Eur Spine J (2010) 19:957–962
DOI 10.1007/s00586-010-1289-x

ORIGINAL ARTICLE

A study of the test–retest reliability of the self-perceived general recovery and self-perceived change in neck pain questions in patients with recent whiplash-associated disorders

Trung Ngo · Maja Stupar · Pierre Côté · Eleanor Boyle · Heather Shearer

Received: 30 September 2009 / Revised: 14 December 2009 / Accepted: 15 January 2010 / Published online: 4 February 2010

Abstract The objectives of this study were to determine the test–retest reliability of two self-perceived recovery questions in patients with recent whiplash-associated disorders (WAD), and to assess whether remembering previous answers influences reliability. The self-perceived general recovery and self-perceived change in neck pain questions were administered to 46 patients with recent WAD 6 weeks after recruitment and again 3–5 days later. At follow-up, we also asked participants if they remembered their previous answers. We used the intra-class correlation coefficients (ICC) to measure the reliability of the original ordinal response structure and kappa statistics for dichotomized responses. The ICC [95% confidence intervals (CI)] for the general recovery and for the change in neck pain questions were 0.70 (0.60–0.80) and 0.80 (0.72–0.87), respectively. The kappa statistic (95% CI) for the general recovery question was 0.81 (0.64–0.99) when recovery was defined as "completely better" or "much improved". The kappa statistic (95% CI) for the change in neck pain question was 0.80 (0.62–0.99) when recovery was defined as "very much better" or "better". Our analysis suggests that the test–retest reliability may be higher for participants who remembered their previous responses. In conclusion, our results suggest that self-perceived recovery questions have adequate reliability for use in epidemiological research of WAD.

Keywords Self-perceived recovery · Whiplash-associated disorders · Test–retest reliability · Patient outcomes · Outcomes research

T. Ngo · M. Stupar · P. Côté · E. Boyle · H. Shearer
Toronto Western Research Institute, University Health Network, Toronto, Canada

T. Ngo · P. Côté
Canadian Memorial Chiropractic College, Toronto, Canada

M. Stupar · P. Côté
Department of Health Policy Management and Evaluation, Faculty of Medicine, University of Toronto, Toronto, Canada

P. Côté · E. Boyle
Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

P. Côté (✉)
Toronto Western Hospital, Med West Building, 2nd Floor, Rm 320, Box 36, 750 Dundas St. West, Toronto, ON, Canada M6J 3S3
e-mail: pcote@uhnresearch.ca

Introduction

Epidemiological studies rely on self-report outcome measures to track the recovery of individuals with whiplash-associated disorders (WAD). A method that is increasingly used to follow the progress of these individuals consists of asking them to provide an assessment of self-perceived recovery or self-perceived change in neck pain and disability [1–4]. These measures provide clear advantages over other outcome measures. First, they resemble the evaluation method used in clinical practice and therefore provide information that is relevant to clinicians [4]. Second, these questions allow for patients to provide a personal appraisal of their recovery. This is important because recovering from musculoskeletal injuries is an individualized process that is dependant on a patient's assessment of her/his progress [5]. According to Beaton et al., recovery can take on one of three meanings: (1) resolution of symptoms, (2) readjustment of life to participate in daily activities, or (3) redefining the meaning of health to enable

one to participate in daily activities. However, self-perceived recovery measures are liable to measurement error and may be sources of misclassification bias in epidemiological studies. To date, only one study has documented the test–retest reliability of the self-perceived change in pain and disability questions [4]. In a sample of patients with arthritis, Fischer documented that the 1-week test–retest correlation was 0.81 for the self-perceived change in disability and 0.58 for the self-perceived change in pain. This study of arthritic patients demonstrates that while misclassification occurs, its magnitude varies with the construct that is being measured.

