Title:

Validity and reproducibility of the STarT Back Tool (Dutch version) in patients with low back pain in primary care settings

Authors:

Jasper D. Bier[1, 2]
Raymond W. J. G. Ostelo[3, 4]
Miranda L. van Hooff[5]
Bart W. Koes[1]
Arianne P. Verhagen[1]

1. Department of General Practice, Erasmus MC, PO Box 2040, 3000 CA, Rotterdam, The Netherlands
2. Fysiotherapie Fascinatio, Capelle aan den IJssel, The Netherlands
3. Department of Epidemiology and Biostatistics, VU University Medical Centre Amsterdam and the EMGO Institute for Health and Care Research, The Netherlands
4. Department of Health Sciences, Faculty of Earth and Life Sciences, VU University Amsterdam, The Netherlands.
5. Sint Maartenskliniek, Nijmegen, The Netherlands

Correspondence to:
Jasper D. Bier, Department of General Practice, Erasmus MC, PO Box 2040, 3000CA Rotterdam, The Netherlands
T +31 6 26536787 **E** j.bier@erasmusmc.nl **URL** www.erasmusmc.nl

Total number of pages:
Manuscript = 21 pages (including title page, abstract, figure / table legends and references)
Figures = 1
Tables = 5

ABSTRACT

Objective: The purpose of this study was to translate and to investigate the reliability and validity of the STarT Back screening tool (SBT) in the primary care setting among patients with non-specific low back pain (LBP).

Design: The SBT was formally translated into the Dutch language following a multistep approach for forward and backward translation. General practitioners and physiotherapists included patients with LBP.

Methods: Patients completed a baseline a questionnaire and a follow-up at 3 days and 3 months. The construct validity was calculated with the Pearson's correlation coefficient. The reproducibility was assessed using the quadratic weighted kappa and the specific agreement. Predictive validity was assed using relative risk ratios for persisting disability at 3 months. Content validity was analyzed using floor- and ceiling effects.

Results: In total, 184 patients were included; 52.2% of patients were categorized in the "low-risk" subgroup, 38.0% "medium-risk" and 9.8% as high risk. For the construct validity we found, as expected, a moderate to high Pearson's correlation for question 3 to 9 and a low correlation for question 1 and 2 with their respective reference questionnaires. The reproducibility had a quadratic weighted kappa of 0.65 and the specific agreement of 82.4% for "low-risk", 53.3% for "medium-risk" and 33.3% for "high-risk". For the predictive validity for persisting disability we found a relative risk ratio for "medium-risk" of 1.8 (95% confidence interval [CI] 1.0 – 3.1) and 2.7 (95% C.I. 1.4 – 4.9) for "high-risk" compared with "low-risk". For the content validity, we found that no floor and ceiling effects were present.

Limitations: There was a relatively small sample size for the retest reliability study, patients were not compared between physical therapists and GPs, as there were not enough patients in both groups. For practical reasons, the patients filled out the baseline questionnaire after receiving the first treatment/consultation; however, the questionnaire is intended to be filled in before the first consultation/treatment.

Conclusion: The SBT is successfully translated into Dutch. The psychometric analysis showed acceptable results and, therefore, the SBT as a valid screening tool for patients with LBP in Dutch primary care.

Keywords:
GP, physical therapy, low back pain, classification, validation

Word count article: 3569

BACKGROUND

Low back pain (LBP) is a major public health problem. Globally it is the most prevalent musculoskeletal disorder causing disability.[1] In the Netherlands the point prevalence of LBP is found to be 26.9%.[2] LBP is a condition that is broadly divided into three major subgroups. First, LBP with a specific (serious) underlying pathology, such as: tumors, fractures and infections. Second, LBP caused by nerve root compression as a result of a stenosis or herniated disc. The third group, the majority of people with LBP (85 - 90%), is called non-specific LBP as no cause can be found.[3,4] Despite the fact that non-specific LBP is regarded self-limiting as it often resolves within six weeks, more recent prognostic studies concluded that around 40% of patients with LBP recovery will take longer than 12 weeks.[4–6] LBP is a burden on the health care system consuming in the Netherlands between 385 and 455 million euro's in direct medical costs and between 3 and 3.1 billion euro's in indirect costs .[7] In the United States reports on combined direct and indirect costs for LBP vary between $86 billion and $238 billion dollar.[8–10]

Although it has been suggested that patients with non-specific LBP are not a homogeneous patient group, defining subgroups is challenging but important for targeting treatment to the individual patient.[3,11,12] So far sub-grouping based on a patho-anatomical source of the pain appears to be of limited value because often an anatomical structure as cause of pain cannot be found.[4] Certain psychosocial factors are known to influence patients' recovery. Subgrouping patients based on psychosocial factors may result in successful risk stratification. The STarT (Subgroup Targeted Treatment) Back Screening Tool (SBT) is a tool using different function, psychosocial and comorbid factors for subgrouping. It is developed in England, to allocate primary care patients with LBP into three subgroups concerning their prognosis: low, moderate or high risk for persisting disability[13] and to apply the appropriate stratified care.[14]

The SBT consists of 9 questions, 8 true/false questions and 1 question with a 5-point Likert-scale as answer option. The validity of the SBT is often studied using a principal component factor analysis. In the United Kingdom (UK) study, as well as the Finnish, French, German and Persian studies it resulted in two subscales: biological (question 1 to 4) and psychosocial (question 5 to 9).[13,15–18] The psychosocial subscale is then viewed as a distress subscale with a Cronbach's alpha ranging from 0.52 (Finnish), 0.55 (German), 0.72 (UK), 0.74 (French), and 0.81 (Persian).[13,15–18] The discriminant validity has been determined by calculating the area under the curve (AUC) of the overall score with the Roland Disability Questionnaire (RDQ) (0.76-0.92).[13,16] The psychosocial subscale of the Pain Catastrophizing Scale (PSC) (0.70-0.83) or the Tampa Scale of Kinesiophobia (TSK) (0.81).[13,18] Other studies calculated the AUC for each separate question resulting in AUCs ranging from 0.74 – 0.86.[16,19] The SBT is a questionnaire formed by combining known factors for delayed recovery of back pain. Based on these independent factors it aims to predict poor disability with each factor adding to the likelihood of a poor prognosis, this is called a formative model. In our formative model approach it is unnecessary to calculate internal consistency and the AUC against the overall score or the psychosocial subscale as we approach it as independent factors and not as a coherent factors.

