



# Psychometric properties of the Swedish version of the Treatment Outcome Satisfaction Questionnaire

Riccardo Lo Martire<sup>1,2</sup>  · Angela Lis<sup>3,5</sup> · Eva Skillgate<sup>4</sup> · Eva Rasmussen-Barr<sup>1,4</sup> 

Received: 27 June 2016/Revised: 17 October 2016/Accepted: 11 November 2016/Published online: 25 November 2016  
© The Author(s) 2016. This article is published with open access at Springerlink.com

## Abstract

**Purpose** Patient satisfaction is an outcome measure for low-back pain (LBP) interventions which allows clinicians to design patient-oriented treatments. The Treatment Outcome Satisfaction Questionnaire (TOSQ) is an English instrument constructed for such evaluations, and no equivalent instruments exist for the Swedish population. This study, therefore, translated TOSQ into Swedish and assessed the translated version's psychometric properties for patients with LBP.

**Methods** A cross-cultural adaptation was used to translate TOSQ into Swedish. Subsequently, data from 131 patients with LBP whom undergone physiotherapy were consecutively aggregated and analyzed in a Rasch rating scale model with person measures standardized at 0–100 logits to evaluate the translated scale's validity. Finally, test-retest reliability of the Swedish version of TOSQ (TOSQ-

S) was quantified via an intraclass correlation coefficient (ICC) and the standard error of measurement (SEM) in 41 patients.

**Results** TOSQ was successfully translated into Swedish; however, while some Rasch model indices supported the translated scale's unidimensionality, one out of eight items and 12 out of 131 subjects misfitted the model. Scale optimization resulted in a 6-item subconfiguration, for which all items fitted the model, person misfits were reduced to ten subjects, and the person separation index increased from 1.86 to 2.04. ICC and SEM estimates suggested acceptable reliability for the six-item TOSQ-S at 0.66 and 6.6 logits, respectively.

**Conclusions** A six-item TOSQ-S configuration showed acceptable psychometric properties and is suitable for measuring treatment outcome satisfaction of physiotherapy in patients with LBP.

✉ Riccardo Lo Martire  
lomartire@kth.se

<sup>1</sup> Division of Physiotherapy, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Alfred Nobels allé 23, Huddinge 141 83, Sweden

<sup>2</sup> Centre for Naval Architecture, Department of Aeronautical and Vehicle Engineering, KTH Royal Institute of Technology, Teknikringen 8, Stockholm 100 44, Sweden

<sup>3</sup> Occupational and Industrial Orthopaedic Center (OIOC), Hospital for Joint Diseases, New York University Langone Medical Center, 63 Downing Street, New York 100 14, NY, USA

<sup>4</sup> Musculoskeletal & Sports Injury Epidemiology Center, Institute of Environmental Medicine, Karolinska Institutet, Box 210, Stockholm 171 77, Sweden

<sup>5</sup> Department of Physical Therapy, Seton Hall University, 400 South Orange Ave, South Orange 070 89, NJ, USA

**Keywords** Low-back pain · Physiotherapy · Rasch analysis · Reliability · Validity

## Introduction

Low-back pain (LBP) is the single leading cause for disability globally [1, 2] and, consequently, one of the most common reasons for seeking healthcare in the western world [3]. Physiotherapy is one of the main conservative LBP treatments, and intervention results are regularly evaluated by patient-reported outcome measures targeted at several domains; patient satisfaction being one [4, 5].

Patient satisfaction is a multidimensional concept constituting of the satisfaction with the physical environment, the patient-care provider interaction, and the treatment outcome [6, 7]. Of these, the last in principle reflects the

patient's attitude towards the health care intervention [8, 9]. It is thus essential, as it incorporates the patient's view into everyday clinical practice [10], and allows the clinician to design patient-oriented interventions [5]. Hitherto, treatment outcome satisfaction of LBP interventions has typically either been assessed by global items [11–13] or instruments intended for other purposes [14]. The absence of psychometrically sound instruments tuned for patients with LBP provides a conceivable clarification as to why.

In 2010, the 'Treatment Outcome Satisfaction Questionnaire' (TOSQ) was developed to assess outcome satisfaction in patients with LBP following physiotherapeutic interventions [15]. TOSQ underwent a meticulous development process during which construct validity was examined based on a sample of physiotherapy outpatients in the New York metropolitan area [15]. To allow integration of TOSQ into the Swedish primary health care, a Swedish translation of the instrument is necessary. Whereas cross-cultural adaptation is suitable for this purpose, incorporating both language and cultural differences into the translation process, the translation nevertheless risks influencing the TOSQ properties considerably, and rendering it necessary to validate the translated version [16].