Despite their use in studies of WAD, we know little about the reliability of the self-perceived recovery and change questions in this population. This is problematic because the probability of misclassification bias is directly related to the reliability of a measure [6]. Epidemiological studies that use these questions are therefore at risk of estimating effect sizes that are biased toward, or away from the null depending on the type misclassification (non-differential vs. differential) present in a study. Moreover, the self-perceived recovery questions have been used as an external criterion to establish the validity of other self-report measures of health status in patients with whiplash injuries [7]. Using an external criterion that is prone to measurement error can lead to erroneous conclusions about the validity of health outcome measures.

The primary objective of this study was to determine the test–retest reliability of the self-perceived general recovery and self-perceived change in neck pain questions in a sample of patients with recent WAD. Our secondary objectives were to determine the reliability of dichotomized response options to these questions and to assess the impact of memory on their reliability.

Materials and methods

Design and study sample

We conducted a test–retest reliability study. Individuals who were involved in traffic collisions and made an automobile insurance claim to AVIVA Canada between January 2008 and December 2008 were eligible for the study. We included individuals who were: (1) at least 18 years of age; (2) resident of the Greater Toronto area, Mississauga, Burlington, Cambridge or Kitchener; (3) and diagnosed with WAD grade I, II, or III [8] by a study coordinator within 3 weeks of their traffic collision. Potential participants were excluded if they were unable to provide written informed consent, unable to complete the interview in English, had WAD grade IV injuries (fracture or dislocation of the cervical spine), or if they had a history of neck surgery.

Self-perceived recovery questions

We studied the test–retest reliability of two self-perceived recovery questions. The first question addressed general recovery from all injuries sustained during the traffic collision: "How well do you feel you are recovering from your injuries?" [1, 3]. Participants were asked to rate their overall recovery on a seven-item Likert scale that included the following choices: completely better, much improved, slightly improved, no change, slightly worse, much worse or worse than ever. The second question focused on change in neck pain: "How do you feel your neck pain has changed since the injury?". Participants were asked to rate the change in their neck pain on a seven-item Likert scale that included the following choices: very much better, better, slightly better, no change, slightly worse, worse, or very much worse.

Study procedures

Participants recruited for this study were first invited to participate in a randomized controlled trial of the effectiveness of three standard treatments for whiplash injuries in Ontario [2]. After consenting to participate in the trial, participants were administered a questionnaire that included questions on demographics, pain level, ability to function, general health, mood, working status and health care utilization [2].

Six weeks after completing the initial questionnaire, participants were interviewed in person or by telephone by a follow-up interviewer who was blind to their initial health status and administered the change in neck pain and general recovery questions. Three to five days following the administration of the 6-week follow-up, participants were contacted again by telephone by a different interviewer and were re-administered the two questions. A minimal interval of 3 days was selected because it minimizes recollection bias when studying conditions that fluctuate in time [9].

We randomly assigned the order of administration of the questions at the second interview to minimize sequence bias. Simple randomization with mixed block size was used to determine the order of question administration. We also collected data to assess the effect of memory on reliability. Participants were asked whether they recalled the answers to the questions that were administered 3 days earlier. Specifically, we asked participants: "Do you remember being asked some of these questions 3 days ago?" and "Do you remember your answers to the questions?".

The study was approved by the Review Ethics Boards of the University Health Network and of the Canadian Memorial Chiropractic College.

Statistical analysis

We estimated that a sample size of 46 subjects was required to measure an Intra-class Correlation Coefficient (ICC) of 0.9, with a power of 0.8 at a significance level of 0.05 [10]. The ICC and 95% confidence intervals (CI) were used to determine the test–retest reliability of the two versions of the perceived recovery questions using their original seven-item responses. The ICCs were computed using Model II for multiple-raters [6, 11, 12]. We also computed weighted kappa coefficients and 95% CI using quadratic weights to determine whether the distribution of responses influenced the reliability as measured by the ICC.

We assessed the reliability of dichotomizing the seven-item responses into "recovered" versus not "recovered" [1, 3]. We defined "recovered" as "completely better" on the general recovery question and as "very much better" on the change in neck pain question. Participants classified as "not recovered" answered any of the other response options. We tested whether varying thresholds for classifying a patient as "recovered" impacted reliability. Therefore, in a secondary analysis, we defined "recovered" as "completely better" or "much improved" on the general recovery question and as "very much better" or "better" on the change in neck pain question. Those categorized as "not recovered" answered any of the remaining response options.