The SBT's ability to predict poor disability at 6 months had sensitivity scores ranging between 39.6% and 80.1% and the specificity scores ranging from 65.4% to 94.6%.[13] The English SBT has been found to be a reliable tool in the UK with a quadratic weighted kappa of 0.79.[13] It has been translated in several languages since its initial English publication in 2008.[16–22] No study has been published on the SBT to evaluate the validity and reliability in Dutch primary healthcare. Our aim is to evaluate the validity and reliability of the STarT Back Tool Dutch Version in Dutch primary care.

METHOD

Translation of the SBT

The original SBT[13] (Appendix A and B) was formally translated following the multistep approach of Beaton et al.[23] and the guidelines of Streiner and Norman.[24] Two Dutch native speakers independently performed a forward translation. After synthesis of a draft Dutch translation of the SBT, this version was backward translated into English by both a Dutch and an English native speaker. An expert committee was formed consisting of one translator who is also a clinical epidemiologist, one backward translator and one clinician (orthopedic spine surgeon). The group examined the forward and backward translations and consolidated these to produce a 'pre-final' version of the Dutch SBT. As it became apparent that two different study groups were preparing Dutch translations a second expert meeting, consisting of a representative of each study group (RO and MvH) was held. A 'combined pre-final' version was compiled based on all previous documents and differences were resolved through consensus. The only difference was found in the translation of question 1 'spread down my leg(s)'. We discussed whether to use 'naar één of beide benen' (i.e. 'to one or both legs') ' or 'naar mijn benen' (i.e. 'to both legs'). As in the original English version the 's' of 'legs' is written between brackets, consensus was reached to use 'naar één of beide benen' (i.e. 'to one or both legs') in the 'combined pre-final' version.

Pre-final testing

To test the 'combined pre-final' version, twenty consecutive Dutch-speaking patients with LBP at the outpatient department of a secondary and tertiary spine referral center, completed this version. In addition, a possibility was made to give comments and suggestions to improve. After completion, they were briefly interviewed about their thoughts of what was meant by each question and the chosen answer. They were also asked for their general comments on the questionnaire (e.g. lay-out, wording, ease of understanding and completion, ambiguities). As no further comments or suggestions to improve were given, the expert group upgraded the 'pre-final' version to the final version. [25] The Dutch version of the SBT is found in Appendix A and B

Design

The final translated version was subsequently used in this clinimetric study as part of a prospective cohort (PRINS study; Prevalence of RIsk groups in Neck- and back pain patients according to the STarT back screening tool) including patients with LBP (and neck pain for a parallel study) of any duration in primary care. This is the first article published on this cohort. Patients received regular care by their

general practitioner (GP) or physiotherapist. In the Netherlands patients have direct access to physical therapist care and therefore this is regarded primary care as is GP-care. Patients were asked to answer baseline and follow-up questionnaires. A power analysis power showed that 100 patients were needed for a reliability study. The study was approved by the medical ethics committee of the Erasmus University, Rotterdam, The Netherlands. (MEC-2014-256). For this study we only use the data of the LBP patients of the PRINS-cohort.

Participants

General Practitioners and Physical Therapists We asked GP's and physical therapists that had previously showed their interest in the SBT to participate in the study and asked them to invite colleagues. Information about the study protocol was given through several meetings, by phone, or by digital/paper documentation. Participating GP's and therapists received the study protocol and a folder with patient information brochures and informed consent forms.

Patients The inclusion period for patients started November 2014 to May 2015. When a patient consulted their physical therapist or GP for their back pain they were asked to participate in the PRINS study. Other inclusion criteria were that the patient was 18 years or older, could speak, read and write in Dutch and had an email address. Patients were excluded if during the consultation the GP or therapist found red flags indicating a possible specific underlying pathology (e.g. infection, fracture, cauda equina or tumor) responsible for the LBP.
Patients were given oral and written information about the procedure of data collection and the aim of the study. They were given an informed consent form. When the patient signed the informed consent form and handed it back to their therapist or GP they registered the patient online. The patient immediately received an email with a link to the baseline questionnaire. When necessary a reminder was sent within a few days.