Traditionally, approaches based on classical test theory were used to assess psychometric properties of single-construct instruments; however, such methods are subject to two major limitations: they do not consider item hierarchy and they assume additivity of rating scale data [17]. Rasch analysis provides an alternative that is not limited by these assumptions, since it transforms ordinal item scores to log-odds (logits) interval data, used to determine the relation between person ability and item requirement in the measured construct [18]. In addition, it is necessary to quantify the test-retest reliability of the translated TOSQ items to determine result precision. Hence, this study translated TOSQ into Swedish and assessed the psychometric properties of the Swedish version of TOSQ (TOSQ-S) in patients with LBP.

## Methods

### Design and participants

In a two-step design, TOSQ was cross culturally adapted from English to Swedish, and TOSQ-S was subsequently psychometrically evaluated.

Following approval by the Regional Ethical Research Committee, data were consecutively aggregated from patients with LBP in conjunction with their rehabilitation at three specialist physiotherapy clinics in Stockholm, Sweden. Patients aged between 18 and 75 years seeking care for LBP as a primary complaint were included, and excluded if they did not understand Swedish. In total, 131 patients, 44

males and 87 females, with a mean (SD) age of 49 (14) years enrolled for participation. Rehabilitation comprised a median of seven occasions of physical training, manual therapy, and home exercises adapted to the patients' needs.

At rehabilitation completion, participants signed an informed consent, completed TOSQ-S without interference in a private room, and left the completed questionnaire in a sealed, coded envelope. Forty-one subjects also completed the questionnaire at home 1-week post-rehabilitation, and mailed the completed questionnaire in a coded envelope.

### Treatment Outcome Satisfaction Questionnaire

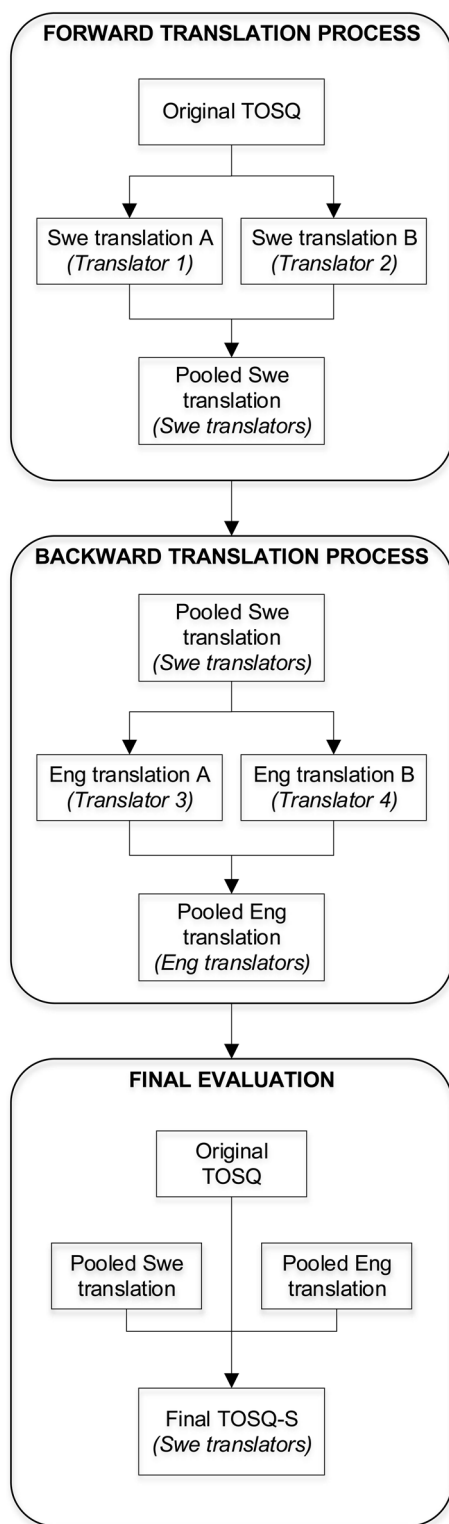
TOSQ assesses treatment outcome satisfaction via eight aspects of the patient's experience at the end of physiotherapy treatment. It reveals information about patients' understanding of their LBP, how they feel about their response to treatment and the improvement achieved, their perception of their ability to cope and/or carry on with their lives, the effort they placed in the rehabilitation process, and their ability to take care of themselves and control their LBP problem [15]. In total, TOSQ contains ten items scored on a seven-point Likert scale, of which two are global and eight are specific [15]. A summary score between 0 and 48 can be calculated for the latter [15].