Finally, we conducted secondary analyses to determine if memory had an impact on reliability by stratifying participants according to their answers to the recall questions. All statistical analyses were performed using SAS [13] and STATA [14].

Results

Sample characteristics

Our sample included 46 participants. Three participants were not enrolled in the randomized controlled trial. The mean age of our sample was 43.2 years and 71.7% were female (Table 1). Most participants (69.6%) had grade II WAD. The mean whiplash disability questionnaire and pain intensity scores decreased between enrollment and the 6 weeks follow-up (Table 1). At enrollment, no participants had hired a lawyer to assist them with their claim. However, 6 weeks later, three participants had hired a lawyer (one participant refused to answer the question). Answers to the global recovery and to the change in neck pain questions suggest that most participants reported improvement 6 weeks after their injury.

Test–retest reliability

The ICCs (95% CI) for the general recovery and for the change in neck pain questions were 0.70 (0.60–0.80) and 0.80 (0.72–0.87), respectively. The weighted kappa statistics were similar at 0.70 (0.42–0.98) and 0.80 (0.51–1.0) for the general recovery and for the change in neck pain questions, respectively.

Reliability of dichotomized response options

The kappa statistic (95% CI) for the general recovery question was 0.85 (0.64–1) when "recovered" was defined "completely better" and 0.81 (0.64–0.99) when defined as "completely better" or "much improved". For the change in neck pain question, the kappa statistic (95% CI) was 0.46 (0.20–0.74) when "recovered" was defined as "very much better" and 0.80 (0.62–0.99) when defined as "very much better" or "better".

Test–retest reliability stratified by responses to the recall question

Twelve subjects were not asked the memory questions because we started administering these questions shortly after the onset of the study. We stratified study participants according to their responses to the recall questions (remember vs. do not remember previous answers). Eighteen participants remembered their previous responses and 16 did not remember (Table 2). The kappa coefficient was 1 for participants who remembered their previous answers to the change in neck pain question; 0.74 (0.41–1) for those who did not remember and 0.66 (0.22–1) for participants who were not asked the question. Similarly, the kappa coefficient was 1 for participants who remembered their previous answers to the general recovery question; 0.88 (0.64–1) for those who did not remember and 0.50 (0.02–0.98) for participants who were not asked the question.

Discussion

The primary objective of our study was to measure the test–retest reliability of two questions used to assess recovery in patients with WAD. We found that the change in neck pain question (ICC = 0.80; 95% CI 0.72–0.87) and the general recovery question (ICC = 0.70; 95% CI 0.60–0.80) had similar reliability when the original response structure of the questionnaire (seven items) was analyzed.

Table 1 Characteristics of participants at enrollment and 6-week follow-up ($N = 46$)

Characteristic	Enrollment	6-week follow-up
Female [*n* (%)]	33 (71.7)	
Age (years)		
Mean (SD); range	43.2 (13.7); 19.6–73.5	
Time since injury (days)		
Mean (SD); range	5.7 (4.7); 0–19	
WAD grade		
I [*n* (%)]	14 (30.4)	
II [*n* (%)]	32 (69.6)	
Highest level of education [*n* (%)]		
High school or less	8 (17.4)	
Post-secondary or some university	12 (26.1)	
Technical school graduate	8 (17.4)	
University graduate	18 (39.1)	
Income [*n* (%)]		
$0–$49,999	24 (52.2)	
$50,000–$59,999	8 (17.4)	
$60,000–$79,999	4 (8.7)	
$80,000+	8 (17.4)	
Did not respond	2 (4.4)	
Lawyer involvement in the claim (%)	0	3 (8.7)[b]
Pain intensity, Mean (SD)[a]		
Neck	5.6 (2.1)	2.9 (2.7)
Shoulder	3.9 (3.0)	2.5 (2.9)
Low back	3.4 (3.2)	2.5 (3.1)
Headache	3.3 (3.2)	1.8 (2.5)
Arm	2.1 (2.7)	1.7 (2.7)
WDQ score, Mean (SD)	47.5 (29.5)	31.2 (29.8)
General recovery question [*n* (%)]		
Completely better		8 (17.4)
Much improved		20 (43.5)
Slightly improved		13 (28.3)
No change		3 (6.5)
Slightly worse		0
Much worse		1 (2.2)
Worse than ever		1 (2.2)
Change in neck pain question [*n* (%)]		
Very much better		16 (34.8)
Better		14 (30.4)
Slightly better		13 (28.3)
No change		2 (4.3)
Slightly worse		1 (2.2)
Worse		0
Very much worse		0

