over a period of time is not sufficient, and can be meaningless, if these feelings fluctuate. In fact, it is well established that reports of average intensity are strongly influenced by peak and end-of-period intensity.[27] When feelings are measured by interference with or impact on activity, it may not even be clear if the questions pertain to frequency or average degree of interference.

It is suggested here that the above problems with score interpretability could potentially be mitigated by asking about intensity and frequency of feelings *jointly*, in order to estimate the amount of time in each level of intensity/interference. In this model, the patient is asked several linked questions about a particular feeling to complete the frequency × severity matrix. For example, we may ask about the amount of time with any pain, and with pain that interferes with none, some, most, and all activities. The results can be presented graphically, eg, as a single "stacked" bar with different colors representing different severity levels and the size (length) of each part of the bar representing the time in each level (Figure 2). A summary measure for each feeling can be derived as a weighted sum, whereby each intensity level is given a weight and is multiplied by the proportion of time in that level. The weights may not be equally spaced and can be derived empirically, for example, using preferences.[28] If psychometrically or conceptually justified, scores for different feelings can be combined into an overall measure of a broader concept in the same way individual items are combined in conventional measures.

Feasibility and Measurement Properties

The approach to measuring abilities discussed here was first described by Kopec more than 2 decades ago.[23] In a more recent paper, Kopec et al provided data from an online population survey among 1089 members of the Canadian

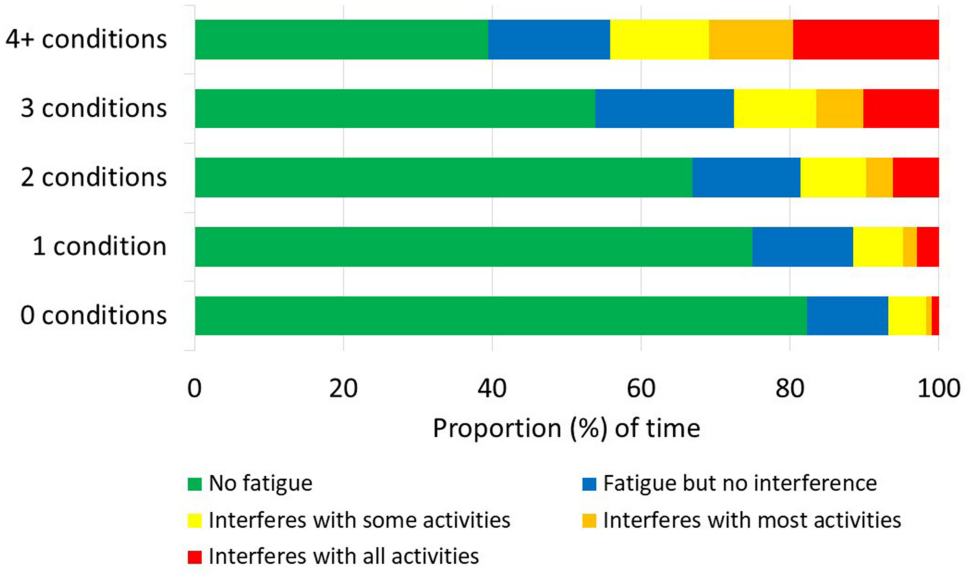

Figure 3 Average proportion of time in different levels of fatigue by number of conditions among members of the Canadian Association of Retired Persons.

Association of Retired Persons in which this method was applied to measure abilities to walk, run, and lift objects.[29] The authors have shown that the approach is feasible and the scores are reliable and valid. Test-retest correlation coefficients were 0.89 for walking, 0.88 for running, and 0.81 for lifting. Men reported higher abilities than women. Scores correlated with the domains of the SF-36[30] and CAT-5D-QOL questionnaire,[31] as expected. The study also provided detailed population data on these abilities not available from current PROMs, including IRT-based measures. For example, the study found that about 14% of older Canadians had difficulty walking 10 m, 22% had difficulty waking 100 m and 38% had difficulty walking 1 km, while 15% were unable to run 10 m and 51% had difficulty lifting 10 kg.[29]