### Cross-cultural adaptation

In accordance with published guidelines [16], the cross-cultural adaptation followed a forward-and-backward translation process in which individual items, response options, and questionnaire instructions were adapted to Swedish (Fig. 1). Two native Swedish speakers independently carried out translations from English to Swedish: one a physiotherapist and researcher and the other a researcher unfamiliar with physiotherapy. Next, the two translations were checked for inconsistencies, compared, adjusted, and pooled into one. Subsequently, two English-speaking translators: one with clinical expertise and one without clinical expertise, and none with knowledge of TOSQ, independently back-translated the Swedish translation of TOSQ to English. The two translations were then reviewed, compared to the original TOSQ to ascertain conceptual and semantic equivalence, combined into one final back-translated version, and used to construct the final TOSQ-S. Finally, TOSQ-S was pre-tested for accuracy of wording and comprehension in a cohort of 15 outpatients with LBP seeking physiotherapy.

### Statistical analysis

A polytomous Rasch rating scale model (Winsteps<sup>®</sup> Rasch measurement program v3.92.1, John M Linacre, Oregon,



**Fig. 1** Cross-cultural adaptation process of the Swedish version of the Treatment Outcome Satisfaction Questionnaire (TOSQ-S). *Swe* Swedish, *Eng* English

USA) with person measures scaled at 0–100 logits was used to evaluate the psychometric properties of TOSQ-S. The eight specific TOSQ-S items were first analyzed, and

subsequently, scale structure optimization was attempted starting with all ten TOSQ-S items. Analyses were initiated with an assessment of the scale’s categorical structure, and acceptability was set to: minimum ten observations per category, monotonical categorical advancement, and categorical fit to the model [19]. Next, person and item fit to the model was examined. Model fit was evaluated via the infit and outfit mean square (MNSQ) values, with 1.0 suggesting perfect fit and lower or higher values indicating overfit and underfit, respectively [18]. Significant MNSQ values smaller than 0.6 or larger than 1.4 were considered to misfit the model [20], and less than 5% of misfitted items and persons, respectively, was considered acceptable. Misfitted items were discarded one-by-one starting from the highest MNSQ value, until an acceptable internal scale validity was achieved. Misfitted persons were investigated to elucidate possible explanations for misfit, but were not omitted to maintain external validity. Scale unidimensionality was assessed via the proportion of total variance accounted for by the Rasch model and by a principal component analysis of the residuals (PCAR). Unidimensionality criteria were that observed variance met approximated expected model variance [21] and that eigenvalues of any residual components were less than 2.0 [22]; corresponding to that of less than two items in the latent residual dimension. The scale targeting, meaning the sample’s satisfaction level relative to items’ satisfaction requirement, was assessed via person–item maps [23]. Person separation reliability, meaning the scale’s reliability in distinguishing between persons according to their satisfaction level, was determined via the person separation index, with 1.50 considered acceptable and 2.00 good, as it can discern two and three satisfaction levels, respectively [23].

Test–retest reliability of individual items was estimated by the linearly weighted kappa and Krippendorff’s alpha for ordinal scale data [24, 25]. The overall scale reliability was calculated from the person logits measure, and quantified by both an intraclass correlation coefficient (ICC) based on a single-measurement two-way random effects model of absolute agreement, and the standard error of measurement (SEM, i.e., the within-subject standard deviation between trials) [26]. To ascertain that inter-trial differences were related to persons and not items, item measures were anchored with the estimates of the first trial in the retest Rasch model. Reliability estimates were interpreted as  $\leq 0.40$  = poor,  $0.40$ – $0.59$  = fair,  $0.60$ – $0.74$  = good, and  $\geq 0.75$  = excellent [27]. Reliability analyses were computed via the packages ‘rel’ (v1.1.0, Lo Martire 2016, <https://CRAN.R-project.org/package=rel>) and ‘irr’ (v0.84, Gamer et al. 2012, <https://CRAN.R-project.org/package=irr>) in R (v3.1.3, R Core Team 2016, R Foundation for Statistical Computing, Vienna, Austria). Statistical significance was set at  $p < 0.05$  for all analyses.

## Results

### Cross-cultural adaptation

Table 1 displays the final TOSQ-S. TOSQ was successfully forward-and-backward translated into Swedish with neither semantic nor language ambiguities, and all items were approved by both English and Swedish research members. Pre-testing of TOSQ-S revealed no difficulties.

