ARTICLE

Assessing the Amount of Change in an Outcome Measure Is Not the Same as Assessing the Importance of Change

Paul W. Stratford, PT, MSc; Daniel L. Riddle, PT, PhD, FAPTA†*

ABSTRACT

Purpose: To determine whether a difference exists between patients' self-ratings of amount of change and their self-ratings of importance of change. ***Methods:*** Eighty-eight patients receiving treatment of low-back pain completed two global rating of change (GRC) scales 4 to 6 weeks after their initial assessments. The scales were similar in format, differing only in that one asked respondents about the amount of change and the other about the importance of change. ***Results:*** Our analysis was restricted to 86 patients who reported improvement or no change. The chance-corrected agreement between patients' self-ratings of amount of change and their self-ratings of importance of change was low ($\kappa = 0.35$; 95% CI, 0.23–0.48). Of 47 disagreements, 44 reported a greater importance of change than amount of change and 3 reported a greater amount of change than importance of change. ***Conclusions:*** Assessing the amount of change is not the same as assessing the importance of change. When the goal is to estimate important change, the reference standard should direct patients to judge the importance of the change.

Key Words: back pain; evaluation research; outcome assessment; validation studies.

RÉSUMÉ

Objectif : Établir s'il existe une différence entre l'autoévaluation de la somme des changements par le patient et l'importance du changement. ***Méthodologie :*** Un échantillon de 88 personnes recevant des traitements pour une lombalgie a réalisé deux évaluations à l'aide d'une échelle d'évaluation globale de changement (EGC) de 4 à 6 semaines après leur évaluation initiale. Les échelles avaient une forme similaire, à ceci près que l'une visait à évaluer la somme des changements et l'autre, l'importance du changement. ***Résultats :*** Notre analyse s'est limitée à 86 patients qui ont fait part d'améliorations ou qui n'avaient constaté aucun changement. Les écarts corrigés entre l'autoévaluation de la somme des changements et de l'importance du changement par le patient étaient faibles ($\kappa = 0{,}35$; CI 95 %, CI 0,23–0,48). Dans 47 cas d'évaluations discordantes, 44 ont signalé un changement d'une plus grande importance que la somme des changements et dans 3 cas, la somme des changements était plus grand que l'importance du changement. ***Conclusions :*** Évaluer la somme des changements n'est pas la même chose qu'évaluer l'importance du changement. Lorsque l'objectif est d'évaluer un changement important, la norme de référence devrait amener les patients à mesurer l'importance du changement.

The past several decades have seen a substantial increase in the number of patient-reported outcome measures (PROMs).[1–5] Unlike scores from more traditional clinical measures, which have intuitive meaning to clinicians, PROM scores are often difficult to interpret. Investigators have therefore sought to supplement validation studies with investigations aimed at enhancing the interpretability of PROM scores and change scores. Specifically, there has been great interest in identifying the magnitude of a clinically important change (also referred to as the minimal clinically important change, MCIC; minimal clinically important difference, MCID; or minimal clinically important improvement, MCII). In many cases, efforts to estimate this quantity are complicated by the fact that no true gold standard exists for the attribute or characteristic of interest (e.g., health-related quality of life or functional status). The lack of a gold standard for PROMs has resulted in the application of several reference standards of change, including the retrospective global rating of change (GRC),[1] the prognostic rating of change,[6] and comparison with another measure's change scores.[7] Of these reference standards, the GRC is reported most frequently. However, many investigations have equated amount of change with importance of change, without inquiring as to whether the reported change was important.[4,7–11] This practice seems

From the: *School of Rehabilitation Science and Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont.; †Departments of Physical Therapy and Orthopaedic Surgery, Virginia Commonwealth University, Richmond, Va., USA.

Correspondence to: Paul W. Stratford, School of Rehabilitation Science, Institute for Applied Health Sciences, McMaster University, 1400 Main St. W., Rm 430, Hamilton, ON L8S 1C7; stratfor@mcmaster.ca.

Contributors: Both authors designed the study, collected the data, and analyzed and interpreted the data; drafted or critically revised the article; and approved the final draft.

Competing interests: Daniel L. Riddle is the recipient of a grant from the National Institutes of Health and is Deputy Editor in Chief of *Physical Therapy*.