Treatment The clinician was blinded for the results of the questionnaires including the score on the SBT. The patients received usual care by their GP or physical therapist. We asked the clinician to treat their patient according to their guideline. The guideline advises the GP provide advice and, if necessary, analgesics to patients in the acute phase. In case of persisting pain GPs can refer the patient to the physical therapist. Guideline recommendations for PTs differ based on the course of pain. In a normal course of pain the PT is advised to give reassurance and information to the patient. In case of an abnormal course of pain the therapist should provide evidence based interventions such as exercise therapy, mobilization, manipulation and/or massage.[26,27]

Measurements

Baseline At baseline (T0) patients filled out a questionnaire consisting of demographic variables (such as age, gender) and the SBT. Furthermore, we measured the average pain in the past week using the 11-point Numeric Pain Rating Scale (NPRS)[28] ranging from 0 (no pain) to 10 (worst imaginable pain). Disability was operationalized using the RDQ[29,30] consisting of 24 statements with a "yes" or "no" answer option. The

total score ranges from 0 to 24, a higher score indicating more disability. We measured Fear of movement/(re)injury using the TSK[31] consisting of 17 statements with four answer options varying from "highly disagree" to "highly agree". The total score ranges from 17 to 68, a higher score indicating a higher level of kinesiophobia. To assess the level of catastrophizing we used the PCS, which consists of 13 statements with each a 5-point Likert scale answers option ranging from "not at all" to "always".[32] The total score ranges from 0 to 52, a higher score indicating a higher level of catastrophizing. Finally we assessed quality of life using the EQ-5D[33] consisting of six questions. The first 5 questions have a 3-point Likert scale answer options ranging van "no problems" to "severe problems" and the sixth question is a health status question ranging from "worst imaginable health" to "best imaginable health", score ranges from 0 to 100.

Follow-up Three days after inclusion (T1) a repeat-questionnaire was sent in order to investigate the re-test reliability of the SBT. It consisted of the SBT, the NPRS and the general perceived effect (GPE) scale to measure recovery: "To what degree have you improved since filling out the baseline questionnaire?" The answer options range from "fully recovered" to "worse than ever" on a 7-point Likert scale. The time interval was considered long enough to reduce recall bias and short enough to prevent substantial improvement.[34] This repeat questionnaire was send to patients that were included during the last 3 months of the inclusion period.
Three months after inclusion (T2), the patients received a follow-up questionnaire consisting of the GPE and RDQ. At the same time, we sent a questionnaire to the GP to ask about the number of visits, prescribed medication, referrals to physiotherapists or medical professionals and requested diagnostic imaging and blood tests. We sent a similar questionnaire to the PT to ask about treatment data such as date of first and last treatment, number of treatments, questionnaires used and the aim and means of treatment. All questionnaires were handled and stored though lime survey 2.05 (Lime-Survey GmbH, Hamburg, Germany).

Statistical analysis
First we analyzed the data to describe the characteristics of the GP's, physical therapists, and the patient population using frequencies, means and standard deviations. The prevalence of the 3 risk profiles according to the SBT-scores are reported.

For the **construct validity** we first analyzed at the characteristics across SBT risk profile to determine the discriminant validity. Next, we calculated the Pearson's correlation coefficient between specific items of the SBT and their respective reference questionnaires based on the comparability of the domains of measurement.[35,36] A priori we expected a moderate ($r \geq 0.3$, <0.5) to high ($r \geq 0.5$) correlation between the SBT activity-questions 3 and 4 with the RDQ, kinesiophobia-question 5 with the TSK, catastrophizing-question 6, 7 and 8 with the PCS and the bothersome-question 9 with the NPRS. We expected a low correlation ($r < 0.3$) between question 1 and 2 and the NPRS as these focus on the location of the pain and not the intensity of pain.

We calculated the **reproducibility** (evaluating the agreement between two measurements) in the patient group that remained stable between baseline (T0) and T1. We asked the patients after three days to fill out the questionnaire a second time. Patients were considered stable when they scored "slightly improved", "no change" or "slightly worsened" on the GPE at second measurement. As there is some doubt in the literature whether the GPE actually can detect change, we combined the stable GPE score with a stable pain score meaning the NPRS on T1 was plus or minus one point compared to baseline.[3] We calculated the quadratic weighted kappa and the specific agreement. The quadratic weighted kappa will be interpreted as ≤0 = poor agreement; .01–.20 = slight; .21–.40 = fair; .41–.60 = moderate; .61–.80 = substantial and .81–1 = almost perfect agreement.[37] The specific agreement is calculated for each risk profile separately.[35] For example; patients who are "low-risk" on baseline and follow-up are calculated as a proportion of patients that were "low-risk" on either of the two measurements. In collaboration with Henrika de Vet we modified the specific agreement to fit a 3x3 table as shown in table 1 because the original method is done in a 2x2 table.

We determined the **predictive validity** by reporting the relative Risk Ratio (RR) for "medium-risk" and "high-risk", both compared to "low-risk" in their ability to predict the outcome on three months. We defined persisting disability as a RDQ of ≥ 7, this is equal to the cut-off used in the original study where it was the median of the baseline scores.[13] Persisting pain is defined as a NPRS above the baseline median and recovery is defined as either "completely recovered" or "much improved" on the GPE.

Limited **content validity** is indicated by the presence of more than 15 percent of the patients reached either the floor (0/9 points) or ceiling effects (9/9 points) on the SBT.[34] .

To measure the construct validity and reliability a sample size of at least 50 persons is advised. [34]

RESULTS

Patient population

In total, 41 GPs and 70 physical therapists signed up to participate and 12 GPs and 33 physical therapists actually included patients. They included 370 patients of which 184 with LBP and 100 with neck pain for the parallel study, 86 patients did not fill out the baseline questionnaire and were excluded from the analysis. Loss to follow up at three months was 34 (18%). (figure 1) Patients that were lost to follow up showed comparable baseline characteristics compared to the responders. Of the LBP patients, at baseline 96 (52%) patients were categorized as "low-risk", 70 (38%) as "medium-risk" and 18 (10%) as "high-risk" (table 2). We found no differences between the groups concerning age, gender of whether they were included by the GP or physical therapist.