### Rasch analysis

Figure 2 shows the response distribution of the 131 patients across the TOSQ-S items. Whereas the responses were skewed towards high treatment outcome satisfaction, no patients had maximal or minimal scores. Nonetheless, the eight-item TOSQ-S Rasch model's lowest response category only had one observation, rendering that category's estimates unstable.

Moreover, while the average categorical measures advanced monotonically, the second and third lowest categories underfitted the model ( $MNSQ_{\text{outfit}} \leq 1.81$ ), and

collapsing them did not resolve the misfit ( $MNSQ_{\text{outfit}} \leq 1.70$ ). In addition, whereas both the variance accounted for by the model (45%; eigenvalue = 6.6) and the PCAR (eigenvalues <1.8) supported unidimensionality, nine persons (7%) and one item (item 8:  $MNSQ_{\text{infit}} = 1.51$ ;  $MNSQ_{\text{outfit}} = 1.55$ ) underfitted, and three persons (2%) overfitted the model. Finally, TOSQ-S mistargeted the sample, having a mean difference of 22 logits between person satisfaction and item satisfaction requirement; however, the person separation index of 1.86 indicated that it could nevertheless distinguish between patients of high and low satisfactions.

To improve TOSQ-S's psychometric properties, data were reanalyzed following stepwise modifications. First, all ten items were analyzed together. Whereas all categories had at least ten observations and their average measures advanced monotonically, the two lowest categories misfitted the Rasch model ( $MNSQ_{\text{infit}} \leq 1.66$ ;  $MNSQ_{\text{outfit}} \leq 2.68$ ), and collapsing them did not resolve the misfit ( $MNSQ_{\text{infit}} = 1.80$ ;  $MNSQ_{\text{outfit}} = 2.83$ ). Again, both the variance accounted for by the model (53%; eigenvalue = 11.2) and the PCAR (eigenvalues <1.9) supported

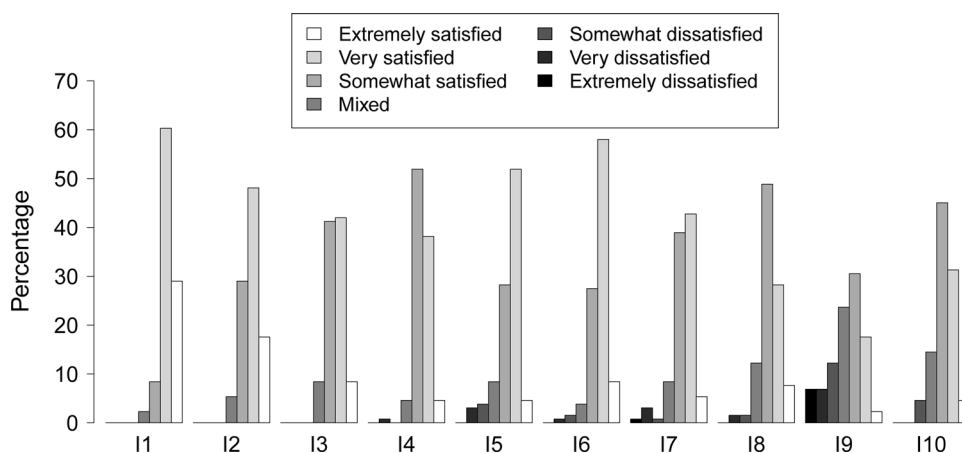
**Table 1** Treatment Outcome Satisfaction Questionnaire

Item	Content
1*	Sammantaget, hur nöjd är du med resultatet av sjukgymnastbehandlingen? [All things considered how satisfied are you with the results of the Physical Therapy treatment?]
2	Hur nöjd är du med hur din rygg har svarat på behandlingen? [How satisfied are you with your back's response to the treatment?]
3	Hur nöjd är du med hur du hanterar din ländryggsmärta? [How satisfied are you with how you cope with your low-back pain?]
4	Hur nöjd är du med din egen insats vid behandlingen? [How satisfied are you with the effort you put into your physical treatment?]
5	Hur nöjd är du med din förmåga att ta del i aktiviteter som är viktiga för dig? (Detta kan inkludera arbete, fritidsaktiviteter, social liv mm.) [How satisfied are you with your ability to take part in activities, which are important to you? (This could include your job, hobbies, social life, etc.)]
6	Hur nöjd är du med hur väl du förstår din ryggsmärta? [How satisfied are you with how well you understand your back pain?]
7	Hur nöjd är du med din förmåga att leva som vanligt? (Även om den är annorlunda än innan dina ryggbesvär startade) [How satisfied are you with your ability to carry on with your life? (Even if it is different than before your back pain started)]
8	Hur nöjd är du med det du gör för att ta hand om dig själv och din rygg? [How satisfied are you with what you are doing to take care of yourself and your back?]
9*	Om du skulle vara tvungen att leva resten av ditt liv med de symptom du har just nu, hur skulle du känna då? [If you had to spend the rest of your life with the symptoms you have right now, how would you feel about it?]
10	Hur nöjd är du med din förmåga att ta kontroll över din ländryggsmärta? [How satisfied are you with your ability to take control over your low-back pain?]