SD standard deviation

[a] Numeric rating scale of 0–10 (0 no pain and 10 worst pain ever)

[b] One participant refused to answer the question; therefore, the denominator for the proportion is 45

While these levels of reliability are commonly considered adequate, they nevertheless indicate that their use results in misclassification. The difference in reliability between the two questions may be related to the construct being measured. It may be easier to repeatedly appraise one's own status when referring to a specific construct such as pain than it is for a complex and multifactorial construct such as general recovery. It is also possible that the differences in reliability are associated with the response options provided for each of the questions. While both questions use a

Table 2 Reliability of dichotomized responses stratified by memory of previous responses

Remembered previous answers	*N*	Kappa (95% CI)
Yes		
General recovery	18	1
Change in neck pain	18	1
No		
General recovery	16	0.88 (0.64–1)
Change in neck pain	16	0.74 (0.41–1)
Not asked		
General recovery	12	0.50 (0.02–0.98)
Change in neck pain	12	0.66 (0.22–1)

CI confidence interval

seven-item Likert scale, the wording of the responses are different ranging from "completely better" to "worse than ever" on the general recovery question and "very much better" to "very much worse" on the change in neck pain question.

It is often necessary to discriminate between patients who have recovered from their injuries and those who have not. Consequently, by dichotomizing the response options, we stratified participants into "recovered" and "not recovered". Our analysis indicates that both questions are equally reliable when the first two response options are used to define recovered ["very much better" or "better" on the neck pain recovery question (kappa = 0.80) and "completely better" or "much improved" on the general recovery question (kappa = 0.81)]. This suggests that participants were consistent in reporting their status within one of the two categories. However, the reliability of the general recovery question (kappa = 0.85) slightly improved while the reliability of the change in neck pain question (kappa = 0.46) worsened when the first response option was used to classify participants as "recovered". While changing the threshold to categorize someone as recovered had little impact on the reliability of the general recovery question, it significantly reduced the reliability of the change in neck pain question. It is possible that this result is due to the daily fluctuations in pain intensity experienced by patients who have not experienced significant improvement in their neck pain. It is also possible that individuals are better at understanding and discriminating between the categories "completely better" and "much improved" than they are at discriminating between categories "very much better" and "better".

The validity of the ICC as a test–retest reliability statistic rests on the assumption that the data are normally distributed [15]. We found that our data were skewed with most respondents selecting the first three response options in both questions. No participants reported to be "slightly worse" on the general recovery question and none reported that their pain had changed and was now "worse" or "very much worse" on the neck pain question. We conducted secondary analyses to determine the impact of this distributional problem and found similar results when using the weighted kappa statistic. This suggests that these distributional violations did not bias the results and that the tests are robust against violations of normality [15, 16].

A 3-day retest interval is short and may lead to subjects remembering their responses to the previous administration of the questions. We assessed the impact of memory and found, as expected, that participants who remembered their answers had perfect reliability. Interestingly, the reliability of both questions in participants who did not remember their previous responses was also high. However, the reliability was lower in participants who were not asked the memory questions. The results of our secondary analysis suggest that the favorable reliability statistics obtained in this study are in part related to a memory effect.