In the same online survey of older Canadians, Kopec et al measured pain, fatigue, anxiety and depression using the approach proposed here.[32] Survey questions were generated following a focus group with patients with chronic conditions. Each feeling was measured using 4 questions with 5 categorical responses each, asking how much of the time the respondent experienced a specified level of each feeling during the last week. The levels were described in terms of activity interference. Examples of observed individual patterns are shown in Figure 2. Average time distributions for symptom levels varied according to gender, number of chronic conditions (Figure 3), as well as use of medication and health services. Scores on all four scales correlated as expected with the Hospital Anxiety and Depression Scale,[33] SF-36, and CAT-5D-QOL domains.

Discussion

In this article, the author briefly reviewed potential problems with the interpretation of scores from modern PROMs and suggested alternative ways to formulate questions and calculate scores for commonly measured health domains, classified as either abilities or feelings. Specific abilities can be measured in terms of ability curves, depicting the relationship between the well-defined performance level and the associated perception of difficulty. A score for each ability can be conceptualized as the area under the curve. If needed, different abilities can be combined into an overall score. For measuring feelings, it is possible to assess frequency and intensity (or interference) jointly in order to determine the amount of time in each level of intensity. The result of measurement can be presented visually as a bar graph. A score for each feeling, which can be interpreted as burden associated with a particular (negative) emotion or symptom, can be obtained as a sum of the time in each level multiplied by intensity weight and expressed as a percentage of the maximum burden. Scoring relative to a population norm is also possible. Similar feelings can potentially be merged into measures of broader concepts. Preliminary evidence of feasibility, validity and reliability of the approaches proposed here has been provided.

It is useful to note two conceptual differences between the instrument design suggested here and the usual approach to PROM design rooted in modern psychometrics. First, most current measures, including those based on IRT, generally assume that items asking about different activities or feelings can measure the same underlying construct and unidimensionality can be demonstrated statistically (this applies to reflective scales; formative indices are not discussed). Here, the design requires that all items ask about the same activity or feeling and unidimensionality is achieved conceptually. Second, standard measures assume that all items measuring the same construct are locally independent, that is, conditional on the construct, responses to different items are not correlated. In the proposed model, items measuring the same ability or feeling are assumed to be logically and causally related and are considered jointly.

The proposed models for measuring abilities and feelings might potentially offer some practical advantages compared with current PROMs, such as greater interpretability and more detail in assessing narrowly defined concepts, although it should be emphasized that these advantages are hypothetical at this time, as empirical evidence is not yet available. These advantages may be more important in a clinical setting, where assessment is often individualized and focused on areas most relevant to the patient.

Using items asking about the same activity or feeling may diminish any potential ambiguity as to what construct is being measured, while specifying the level of performance in physical units may make it easier for the clinical user to imagine what the responses mean in practical terms, thus enhancing score interpretability. Standard PROMs often include items that specify the level of performance. For example, there are 3 items about walking different distances in the SF-36.[29] An important distinction, however, is that these items are treated as independent and are not selected or designed specifically to derive an ability curve. Because the ability curve depicts, in a systematic way, the amount of difficulty for all levels of performance for a specific activity, it may be easier to interpret than a numerical score based on items covering a range of different activities at arbitrarily selected levels. Similarly, a frequency × intensity matrix may provide a more comprehensive and detailed assessment of a specific feeling than separate items about frequency or average intensity because it identifies fluctuations in intensity and, therefore, may better approximate the theoretical ideal, ie, continuous assessment of intensity over time. This might also help reduce peak and end bias associated with standard questions about average intensity.

In the proposed approach, the maximum and minimum possible scores may be more likely to represent absolute extremes than maximum and minimum scores in many current measures. For example, the highest score for ability to walk may be logically described as no difficulty walking a very long distance, such as 10 km, and the lowest as unable to walk a very short distance, such as 10 m. Similarly, the highest vs lowest score for anxiety would be obtained when reporting very severe anxiety all the time vs no anxiety at any time. Such extreme scores naturally derive from the design of the instrument and should not depend on a particular item bank or population norms. In addition, it might be easier to design measures such that significant ceiling or floor effects are avoided.