Validity and reproducibility

Construct validity. For each increase in the risk profile we found a corresponding increase in pain, disability, catastophising and kinesiofobia (Table 3) showing that the SBT has good discriminant validity. Next we found a high correlation, between SBT question 9 and the NPRS (r = 0.6), question 3 and 4 with

the RDQ and question 8 with the PCS (all $r = 0.5$). We found a moderate correlation ($r = 0.4$) between question 5 and the TSK and question 6 and 7 and the PCS (both $r = 0.3$). The correlation between question 1 and 2 was absent to low and scored $r = 0.28$ and $r = -0.05$ respectively (table 4). The correlations are as was expected a priori and therefore we conclude that the construct validity is good.

Reproducibility The average time between first (T0) and second (T1) questionnaire was 6 days (range 3-10). In total, 58 patients completed the second questionnaire of which 19 patients were regarded stable compared to baseline. The quadratic weighted kappa for the SBT of 0.65 (95% C.I. 0.34 – 0.96) showed a substantial reproducibility. The "low-risk" group had a specific agreement of 82.4%, "medium-risk" of 53.3% and "high-risk" of 33.3% showing an excellent to fair reproducibility.

Predictive validity. In total, 150 patients completed the T2 questionnaire, of which 76 were regarded as "low-risk" at baseline, 58 as "medium-risk" and 16 as "high-risk". In all three risk profiles a decrease in NPRS and RDQ scores over time was seen. The number of patients with a decrease that met the threshold (RDQ < 7) was highest in the "medium-risk" group (table 5). Persisting pain is set as a NPRS ≥ 6. The RR for "medium-risk" at 3 months as compared to the "low-risk" were 1.8 (95% C.I 1.0 – 3.1) for persisting disability, 1.6 (95% C.I. 0.9 – 3.0) for persisting pain and 1.0 (95% C.I. 0.7 – 1.3) for recovery. For "high-risk" compared to the low-risk group the RR were 2.7 (95% C.I. 1.4 - 4.9) for persisting disability, 3.4 (95% C.I. 2.3 – 6.8) for persisting pain and 0.6 (95% C.I. 0.3 – 1.2) for perceived recovery. An RR of 3.4 means that patients with "high-risk" had 3.4 times higher chance for persisting low back pain compared to patients with "low-risk". Some confidence intervals include 1 (= equal risks) making it statistically insignificant.

Content validity We analyzed 184 baseline questionnaires concerning the SBT in determining floor and ceiling effects. Nine patients (4.9%) scored zero and 2 patients (1.1%) scored 9 points implying no floor or ceiling effects are present and therefore the SBT showed a good content validity.

DISCUSSION

Main findings

The SBT is a formative model aiming to give a prognosis on poor disability. The construct validity showed correlations as a priori was expected between SBT items with their respective reference questionnaires (NPRS, RDQ, TSK and PSC). The re-test reliability is moderate to good, and the RR demonstrates an increased chance for persisting disability and pain with an increase of the risk profile. An expert committee found the questions to be relevant and 20 patients used the SBT and comprehended all questions. Furthermore the absences of floor and ceiling effects confirmed a good content validity.

Interpretation of findings

The specific agreement, as a measurement to determine the reproducibility, shows a fairly accurate intra-observer consistency for patients with a "low-risk" score. The accuracy decreases as the risk-profile

increases. This might be due to the relatively low number of patients in this high-risk category. Also, in "high-risk" patients multiple psychosocial factors are present, which can be influenced during therapy by addressing an active health behavior and the unlikeliness of a serious underlying condition.[38] The latter is probably less of influence as the questionnaire at baseline is given after the first treatment during which the psychosocial factors and the active health behavior are likely to have been addressed. Patients might have been influenced by this information during the primary consultation and therefor shifted from the "high-risk" to the "medium-risk" group before completing the baseline questionnaire. A previous study suggests that assignment to a risk category following a short delay may more successfully predict final outcomes than when administered during initial assessment.[39]

For the reproducibility analysis the conditions (time, pain, perceived recovery) were set a priori to ensure 'stable patients'. Due to the natural course of the pain, patients might be recovering between both measurements, shifting to a lower risk-profile and explaining the higher score in the "low-risk" group. The Kappa is influenced by a skewed distribution due to the large proportion of patients with "low-risk". Nevertheless the Kappa shows that the SBT is able to distinguish sufficiently between risk groups.[35]

Within the reproducibility analysis we found that 4 out of the 5 patients shifted from "high-risk" to "medium-risk" within the first week. These patients had only one consultation in this period and therefore might have been susceptible to change.

We used relative risks (RR) to calculate the additional risk of "high-risk" and "medium-risk" compared to "low-risk". Predicting persisting disability gave the best results, in accordance with the developers aim. Poor disability is defined as a RDQ score of 7 or more, like the original study and other comparable studies.[13,15–18,38] In interpreting the predictive value it has to be taken into account that clinicians applied 'usual care'. There was no standardized or stratified therapy protocol for the clinicians to use. We asked the GP or physical therapist the follow the national guidelines, but recent studies show that guidelines are often not followed by the clinician or the patient.[40,41] The clinician was free to apply their usual care and adjust their therapy in the way they seemed fit.

The confidence intervals of the "medium-risk" and "high-risk" risks for persisting pain and disability show some overlap, which might suggest a lack of independence, but may also be the result of a lack of power. Furthermore, it has to be taken into account that clinicians applied 'usual care' and not the advised approach possibly influencing the outcome, which might explain the overlap.

Findings in the context of other literature

When comparing our results with the results from the UK study we have to keep in mind that the healthcare system is different between the countries. Despite these differences, we included, in line with the UK study, all LBP patients disregarding duration of complaints or previously provided healthcare. In contrast to the UK study, in our study not all patients were seen by their GP as the PT also included and treated patients via direct access.