Physiotherapy Canada **2013; 65(3);244–247; doi:10.3138/ptc.2012-16**

to have gone unnoticed: we found no investigations examining whether GRC ratings of amount of change and GRC ratings of importance of change are equivalent. In this study, therefore, we examine the relationship between patients' perceptions of amount of change and their perceptions of the importance of improvement.

In the form first popularized by Jaeschke and colleagues (1989),[1] the GRC had two steps. The first step asked whether the patient was better, about the same, or worse; if the response was *better* or *worse*, the second step asked the patient to rate how much better or worse on a 7-point scale. Effectively, then, this scale had 15 points, ranging from −7 to 7. Previous work by Jaeschke (reference not provided in Jaeschke and colleagues[1]) suggests that a change rating of ≥3 points on the GRC is associated with a clinically important change in the patient's condition. Since Jaeschke and colleagues' initial report, various GRC scales, with differing numbers of scale points and different descriptors assigned to those points, have frequently been applied as reference standards for estimating MCIC or MCII. However, many of the modified GRCs used to estimate MCIC or MCII ask about only the amount of change and do not inquire about the importance of that change.[4,7–14]

The purpose of our study was to determine whether a difference exists between patients' ratings of amount of change and their ratings of importance of change. Although there have been challenges to the validity of the GRC,[7,15] our intent in this study was to comment on the relationship between estimates of amount of change and estimates of importance of change when the GRC is used, not to critique the appropriateness of GRC as a reference standard for change.

METHODS

Study design

This investigation represents a secondary analysis of data gathered as part of a study reported elsewhere, for which ethics approval was obtained.[16] The purpose of the original study was to compare the ability of three low-back pain (LBP) questionnaires to assess change over time.[16] Clients completed the questionnaires at their initial physical therapy visit and following 4 to 6 weeks of treatment. At the 4- to 6-week follow-up assessment, clients also completed two GRC scales. One scale inquired about amount of change; the other asked about the importance of change.

Measures: Global ratings of change

The "amount of change" scale was similar to the 15-point scale reported by Jaeschke and colleagues.[1] This instrument asked respondents whether they were better, about the same, or worse;[16] if the response was *better* or *worse*, respondents were asked to choose one of the following terms to describe how much they had improved or deteriorated: (1) a tiny bit, almost the same; (2) a little bit; (3) somewhat; (4) moderately; (5) quite a bit; (6) a great deal; (7) a very great deal. Effectively this created a 15-point scale (7 levels of deterioration, 7 levels of improvement, and 1 level representing no change). The "importance of change" scale asked clients to rate the extent to which the change was important, using a similar spectrum of descriptive terms: (1) a tiny bit important; (2) a little bit important; (3) somewhat important; (4) moderately important; (5) quite important; (6) a great deal important; (7) a very great deal important.

Thus, the 15-point "importance of change" scale was identical to the "amount of change scale" except that, rather than asking about the amount of change, it asked about the importance of change. The "amount of change" scale was completed first, followed by the "importance of change" scale.

Data analysis

We calculated kappa as a representation of chance-corrected agreement for participants' responses to the GRCs. We chose to apply kappa (κ) rather than weighted kappa (κ_w) in our analysis because of inconsistencies in the literature in the number of response options on amount of change scales,[4,8,11,13] the descriptors accompanying the response options,[4,8,11] and the choice of cut-point used to classify improvement as important or not important.[4,11,17] Given these uncertainties, we believe that kappa provides a better impression of the consequence of not directly asking whether the change is important, because it is unclear at what demarcation point partial credit should be assigned, if at all. We applied Bowker's test for symmetry to examine whether the patterns of disagreements above and below the main agreement diagonal (see Table 1) were similar.[18] Based on chance alone, one would not expect these disagreements to differ significantly; differences were considered statistically significant at $p < 0.05$ (2-tailed). Parameter estimates were accompanied by 95% CIs; all analyses were performed using STATA 12.1 (StataCorp, College Station, TX).