It is unknown at this time if the proposed strategy may be preferred in dealing with response shift, and generally, assessment of change in abilities or feelings over time. It is possible that greater specificity in defining performance levels when assessing abilities may help reduce potential bias in responses. With respect to measuring feelings, empirical data are needed to determine if the suggested design is subject to bias in measuring change to the same extent as standard measures.

Finally, this method of formulating questions may facilitate translation and cultural adaptation of PROMs. Although more research is needed, it seems possible that a set of linked questions that explicitly pertain to difficulty with a single activity across the full range of well-defined performance levels, or a single feeling across a full range of frequency and intensity, might be easier to translate to other languages and achieve cultural comparability than an item bank that covers a wide variety of activities or feelings at arbitrary or unspecified levels.

There are, however, several practical and conceptual issues related to the measurement approach proposed here that require more research and discussion. First, it is not obvious that there is a practical need for greater granularity in measuring abilities or feelings. For many users it may be more important to ensure comprehensiveness in assessing broader domains by including items covering a wider range of topics in a single scale. Second, traditional scales, by treating all items as independent measures of the same concept, offer greater flexibility in item selection and in generating scales of desired length, reliability, or information, especially when computerized adaptive testing is used.

Third, traditional scales of the same length may be more reliable and discriminating than the measures proposed here. Fourth, defining performance levels in physical units may be awkward for some types of activities, especially complex physical activities and mental activities. Fifth, responding to questions about time in different levels of intensity for a particular feeling may require more cognitive effort than responding to questions about frequency or average intensity independently. More research on how respondents interpret such questions is needed. Sixth, when questions are conceptually linked rather than independent, it may not be obvious how to deal with inconsistent responses, for example, non-monotonic ability curves or logically impossible distributions of time in different intensity levels. Seventh, the lack of independence between the items may require new psychometric models to analyze the properties of individual items.

Furthermore, critics might argue that the current proposal does not take into account recent advances in measurement theory, psychometrics, statistical methodology, and data science. Examples include the use of multidimensional IRT models when assessing multiple constructs, latent variable mixture models to improve measurement invariance, or progress in understanding response processes.[15–17] Also, this proposal does not discuss the importance of patient engagement in measure development. Greater involvement of patients and other stakeholders in PROM development has arguably been one of the most important factors in improving the design of PROMs over the past two decades.[1,2]

When responding to such arguments it should be clear that the intention of the viewpoint presented here is not to advance measurement theory or change the way the vast majority of PROMs are developed and used. Rather, the objective is to stimulate discussion about alternatives to the standard PROM design that might help improve their interpretability and thereby increase their use, especially in a clinical setting. In selecting PROMs, there are often trade-offs between scale length, content validity, reliability, discrimination, interpretability, and other characteristics of the available instruments. A measure that is less reliable or discriminating may still be preferable if it is easier to interpret.[34,35]

The proposed approach to health measure design might encourage new lines of research. For example, measuring feelings the way suggested here may facilitate the study of the relationship between language and emotions and help determine to what extent the adjectives and other phrases commonly used to describe feelings are semantically distinguishable in different languages and have a biological basis.[36] Another potentially fruitful research avenue might be the study of the mathematical relationships between performance levels, conceptualized as a stimulus, and perception of difficulty, regarded as a response, somewhat analogous to research in psychophysics.[37] The proposed strategy may also help in developing "gold standard" methods of measuring abilities, whereby the individual is asked to perform a given activity at systematically varied levels and the perception of difficulty is elicited multiple times during the test.

In conclusion, this article suggests possible modifications to the standard methods of asking questions about abilities and feelings in self-reported measures of health outcomes. As the importance of measuring outcomes from a patient's perspective is increasingly appreciated, it is hoped that the ideas presented here will generate some interest among PROM users and developers.