The distribution in risk-profiles in our study was well comparable with the distribution in the UK study.[13] All other cohorts all had a shift towards "high-risk" at the expense of "low-risk".[16–19,21] For each increase in risk profile we found an increase in pain, PCS and TSK, this discriminant validity is also found in another studies.[42–44] Other validation studies such as the Finnish, German, French, and Persian followed the same

method as the initial UK study by using the Area Under the Curve (AUC) to determine the validity thus making it easier to compare.[16,18] In our study we refrained from using the AUC because we chose not to dichotomize the scores of the questionnaires. We used the Pearson's correlation coefficient giving us the correlation information needed, although this made it more difficult to compare our results to other studies. We compared individual questions with their reference questionnaire; the UK study used the total SBT score or the psychosocial subscale to calculate the AUC.

The quadratic weighted kappa for the re-test reliability in our study is lower than the one in the UK study (0.79 for the stable patients), but comparable to the German version (0.67).[13,18] This might be due to our small sample size of 19 compared to 295 and 410 in the previous mentioned studies. Our data is also more skewed towards "low-risk" as a result of a higher percentage of patients in this group, which influences the kappa. Besides using the quadratic weighted kappa we also calculated the specific agreement.[35] No other studies used this measurement therefore we can't compare results. Our findings are in accordance with all other studies that evaluated translations of the SBT to be a reliable and valid instrument.[13,15,16,18,19]

Strengths and limitations

The strength of this study is that we successfully translated the SBT into Dutch and determined the construct validity, reproducibility, predictive validity and content validity. The advised minimum sample size was met for the validity section. A limitation is that we had a relatively small sample size for the re-test reliability study. Also we were not able to compare patients between physical therapist and GP, as we did not have enough patients in both groups. Another limitation is that for practical reasons the patient filled out the baseline questionnaire after receiving the first treatment/consultation. The questionnaire is intended to be filled in before the first consultation/treatment because the patient might change its cognition and therefore influence the results.

Clinical and/or research implications

The STarT Back tool has been translated and validated for use in Dutch primary care. It can be used to, in an early stage, predict persisting disability. More important is that is can be used to match the patient to the advised treatment. Further research is needed to determine if this stratified care leads to a faster recovery and in its turn leads to lower healthcare consumption and lower costs. To further determine the predictive validity future studies might include a non-intervention (natural course) group.

CONCLUSION

The SBT is successfully translated in Dutch and according to the psychometric analysis it showed to be a sufficiently valid and reliable instrument.

ACKNOWLEDGEMENTS

The authors would like to thank all of the general practitioners and physical therapists that included patients for this study. A special thank goes out to Nynke Wildervank for her contribution and to Steven Constandse, Guido Iken, Joost van Broekhoven and Frans van der Kooij for their extra efforts to reach the needed sample size.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests regarding the publication of this paper. This study was undertaken with financial support of CZ healthcare insurance company and Dutch Arthritis Association.

REFERENCES

1. Vos T, Flaxman AD, Naghavi M, et al. 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*. 2012;380(9859):2163-2196. doi:10.1016/S0140-6736(12)61729-2.
2. Picavet HSJ, Schouten JS a G. Musculoskeletal pain in the Netherlands: Prevalences, consequences and risk groups, the DMC3-study. *Pain*. 2003;102(1-2):167-178. doi:10.1016/s0304-3959(02)00372-x.
3. Kamper SJ, Ostelo RWJG, Knol DL, Maher CG, de Vet HCW, Hancock MJ. Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. *J Clin Epidemiol*. 2010;63(7):760-766.e1. doi:10.1016/j.jclinepi.2009.09.009.
4. van Tulder M, Becker A, Bekkering T, et al. Chapter 3 European guidelines for the management of acute nonspecific low back pain in primary care. *Eur Spine J*. 2006;15(S2):s169-s191. doi:10.1007/s00586-006-1071-2.
5. Henschke N, Maher CG, Refshauge KM, et al. Prognosis in patients with recent onset low back pain in Australian primary care: inception cohort study. *BMJ*. 2008;337(jul07_1):a171. doi:10.1136/bmj.a171.
6. da C Menezes Costa L, Maher CG, Hancock MJ, McAuley JH, Herbert RD, Costa LOP. The prognosis of acute and persistent low-back pain: a meta-analysis. *CMAJ*. 2012;184(11):E613-E624. doi:10.1503/cmaj.111271.
7. Lambeek LC, van Tulder MW, Swinkels ICS, Koppes LLJ, Anema JR, van Mechelen W. The trend in total cost of back pain in The Netherlands in the period 2002 to 2007. *Spine (Phila Pa 1976)*. 2011;36(13):1050-1058. doi:10.1097/BRS.0b013e3181e70488.
8. Mafi JN, McCarthy EP, Davis RB, Landon BE. Worsening trends in the management and treatment of back pain. *JAMA Intern Med*. 2013;173(17):1573-1581. doi:10.1001/jamainternmed.2013.8992.
9. Martin BI, Deyo RA, Mirza SK, et al. Expenditures and health status among adults with back and neck problems. *JAMA*. 2008;299(6):656-664. doi:10.1001/jama.299.6.656.
10. Ma VY, Chan L, Carruthers KJ. Incidence, prevalence, costs, and impact on disability of common conditions requiring rehabilitation in the United States: stroke, spinal cord injury, traumatic brain injury, multiple sclerosis, osteoarthritis, rheumatoid arthritis, limb loss, and back pa. *Arch Phys Med Rehabil*. 2014;95(5):986-995.e1. doi:10.1016/j.apmr.2013.10.032.
11. Fritz JM, Brennan GP, Clifford SN, Hunter SJ, Thackeray A. An examination of the reliability of a classification algorithm for subgrouping patients with low back pain. *Spine (Phila Pa 1976)*. 2006;31(1):77-82. doi:10.1097/01.brs.0000193898.14803.8a.
12. Kent P, Keating JL. Classification in nonspecific low back pain: what methods do primary care clinicians currently use? *Spine (Phila Pa 1976)*. 2005;30(12):1433-1440. doi:00007632-200506150-00015 [pii].
13. Hill JC, Dunn KM, Lewis M, et al. A primary care back pain screening tool: Identifying patient subgroups for initial treatment. *Arthritis Rheum*. 2008;59(5):632-641. doi:10.1002/art.23563.
14. Foster NE, Hill JC, Hay EM. Subgrouping patients with low back pain in primary care: are we