RESULTS

Participant characteristics

A total of 88 clients referred for physical therapy with non-specific LBP took part in the original investigation. The sample's mean age was 41 (SD 11.6) years; mean duration of symptoms was 48 (SD 36) days. Of the 88 participants, 76 had work-related injuries. Because only two participants reported worsening on both the "amount of change" and "importance of change" scales, a sample size too small to allow comment on deterioration, data for these participants were deleted from subsequent analyses; the discussion below therefore focuses on the relationship between amount of improvement and importance of improvement.

GRC score comparison

Table 1 summarizes responses to the "amount of improvement" and "importance of improvement" global

Table 1 Responses to "Amount of Improvement" and "Importance of Improvement" Global Rating of Change Questions (No. of Responses)

		Importance of improvement								
		0	1	2	3	4	5	6	7	Total
Amount of improvement	0	**6**	0	0	0	0	0	0	0	6
	1	0	**1**	0	0	0	0	0	0	1
	2	0	0	**5**	1	0	2	1	0	9
	3	0	0	0	**3**	2	1	1	0	7
	4	0	0	0	0	**3**	0	4	1	8
	5	0	0	0	0	3	**7**	14	6	30
	6	0	0	0	0	0	0	**7**	11	8
	7	0	0	0	0	0	0	0	**7**	**7**
	Total	6	1	5	4	8	10	27	25	86

1 = a tiny bit improved/important; 2 = a little bit improved/important;
3 = somewhat improved/important; 4 = moderately improved/important;
5 = quite improved/important; 6 = a great deal improved/important;
7 = a very great deal improved/important.

ratings. The main diagonal displays the agreement frequencies between the two scales ($\kappa = 0.35$; 95% CI, 0.23–0.48). Cell values above the main diagonal indicate the number of clients who rated the importance of improvement higher than the amount of improvement; cell values below the main diagonal represent the number of clients who rated the amount of improvement higher than the importance of improvement. Of the 47 discordant cell values, 44 show higher ratings for importance of improvement than for amount of improvement. Bowker's test reveals that this disagreement pattern represents a statistically significant asymmetry ($\chi^2_{12} = 47.0$, $p < 0.001$).

DISCUSSION

The goal of our study was to examine the agreement between clients' ratings of amount of change and their ratings of the importance of change. Our point estimate of kappa was 0.35; our results also show that when disagreements occurred, ratings of importance of improvement were higher than ratings of amount of improvement, and that this was particularly evident for amount-of-change ratings >4. The consequences of these findings are potentially important: studies using GRC scales that inquire only about the magnitude of change, without assessing its importance, may underestimate MCII relative to studies that include an assessment of importance of change. Since knowledge of both amount of change and importance of change is essential to clinical decision making, and one is not a surrogate measure for the other, both should be assessed.

LIMITATIONS

Our study has several limitations. First, our "importance of change" scale mirrored the format of the 15-point "amount of change" scale; although this allowed a direct comparison between scales, it does not address the dilemma of how important is important enough. For example, is "somewhat important" important enough to be considered "important"? Second, because clients were always asked about the importance of change after being asked about the amount, we do not know whether this affected our results. Third, the interpretation of our results is specific not only to the response format of the applied scales but also to the clients and setting. Messick reminds us that reliability and validity are not properties of a measure but of a measure's scores: that is, they exist in a context.[19(p.14)] Thus, we do not know the extent to which our findings are generalizable to other scale formats, conditions, or clinical settings. Finally, our study examined the relationship of responses for improvement ratings only; small sample size prevented us from considering responses from clients who reported deterioration.

CONCLUSION

Our findings show that patients' ratings of amount of change cannot be used interchangeably with ratings of importance of change. When the goal is to estimate important change, our data indicate that a reference standard should be used that directly asks patients/clients to judge the importance of the change.

KEY MESSAGES

What is already known on this topic

Investigators have often equated a rating of *amount of change* with a rating of the *importance of change* without justifying this assumption.

What this study adds

Our findings show that an *amount of change* rating is not interchangeable with an *importance of change* rating.