Disclosure

Dr Jacek A Kopec reports grants from Canada Foundation for Innovation, during the conduct of the study.

References

1. Black N, Burke L, Forrest CB, et al. Patient-reported outcomes: pathways to better health, better services, and better societies. *Qual Life Res*. 2016;25(5):1103–1112. doi:10.1007/s11136-015-1168-3
2. Calvert M, Kyte D, Price G, Valderas JM, Hjollund NH. Maximising the impact of patient reported outcome assessment for patients and society. *Br Med J*. 2019;364:k5267. doi:10.1136/bmj.k5267
3. Nunnally JC. *Psychometric Theory*. McGraw-Hill; 1967.
4. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 5th ed. Oxford University Press; 2014.
5. Hawkins M, Elsworth GR, Osborne RH. Application of validity theory and methodology to patient-reported outcome measures (PROMs): building an argument for validity. *Qual Life Res*. 2018;27(7):1695–1710. doi:10.1007/s11136-018-1815-6
6. Stover AM, McLeod LD, Langer MM, Chen WH, Reeve BB. State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. *J Patient-Rep Outcomes*. 2019;3(1):50. doi:10.1186/s41687-019-0130-5
7. Hays RD, Lipscomb J. Next steps for use of item response theory in the assessment of health outcomes. *Qual Life Res*. 2007;16(Suppl 1):195–199. doi:10.1007/s11136-007-9175-7