getting any better at it? *Man Ther*. 2011;16(1):3-8. doi:10.1016/j.math.2010.05.013.

15. Bruyère O, Demoulin M, Beaudart C, et al. Validity and reliability of the French version of the STarT Back screening tool for patients with low back pain. *Spine (Phila Pa 1976)*. 2014;39(2):E123-E128. doi:10.1097/BRS.0000000000000062.
16. Abedi M, Manshadi FD, Khalkhali M, et al. Translation and validation of the Persian version of the STarT Back Screening Tool in patients with nonspecific low back pain. *Man Ther*. 2015:1-5. doi:10.1016/j.math.2015.04.006.
17. Piironen S, Paananen M, Haapea M, et al. Transcultural adaption and psychometric properties of the STarT Back Screening Tool among Finnish low back pain patients. *Eur Spine J*. February 2015. doi:10.1007/s00586-015-3804-6.
18. Karstens S, Krug K, Hill JC, et al. Validation of the German version of the STarT-Back Tool (STarT-G): a cohort study with patients from primary care practices. *BMC Musculoskelet Disord*. 2015;16:346. doi:10.1186/s12891-015-0806-9.
19. Morso L, Albert H, Kent P, et al. Translation and discriminative validation of the STarT Back Screening Tool into Danish. *Eur Spine J*. 2011;20(12):2166-2173. doi:10.1007/s00586-011-1911-6.
20. Pilz B, Vasconcelos RA, Marcondes FB. The Brazilian version of STarT Back Screening Tool – translation , cross-cultural adaptation and reliability *. 2014;18(5):453-461.
21. Luan S, Min Y, Li G, et al. Cross-cultural Adaptation, Reliability, and Validity of the Chinese Version of the STarT Back Screening Tool in Patients With Low Back Pain. *Spine (Phila Pa 1976)*. 2014;39(16):E974-E979. doi:10.1097/BRS.0000000000000413.
22. Bruyere O, Demoulin M, Brereton C, et al. Translation validation of a new back pain screening questionnaire (the STarT Back Screening Tool) in French. *Arch Public Heal*. 2012;70(1):12. doi:10.1186/0778-7367-70-12.
23. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila Pa 1976)*. 2000;25(24):3186-3191. http://www.ncbi.nlm.nih.gov/pubmed/11124735. Accessed March 25, 2016.
24. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales*. 5th ed. Oxford University Press; 2014. https://global.oup.com/academic/product/health-measurement-scales-9780199685219?cc=nl&lang=en&. Accessed March 25, 2016.
25. Apeldoorn A, Hooff ML van, Ostelo RWJG. De STarT Back Screening Tool. *Fysiopraxis (in Dutch)*. 2013;04:32-33. https://issuu.com/kngfdefysiotherapeut/docs/2013-04_fysiopraxis_april_2013.
26. Staal JB, Hendriks EJM, Heijmans M, et al. KNGF-richtlijn Lage rugpijn. 2013.
27. Chavannes AW, Mens JMA, Koes BW, et al. NHG-Standaard Aspecifieke lagerugpijn | NHG. *Huisarts Wet.* 2005;48(3):113-123. https://www.nhg.org/standaarden/volledig/nhg-standaard-aspecifieke-lagerugpijn#note-15. Accessed April 22, 2016.
28. Hjermstad MJ, Fayers PM, Haugen DF, et al. Studies comparing Numerical Rating Scales, Verbal Rating Scales, and Visual Analogue Scales for assessment of pain intensity in adults: a systematic literature review. *J Pain Symptom Manag*. 2011;41(6):1073-1093. doi:10.1016/j.jpainsymman.2010.08.016.