Square brackets denote original English items adapted from Lis [15]

\* Global items not included in scoring. Response categories: extremt nöjd [extremely satisfied], mycket nöjd [very satisfied], ganska nöjd [somewhat satisfied], blandat (ungefär lika nöjd som missnöjd) [mixed (about equally satisfied and dissatisfied)], ganska missnöjd [somewhat dissatisfied], mycket missnöjd [very dissatisfied], extremt missnöjd [extremely dissatisfied]

**Fig. 2** Raw score item frequencies for the Swedish version of the Treatment Outcome Satisfaction Questionnaire ( $n = 131$ )



unidimensionality; however, 14 patients (11%) and two items (Table 2) underfitted, and nine subjects (7%) overfitted the model. The person–item map revealed an increase in the mistargeting with a mean difference of 28.4 logits between person satisfaction and item satisfaction requirement (Fig. 3a), but the person separation index remained at a similar magnitude of 1.89.

Following exclusion of the misfitted items 8 and 9, nine (7%) and four subjects (3%) underfitted and overfitted the model, respectively. The additional exclusion of item 6 reduced the number of underfitted subjects to seven (5%) but increased the number of overfitted subjects to six (5%). Finally, discarding item 10 reduced the number of overfitted subjects to three (2%). The exclusion of these four items resulted in the three lowest response categories having less than ten observations, rendering estimates there unstable; however, their average measures advanced monotonically, and while the third lowest category

overfitted the model ( $MNSQ_{infit} = 0.55$ ;  $MNSQ_{outfit} = 0.37$ ), collapsing the second and third lowest categories this time resolved the misfit. The reduced scale accounted for 60% of the total variance (eigenvalue = 8.9), and no substantive residual dimensions were identified during PCAR (eigenvalues <1.7); both measures again suggesting unidimensionality. The mistargeting remained at a similar magnitude of 26.3 logits (Fig. 3b); however, the person separation index at 2.04 suggested that three distinct subgroups could be reliably discerned by the six-item sub-configuration (Fig. 3b).

### Test–retest reliability

Reliability point estimates suggested excellent, fair, and poor agreements for two, six, and two items, respectively (Table 2). For the six-item TOSQ-S summary score, mean (SD) person measures were 38.3 (20.8) logits the first trial

**Table 2** Psychometric properties of the Swedish version of the Treatment Outcome Satisfaction Questionnaire

Item	Raw score (SD)	10-item Rasch model ( $n = 131$ )			Test–retest reliability ( $n = 41$ )	
		Logits measure (SE)	Infit MNSQ ( $z$ )	Outfit MNSQ ( $z$ )	LW kappa (95% CI)	K alpha (95% CI*)
1	5.2 (0.7)	−16.6 (2.8)	0.82 (−1.5)	0.81 (−1.6)	0.49 (0.28, 0.71)	0.54 (0.25, 0.76)
2	4.8 (0.8)	3.3 (2.5)	0.97 (−0.2)	0.99 (0.0)	0.77 (0.63, 0.91)	0.87 (0.75, 0.95)
3	4.5 (0.8)	15.3 (2.3)	0.68 (−2.4)	0.77 (−1.9)	0.32 (0.13, 0.52)	0.44 (0.14, 0.66)
4	4.4 (0.7)	19.1 (2.3)	0.82 (−1.3)	0.77 (−1.8)	0.76 (0.58, 0.95)	0.77 (0.55, 0.93)
5	4.4 (1.1)	20.8 (2.2)	0.97 (−0.2)	0.87 (−0.9)	0.49 (0.23, 0.75)	0.44 (0.12, 0.69)
6	4.7 (0.8)	8.9 (2.5)	1.19 (1.3)	1.14 (1.0)	0.25 (0.01, 0.50)	0.32 (0.00, 0.60)
7	4.3 (1.0)	22.4 (2.2)	0.84 (−1.1)	0.76 (−1.9)	0.50 (0.33, 0.66)	0.59 (0.31, 0.79)
8	4.2 (0.9)	25.0 (2.1)	<b>1.50 (3.0)</b>	<b>1.43 (2.8)</b>	0.47 (0.24, 0.69)	0.53 (0.22, 0.76)
9	3.3 (1.5)	49.9 (1.7)	1.35 (2.4)	<b>1.79 (4.8)</b>	0.56 (0.37, 0.75)	0.65 (0.33, 0.86)
10	4.2 (0.9)	27.2 (2.1)	0.81 (−1.3)	1.03 (0.3)	0.43 (0.18, 0.68)	0.47 (0.14, 0.70)