Our study has limitations. First, to measure test–retest reliability, we would optimally need a sample of whiplash patients with a stable condition [6]. This is not possible when studying recent whiplash injuries [17]. Therefore, it is likely that our participants still experienced daily fluctuations in their symptoms and in their general status. This could have a negative effect on the reliability of the questions. As previously mentioned, a second limitation of our study is the impact of a memory effect due to the short 3-day retest interval. Third, our sample size of 46 was too small to include several participants who got worse during the study period. Therefore, it is possible that our reliability estimates would have been different if we had enrolled a higher proportion of participants who did not improve. Fourth, the answers to the two questions may have varied because the change in neck pain question included a specific time reference (since the injury) while the global recovery question did not. Finally, the study sample included participants who were enrolled in a randomized controlled trial. Therefore, our findings may not be generalizable to the general population of individuals with WAD.

Conclusion

Our study suggests that the self-perceived general recovery and self-perceived change in neck pain questions have adequate reliability for use in epidemiological research of WAD. From a test–retest reliability perspective, researchers have the option to define "recovery" as "completely better" or "completely better" or "much improved" using the general recovery question. Defining "recovery" with

the self-perceived change in neck pain question should only be done using "very much better" or "better" instead of "very much better".

Acknowledgments This study was funded by an industry grant from AVIVA Canada to the University Health Network. We thank Dr. Marion McGregor and David Soave from the Canadian Memorial Chiropractic College for their assistance with the study.

References

1. Cassidy JD, Carroll LJ, Côté P, Frank J (2007) Does multidisciplinary rehabilitation benefit whiplash recovery? Results of a population-based incidence cohort study. Spine 32:126–131
2. Côté P, Cassidy JD, Carette S, Boyle E, Shearer HM, Stupar M, Ammendolia C, van der Velde G, Hayden JA, Yang X, van Tulder M, Frank JW (2008) Protocol of a randomized controlled trial of the effectiveness of physician education and activation versus two rehabilitation programs for the treatment of whiplash-associated disorders. The University Health Network Whiplash Intervention Trial. Trials 9:75
3. Ferrari R, Rowe BH, Majumdar SR, Cassidy JD, Blitz S, Wright SC, Russell AS (2005) Simple educational intervention to improve the recovery from acute whiplash: results of a randomized, controlled trial. Acad Emerg Med 12:699–706
4. Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H (1999) Capturing the patient's view of change as a clinical outcome measure. JAMA 282(12):1157–1162
5. Beaton DE, Tarasuk V, Katz JN, Wright JG, Bombardier C (2001) "Are you better?" A qualitative study of the meaning of recovery. Arthritis Rheum 45(3):270–279
6. Streiner DL, Norman GR (2003) Health measurement scales: a practical guide to their development and use, 3rd edn. Oxford University Press, Toronto
7. Stewart M, Maher CG, Refshauge KM, Bogduk N, Nicholas M (2007) Responsiveness of pain and disability measures for chronic whiplash. Spine 32(5):580–585
8. Spitzer WO, Skovron ML, Salmi LR et al (1995) Scientific monograph of the Quebec task force on whiplash-associated disorders: redefining "whiplash" and its management. Spine 20:1S–73S
9. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF (2003) A comparison of two time intervals for test-retest reliability of health status instruments. J Clin Epidemiol 56:730–735
10. Walter SD, Eliasziw M, Donner A (1998) Sample size and optimal designs for reliability studies. Stat Med 17:101–110
11. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420–428
12. Portney LG (2000) Foundations of clinical research: applications to practice, 2nd edn. Prentice Hall Health, Upper Saddle River
13. SAS 9.1 for Windows, SAS Institute Inc., Cary
14. Stata/sE 9.2 for windows. StataCorp LP, College Station
15. Gardner PL (1975) Scales and statistics. Rev Educ Res; 45:43–57
16. Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Measurement 33:613–619
17. Carroll LJ, Holm LW, Hogg-Johnson S et al (2008) Course and prognostic factors for neck pain in whiplash-associated disorders (WAD): results of the bone and joint decade 2000–2010 task force on neck pain and its associated disorders. Spine 33:S83–S92