29. Brouwer S, Kuijer W, Dijkstra PU, Goeken LN, Groothoff JW, Geertzen JH. Reliability and stability of the Roland Morris Disability Questionnaire: intra class correlation and limits of agreement. *Disabil Rehabil*. 2004;26(3):162-165. doi:10.1080/09638280310001639713.
30. Roland M, Morris R. A study of the natural history of low-back pain. Part II: development of guidelines for trials of treatment in primary care. *Spine (Phila Pa 1976)*. 1983;8(2):145-150. doi:10.1007/978-1-4471-5451-8_59.
31. Vlaeyen JW, Kole-Snijders AM, Boeren RG, van Eek H. Fear of movement/(re)injury in chronic low back pain and its relation to behavioral performance. *Pain*. 1995;62(3):363-372. http://www.ncbi.nlm.nih.gov/pubmed/8657437.
32. Sullivan MJL, Bishop SR, Pivik J. The Pain Catastrophizing Scale: Development and validation. *Psychol Assess*. 1995;7(4):524-532. doi:Doi 10.1037/1040-3590.7.4.524.
33. Salen BA, Spangfort E V, Nygren AL, Nordemar R. The Disability Rating Index: an instrument for the assessment of disability in clinical settings. *J Clin Epidemiol*. 1994;47(12):1423-1435. http://www.ncbi.nlm.nih.gov/pubmed/7730851.
34. Terwee CB, Bot SDM, 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.
35. de Vet HCW, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's kappa. *Bmj-British Med J*. 2013;346(April):f2125. doi:Artn F2125Doi 10.1136/Bmj.F2125.
36. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed.; 1988. http://www.amazon.com/Statistical-Analysis-Behavioral-Sciences-Edition/dp/0805802835. Accessed March 25, 2016.
37. Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Phys Ther*. 2005;85(3):257-268. http://ptjournal.apta.org/content/85/3/257.long. Accessed March 21, 2016.
38. Hay EM, Dunn KM, Hill JC, et al. A randomised clinical trial of subgrouping and targeted treatment for low back pain compared with best current care. The STarT Back Trial Study Protocol. *BMC Musculoskelet Disord*. 2008;9:58. doi:10.1186/1471-2474-9-58.
39. Newell D, Field J, Pollard D. Using the STarT Back Tool: Does timing of stratification matter? *Man Ther*. 2015;20(4):533-539. doi:10.1016/j.math.2014.08.001.
40. Childs JD, Fritz JM, Wu SS, et al. Implications of early and guideline adherent physical therapy for low back pain on utilization and costs. *BMC Health Serv Res*. 2015;15(1):150. doi:10.1186/s12913-015-0830-3.
41. Bier JD, Kamper SJ, Verhagen AP, Maher CG, Williams CM. Predictors of non-adherence to guideline recommended care in acute low back pain. *Submitt Publ*.
42. Hill JC, Whitehurst DGT, Lewis M, et al. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet (London, England)*. 2011;378(9802):1560-1571. doi:10.1016/S0140-6736(11)60937-9.
43. Butera KA, Lentz TA, Beneciuk JM, George SZ. Preliminary Evaluation of a Modified STarT Back Screening Tool Across Different Musculoskeletal Pain Conditions. *Phys Ther*. February 2016.

doi:10.2522/ptj.20150377.

44. Fuhro FF, Fagundes FRC, Manzoni ACT, Costa LOP, Cabral CMN. Örebro Musculoskeletal Pain Screening Questionnaire Short-Form and STarT Back Screening Tool: Correlation and Agreement Analysis. *Spine (Phila Pa 1976)*. 2016;41(15):E931-E936. doi:10.1097/BRS.0000000000001415.

*Table 1, specific agreement**

Baseline (T0) \ Follow-up (T1)	Low	Medium	High
Low	7 (A)	2 (B)	0 (C)
Medium	1 (D)	4 (E)	0 (F)
High	0 (G)	4 (H)	1 (I)

** "Low-risk" A/(A+(B+C+D+G)/2) = 7/8,5 = 82.4%*

"Medium-risk" E/(E+(B+H+D+F)/2) = 4/7.5 = 53.3%

"High-risk" I/(I+(C+F+G+H)/2) = 1/3 = 33.3%

Figure 1, patient flow

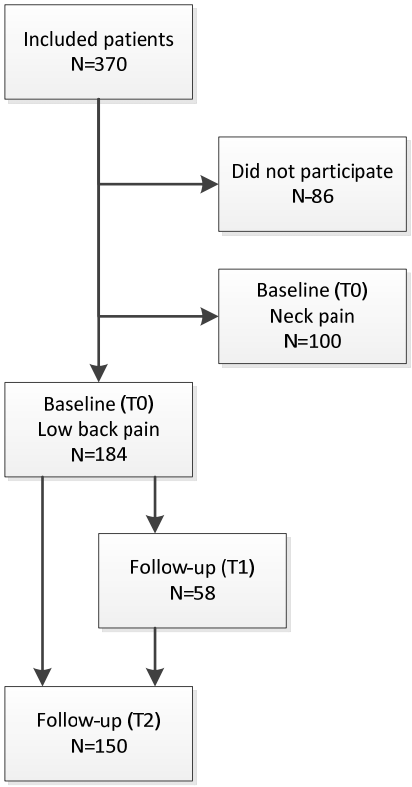

Table 2, baseline characteristics of the study population.*

	Study population (n=184)	**UK-validation sample** (n=500)
Female	103 (56.0)	293 (58.6)
Age in years, mean (SD)	44.6 (14.6)	45 (9.7)
SBT risk profile		
Low	96 (52.2)	234 (47.4)
Medium	70 (38.0)	186 (37.7)
High	18 (9.8)	74 (15.0)
Episode duration		
<1 month	53 (28.8)	83 (16.9)
1 to 3 months	26 (14.1)	94 (19.1)
4 to 6 months	12 (6.5)	77 (15.7)
7 months to 3 years	36 (19.6)	125 (25.5)
>3 years	57 (31.0)	112 (22.8)
SBT score, mean (SD)	3.60 (2.0)	3.83 (2.3)
Pain intensity		
Mild (0-5)	63 (34.2)	325 (66.1)
Moderate (5-7)	88 (47.8)	113 (23.0)
Severe (8-10)	33 (17.9)	54 (10.1)
Disability (RDQ), mean (SD)	9.5 (5.9)	9.1 (5.9)
Referred leg pain	54 (29.3)	303 (60.6)
Comorbid pain in neck/shoulder	124 (67.4)	276 (55.2)
Very or extremely bothered by back	94 (51.1)	276 (55.2)
Fear (TSK), mean (SD)	34.8 (7.1)	39.5 (6.9)
Catastrophizing (PCS), mean (SD)	13.7 (10.3)	