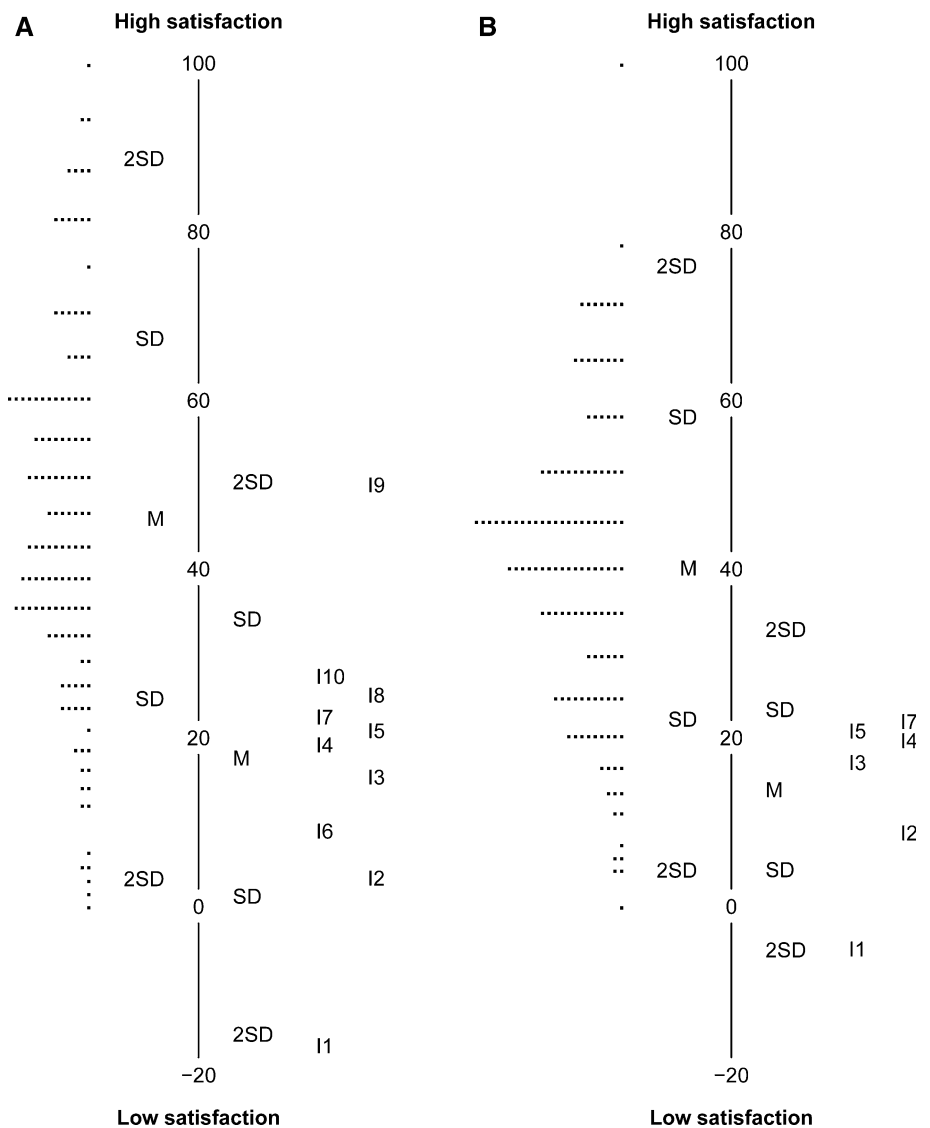
Mean values presented for raw score, logits measure, LW kappa (linearly weighted kappa), and K alpha (Krippendorff's alpha for ordinal scale data)