8. Nguyen H, Butow P, Dhillon H, Sundaresan P. A review of the barriers to using patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs) in routine cancer care. *J Med Radiat Sci*. 2021;68(2):186–195. doi:10.1002/jmrs.421
9. Philpot LM, Barnes SA, Brown RM, et al. Barriers and benefits to the use of patient-reported outcome measures in routine clinical care: a qualitative study. *Am J Med Qual*. 2018;33(4):359–364. doi:10.1177/1062860617745986
10. Spertus J. Barriers to the use of patient-reported outcomes in clinical care (Editorial). *Circ Cardiovasc Qual Outcomes*. 2014;7(1):2–4.
11. Westerman MJ, Hak T, Sprangers MA, Groen HJ, van der Wal G, The AM. Listen to their answers! Response behaviour in the measurement of physical and role functioning. *Qual Life Res*. 2008;17(4):549–558. doi:10.1007/s11136-008-9333-6
12. Sawatzky R, Russell LB, Sajobi TT, Lix LM, Kopec J, Zumbo BD. (2018). The use of latent variable mixture models to identify invariant items in test construction. *Qual Life Res*. 2018;27(7):1745–1755. doi:10.1007/s11136-017-1680-8
13. Wagner AK, Gandek B, Aaronson NK, et al. Cross-cultural comparisons of the content of SF-36 translations across 10 countries: results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol*. 1998;51(11):925–932. doi:10.1016/S0895-4356(98)00083-3
14. Ziegler M, Hagemann D. Testing the unidimensionality of items. Pitfalls and loopholes. *Eur J Psychol Assess*. 2015;31(4):231–237. doi:10.1027/1015-5759/a000309
15. Bass M, Morris S, Neapolitan R Utilizing multidimensional computer adaptive testing to mitigate burden with patient reported outcomes. AMIA Annual Symposium Proceedings; 2015:320–328. eCollection 2015.
16. Sawatzky R, Ratner PA, Kopec JA, Zumbo BD. Latent variable mixture models: a promising approach for the validation of patient reported outcomes. *Qual Life Res*. 2012;21(4):637–650. doi:10.1007/s11136-011-9976-6
17. Zumbo BD, Hubley AM, Eds. *Understanding and Investigating Response Processes in Validation Research*. Springer International Publishing AG; 2017.
18. HealthMeasures. Intro to PROMIS® (Patient-Reported Outcomes Measurement Information System). Available from: https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis. Accessed May 10, 2021.
19. Valderas JM, Alonso J. Patient reported outcome measures: a model-based classification system for research and clinical practice. *Qual Life Res*. 2008;17(9):1125–1135. doi:10.1007/s11136-008-9396-4
20. Schneider S, Junghaenel DU, Keefe FJ, Schwartz JE, Stone AA, Broderick JE. Individual differences in the day-to-day variability of pain, fatigue, and well-being in patients with rheumatic disease: associations with psychological variables. *Pain*. 2012;153(4):813–822. doi:10.1016/j.pain.2012.01.001
21. Mendes de Leon CF, Guralnik JM, Bandeen-Roche K. Short-term change in physical function and disability: the Women's Health and Aging Study. *J Gerontol B Psychol Sci Soc Sci*. 2002;57(6):S355–65. doi:10.1093/geronb/57.6.S355
22. HealthMeasures. PROMIS® (Patient-Reported Outcomes Measurement Information System). List of Adult Measures. Available from: https://www.healthmeasures.net/index.php?option=com_content&view=category&layout=blog&id=113&Itemid=808. Accessed May 10, 2021.
23. Kopec JA. Concepts of disability: the activity space model. *Soc Sci Med*. 1995;40(5):649–656. doi:10.1016/0277-9536(95)80009-9
24. HealthMeasures. PROMIS® (Patient-Reported Outcomes Measurement Information System). Assessment Centre. Computerized Adaptive Test (CAT) Demonstration Page: depression. Available from: https://www.assessmentcenter.net/ac1/Default.aspx?SID=5B6B4AFE-B791-4E30-A232-29EF4B03C6C2. Accessed May 10, 2021.
25. Zijlema WL, Stolk RP, Löwe B, et al. How to assess common somatic symptoms in large-scale studies: a systematic review of questionnaires. *Psychosom Res*. 2013;74(6):459–468. doi:10.1016/j.jpsychores.2013.03.093
26. Lester D. An inventory to measure helplessness, hopelessness, and haplessness. *Psychol Rep*. 2001;89(3):495–498. doi:10.2466/pr0.2001.89.3.495
27. Kahneman D. Evaluation by moments, past and future. In: Kahneman D, Tversky A, editors. *Choices, Values and Frames*. Cambridge University Press; 2000:693–708.
28. Kopec JA, Sayre EC, Rogers P, et al. Multiattribute health utility scoring for the computerized adaptive measure CAT-5D-QOL was developed and validated. *J Clin Epidemiol*. 2015;68(10):1213–1220. doi:10.1016/j.jclinepi.2015.03.020
29. Kopec JA, Russell L, Sayre EC, Rahman MM. Self-reported ability to walk, run, and lift objects among older Canadians. *Rehabil Res Pract*. 2017;2017:1921740. doi:10.1155/2017/1921740
30. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care*. 1992;30(6):473–483. doi:10.1097/00005650-199206000-00002
31. Kopec JA, Sayre EC, Davis AM, et al. Assessment of health-related quality of life in arthritis: conceptualization and development of five item banks using item response theory. *Health Qual Life Outcomes*. 2006;4:33. doi:10.1186/1477-7525-4-33
32. Kopec JA, Russell L, Sayre EC, Rahman MM. How to measure the burden of symptoms that fluctuate over time? *Qual Life Res*. 2016;25(Supp. 1):174.
33. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand*. 1983;67(6):361–370. doi:10.1111/j.1600-0447.1983.tb09716.x
34. Montgomery N, Howell D, Ismail Z, et al. Cancer Care Ontario Patient Reported Outcome Advisory Committee. Selecting, implementing and evaluating patient-reported outcome measures for routine clinical use in cancer: the Cancer Care Ontario approach. *J Patient-Rep Outcomes*. 2020;4(1):101. doi:10.1186/s41687-020-00270-1
35. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34–42. doi:10.1016/j.jclinepi.2006.03.012
36. Lindquist KA, MacCormack JK, Shablack H. The role of language in emotion: predictions from psychological constructionism. *Front Psychol*. 2015;14(6):444.
37. Gescheider G. *Psychophysics: The Fundamentals*. 3rd ed. Lawrence Erlbaum Associates, Inc; 1997.