REFERENCES

1. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials. 1989;10(4):407–15. http://dx.doi.org/10.1016/0197-2456(89)90005-6. Medline:2691207
2. Kopec JA, Esdaile JM, Abrahamowicz M, et al. The Quebec Back Pain Disability Scale: measurement properties. Spine. 1995;20(3):341–52. http://dx.doi.org/10.1097/00007632-199502000-00016. Medline:7732471
3. Roos EM, Lohmander LS. The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis. Health Qual Life Outcomes. 2003;1(1):64. http://dx.doi.org/10.1186/1477-7525-1-64. Medline:14613558
4. de Morton NA, Davidson M, Keating JL. Validity, responsiveness and the minimal clinically important difference for the de Morton Mobility Index (DEMMI) in an older acute medical population. BMC Geriatr. 2010;10(1):72. http://dx.doi.org/10.1186/1471-2318-10-72. Medline:20920285
5. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine. 1983;8(2):141–4. http://dx.doi.org/10.1097/00007632-198303000-00004. Medline:6222486

6. Westaway MD, Stratford PW, Binkley JM. The Patient-Specific Functional Scale: validation of its use in persons with neck dysfunction. J Orthop Sports Phys Ther. 1998;27(5):331–8. Medline:9580892
7. Fletcher KE, French CT, Irwin RS, et al. A prospective global measure, the Punum Ladder, provides more valid assessments of quality of life than a retrospective transition measure. J Clin Epidemiol. 2010;63(10):1123–31. http://dx.doi.org/10.1016/j.jclinepi.2009.09.015. Medline:20303709
8. Demoulin C, Ostelo R, Knottnerus JA, et al. Quebec Back Pain Disability Scale was responsive and showed reasonable interpretability after a multidisciplinary treatment. J Clin Epidemiol. 2010;63(11):1249–55. http://dx.doi.org/10.1016/j.jclinepi.2009.08.029. Medline:20400266
9. Dawson J, Doll H, Boller I, et al. Comparative responsiveness and minimal change for the Oxford Elbow Score following surgery. Qual Life Res. 2008;17(10):1257–67. http://dx.doi.org/10.1007/s11136-008-9409-3. Medline:18958582
10. Las Hayas C, Quintana JM, Padierna JA, et al. Health-Related Quality of Life for Eating Disorders questionnaire version-2 was responsive 1-year after initial assessment. J Clin Epidemiol. 2007;60(8):825–33. http://dx.doi.org/10.1016/j.jclinepi.2006.10.004. Medline:17606179
11. Ornetti P, Dougados M, Paternotte S, et al. Validation of a numerical rating scale to assess functional impairment in hip and knee osteoarthritis: comparison with the WOMAC function scale. Ann Rheum Dis. 2011;70(5):740–6. http://dx.doi.org/10.1136/ard.2010.135483. Medline:21149497
12. Terwee CB, Roorda LD, Knol DL, et al. Linking measurement error to minimal important change of patient-reported outcomes. J Clin Epidemiol. 2009;62(10):1062–7. http://dx.doi.org/10.1016/j.jclinepi.2008.10.011. Medline:19230609
13. Young BA, Walker MJ, Strunce JB, et al. Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. Spine J. 2009;9(10):802–8. http://dx.doi.org/10.1016/j.spinee.2009.06.002. Medline:19632904
14. Wright AA, Cook CE, Baxter GD, et al. A comparison of 3 methodological approaches to defining major clinically important improvement of 4 performance measures in patients with hip osteoarthritis. J Orthop Sports Phys Ther. 2011;41(5):319–27. Medline:21335930
15. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 1997;50(8):869–79. http://dx.doi.org/10.1016/S0895-4356(97)00097-8. Medline:9291871
16. Stratford PW, Binkley J, Solomon P, et al. Assessing change over time in patients with low back pain. Phys Ther. 1994;74(6):528–33. Medline:8197239
17. Hayes KW, Petersen C, Falconer J. An examination of Cyriax's passive motion tests with patients having osteoarthritis of the knee. Phys Ther. 1994;74(8):697–707, discussion 707–9. Medline:8047559
18. Bowker AH. A test for symmetry in contingency tables. J Am Stat Assoc. 1948;43(244):572–4. http://dx.doi.org/10.1080/01621459.1948.10483284. Medline:18123073
19. Messick S. Validity. In: Linn RL, editor. Educational measurement. 3rd ed. Phoenix (AZ): Oryx Press; 1993. p. 13–104.