Bold–italic digits denote misfitted items

SD standard deviation, SE Standard error, MNSQ mean square,  $z$  score, CI confidence interval

\* Bootstrapped with 2000 replicates

**Fig. 3** Item-person map of the Swedish version of the Treatment Outcome Questionnaire (TOSQ-S), with patients (*squares*) satisfaction relative to the satisfaction requirement of individual items (I). A is based on a Rasch model of all ten TOSQ-S items and B on the final six-item TOSQ-S configuration. *M* mean, *SD* standard deviation



and 33.4 (16.4) logits the second trial, with mean (95% CI) ICC and SEM of 0.66 (0.44, 0.80) and 6.6 (3.8, 9.4) logits, respectively. Hence, both ICC and SEM point estimates suggested acceptable reliability, albeit a decrease in the average satisfaction level was observed.

One patient decreased considerably more in satisfaction between trials than other patients. Excluding this patient from the analyses reduced the inter-trial mean discrepancy to 3.3 logits, decreased SEM to 5.5 (3.7, 7.4) logits, and increased ICC to 0.79 (0.64, 0.88).

**Discussion**

This study cross culturally adapted TOSQ from English to Swedish and assessed the psychometric properties of the translated scale in patients with LBP following

physiotherapy completion. Our results showed that a six-item TOSQ-S configuration had acceptable psychometric properties and is suitable for measuring treatment outcome satisfaction of physiotherapy in patients with LBP.

The cross-cultural adaption process resulted in an equivalent version of TOSQ in Swedish. The Rasch model, however, revealed that TOSQ-S had a considerable amount of misfitted items and respondents, but these limitations were largely remedied in the scale optimization procedure. A six-item subconfiguration was able to reliably distinguish between patients of high-, average-, and low-treatment outcome satisfaction, met the criteria for an acceptable proportion of misfitted items, and had slightly higher than 5% misfitted persons. The average person measure sample’s satisfaction level was high relative to the items’ satisfaction requirement, suggesting that TOSQ-S may be mistargeted to the intended population; however, no ceiling



effect was observed (Fig. 3). The skewness towards high sample satisfaction could be a consequence of treatments being conducted at specialist clinics, and the person–item relationship could, therefore, change considerably at general outpatient clinics. Consequently, although our sample size met the recommendation for polytomous Rasch models [28], the three lowest categories had very few observations, which can explain the misfit of one of these categories to the model. The categorical structure is, therefore, a possible area of improvement, but was kept unchanged until further investigated. Unscaled item measure standard errors below 0.2 logits, however, supported that item measures had a precision of 0.5 logits with 95% confidence [28].

On average, the six-item TOSQ-S summary score had acceptable reliability, with ICC suggesting good reliability and SEM corresponding to 7% of the person measure range. Whereas equivalent patient numbers increased and decreased in satisfaction between trials, one patient displayed a systematic decrease across most items, suggesting true decrease in satisfaction for this instance. The considerably higher ICC estimates following exclusion of this patient was likely a result of ICC's dependence on sample variance, as SEM was only marginally affected [26]. Hence, ICC estimates excluding the patient may be more accurate.

Interestingly, item 1 which assessed overall treatment satisfaction had the lowest satisfaction requirement, while item 9 which inquired about how satisfied patients would be living the rest of their lives with the current low-back symptoms had the highest satisfaction requirement; each separated by more than one standard deviation from adjacent items (Fig. 3a). This suggests that global items inquiring about current satisfaction overestimate satisfaction level and thereby may not provide representative satisfaction estimates. In contrast, items inquiring about the long-term perspective appear to be conservative in their estimates. Thus, while item 9 was discarded in the final Rasch model, it, nevertheless, may be valuable as a proxy for treatment success from the patient's perspective; either reflecting complete recovery from symptoms or incomplete recovery paired with symptom acceptance.

Patient satisfaction has been recommended as an outcome measure for LBP interventions [4, 5], and as no such instruments currently exist in Swedish, TOSQ-S can suitably fill the void; both as a measure of health care quality and of the patient's perception of treatment outcome. However, while the initial design of TOSQ intended for a summary score of the specific items [15], the Rasch analysis suggested that another configuration is more suitable for TOSQ-S. We, therefore, recommend that the TOSQ-S summary score be derived from the six-item configuration, and the other items remain for their descriptive value. The permissive inclusion criteria

combined with the indiscriminate consecutive patient recruitment via remittance from health care clinics in areas of different socioeconomic status facilitate the results extrapolation. Nonetheless, it is limited to the Swedish speaking urban population.

## Conclusions

A six-item configuration of the Swedish version of the Treatment Outcome Satisfaction Questionnaire showed acceptable psychometric properties and is suitable for measuring treatment outcome satisfaction of physiotherapy in patients with LBP.

**Acknowledgements** We thank Richard Danell for collecting data in the test–retest part of the study. We also thank the other participating physiotherapists and the patients for their contribution.

## Compliance with ethical standards

Primary data available for journal review if requested.

**Conflict of interest** None declared.

**Ethical approval** This study was approved by Stockholm's Regional Committee for Medical Research Ethics and conformed to the declaration of Helsinki.

**Informed consent** All participants received study information and signed an informed consent.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Vos T et al (2012) Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 380:2163–2196. doi:10.1016/S0140-6736(12)61729-2
- Lin CC et al (2016) The economic burden of guideline-recommended first line care for acute low back pain. *Eur Spine J*. doi:10.1007/s00586-016-4781-0
- Waddell G, Burton AK (2005) Concepts of rehabilitation for the management of low back pain. *Best Pract Res Clin Rheumatol* 19:655–670. doi:10.1016/j.berh.2005.03.008
- Deyo RA et al (1998) Outcome measures for low back pain research. A proposal for standardized use. *Spine (Phila Pa 1976)* 23:2003–2013
- Bombardier C (2000) Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. *Spine (Phila Pa 1976)* 25:3100–3103
- Beattie PF et al (2002) Patient satisfaction with outpatient physical therapy: instrument validation. *Phys Ther* 82:557–565

7. Hudak PL et al (2003) The metaphor of patients as customers: implications for measuring satisfaction. *J Clin Epidemiol* 56:103–108
8. Linder-Pelz S (1982) Social psychological determinants of patient satisfaction: a test of five hypothesis. *Soc Sci Med* 16:583–589
9. Linder-Pelz SU (1982) Toward a theory of patient satisfaction. *Soc Sci Med* 16:577–582
10. Potter M et al (2003) The physiotherapy experience in private practice: the patients' perspective. *Aust J Physiother* 49:195–202
11. Hudak PL et al (2004) Testing a new theory of patient satisfaction with treatment outcome. *Med Care* 42:726–739
12. Werner EL et al (2010) The COPE LBP trial: cognitive patient education for low back pain—a cluster randomized controlled trial in primary care. *BMC Musculoskelet Disord* 11:33. doi:[10.1186/1471-2474-11-33](https://doi.org/10.1186/1471-2474-11-33)
13. Nielsen PR et al (2010) Prehabilitation and early rehabilitation after spinal surgery: randomized clinical trial. *Clin Rehabil* 24:137–148. doi:[10.1177/0269215509347432](https://doi.org/10.1177/0269215509347432)
14. Casserley-Feeney SN et al (2008) Patient satisfaction with private physiotherapy for musculoskeletal pain. *BMC Musculoskelet Disord* 9:50. doi:[10.1186/1471-2474-9-50](https://doi.org/10.1186/1471-2474-9-50)
15. Lis A (2010) Reliability and validity of a treatment outcome satisfaction questionnaire for patients with low back pain (LBP). Dissertation, New York University
16. Beaton DE et al (2000) Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila Pa 1976)* 25:3186–3191
17. Nunnally JC, Bernstein IH (1994) *Psychometric theory*. McGraw-Hill, New York
18. Tesio L (2003) Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *J Rehabil Med* 35:105–115
19. Linacre JM (2002) Optimizing rating scale category effectiveness. *J Appl Meas* 3:85–106
20. Wright BD, Linacre JM (1994) Reasonable mean-square fit values. *Rasch Measurement Trans* 8(3):370
21. Linacre JM (2008) Variance in data explained by Rasch measures. *Rasch Measurement Trans* 22(1):1164
22. Basra MK et al (2015) PFI-14(c): a Rasch analysis refinement of the psoriasis family index. *Dermatology* 231:15–23. doi:[10.1159/000379748](https://doi.org/10.1159/000379748)
23. Duncan PW et al (2003) Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. *Arch Phys Med Rehabil* 84:950–963
24. Krippendorff K (2004) *Content analysis: an introduction to its methodology*. Sage, Thousand Oaks
25. Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 70:213–220
26. Hopkins WG (2000) Measures of reliability in sports medicine and science. *Sports Med* 30:1–15
27. Cicchetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 6:284–290
28. Linacre JM (1994) Sample size and item calibration stability. *Rasch Measurement Trans* 7(4):328