* Values are numbers (percentage) unless otherwise indicated. SBT = STarT Back Tool (0-9), RDQ = Roland Disability Questionnaire (0-24), TSK = Tampa Scale of Kinesiophobia (17-68), PSC = Pain Catastrophizing Scale (0-42). Pain intensity is measured on a Numeric Pain Rating Scale (0-10). UK Validation study preformed by Hill in2008[13]

Table 3, characteristics of patients in the risk profiles*

	Low risk	Medium risk	High risk
SBT, N (%)	96 (52.2)	70 (38.0)	18 (9.8)
RDQ	6.5 (4.9)	11.8 (5.1)	16.7 (3.1)
NPRS	5.2 (1.8)	6.5 (1.5)	7.2 (1.6)
TSK	32.4 (5.9)	35.4 (6.6)	44.6 (6.2)
PCS	10.0 (7.9)	15.0 (9.5)	28.6 (10.6)

* Values are mean scores (SD) unless otherwise indicated. SBT = STarT Back Tool (0-9), RDQ = Roland Disability Questionnaire (0-24), NPRS = Numeric Pain Rating Scale (0-10), TSK = Tampa Scale of Kinesiophobia (17-68), PSC = Pain Catastrophizing Scale (0-52).

Table 4; Pearssons correlation between the STarT Back Tool and their reference questionnaires*

SBT and reference	Correlation			
	A priori	r		Expected
Q1 - NPRS	r < 0.30	0.28	low	Yes
Q2 - NPRS	r < 0.30	-0.05	low	Yes
Q3 - RDQ	r ≥ 0.30	0.48	moderate	Yes
Q4 - RDQ	r ≥ 0.30	0.49	moderate	Yes
Q5 - TSK	r ≥ 0.30	0.38	moderate	Yes
Q6 - PCS	r ≥ 0.30	0.34	moderate	Yes
Q7 - PCS	r ≥ 0.30	0.28	low	No
Q8 - PCS	r ≥ 0.30	0.46	moderate	Yes
Q9 - NPRS	r ≥ 0.30	0.63	high	Yes

* r = Pearssons correlation, NPRS = Numeric Pain Rating Scale, RDQ = Roland Disability Questionnaire, TSK = Tampa Scale of Kinesiophobia, PCS = Pain Catastrophizing Scale.

Table 5, Three month follow-up results*

	Persisting pain		Persisting disability		Recovery	
	NPRS	RR (95% C.I.)	RDQ	RR (95% C.I.)	GPE	RR (95% C.I.)
Low Risk	3.14 (2.38)		3.67 (5.09)		2.28 (0.89)	
Medium Risk	3.38 (2.64)	1.59 (0.85 - 2.96)	5.34 (5.79)	1.80 (1.04 - 3.11)	2.53 (1.17)	0.96 (0.72 - 1.29)
High Risk	5.13 (2.68)	3.39 (2.31 - 6.76)	9.19 (7.54)	2.67 (1.44 - 4.93)	2.56 (0.96)	0.63 (0.33 - 1.22)

* Values are mean scores (SD) unless otherwise indicated. NPRS = Numeric Pain Rating Scale (0-10), RDQ = Roland Disability Questionnaire (0-24), GPE = General Percieved Effects (1-7)

Appendix 1

The Keele STarT Back Screening Tool

Patient name: _____ D ate: _____

Thinking about the **last 2 weeks** tick your response to the following questions:

		Disagree 0	**Agree** 1
1	My back pain has **spread down my leg(s)** at some time in the last 2 weeks	☐	☐
2	I have had pain in the **shoulder** or **neck** at some time in the last 2 weeks	☐	☐
3	I have only **walked short distances** because of my back pain	☐	☐
4	In the last 2 weeks, I have **dressed more slowly** than usual because of back pain	☐	☐
5	It's not really safe for a person with a condition like mine to be physically active	☐	☐
6	**Worrying thoughts** have been going through my mind a lot of the time	☐	☐
7	I feel that **my back pain is terrible** and **it's never going to get any better**	☐	☐
8	In general I have **not enjoyed** all the things I used to enjoy	☐	☐

9. Overall, how **bothersome** has your back pain been in the **last 2 weeks**?

Not at all	Slightly	Moderately	Very much	Extremely
☐ 0	☐ 0	☐ 0	☐ 1	☐ 1

Total score (all 9): _____ Sub Score (Q5-9):_____

Funded by Arthritis Research UK

A; the original English version of the STarT Back Tool

The STarT Back Screening Tool: Dutch Version

Rugscreenings Instrument

Auteur:
- ✓ *Oorspronkelijke versie:* Jonathan Hill et al. © Keele University 01/08/07 (http://www.keele.ac.uk/sbst/)
- ✓ *Nederlandse versie: M van Hooff, W van Lankveld, P Anderson, A Apeldoorn, F van Hartingsveld, R Ostelo (2011)*

Naam: _____ Datum: _____

Antwoord u alstublieft ieder onderdeel. Kruis bij ieder onderdeel het vakje aan dat op u van toepassing is. Soms is het moeilijk om tussen twee vakjes te kiezen, kruis dan het vakje aan dat uw probleem het beste beschrijft. Kruis niet meer dan één vakje per onderdeel aan!

Denk bij het beantwoorden van de volgende vragen telkens aan de situatie **in de laatste 2 weken.**

		Oneens 0	**Eens** 1
1	In de laatste 2 weken **straalde** mijn rugpijn wel eens **uit naar één of beide benen.**	☐	☐
2	In de laatste 2 weken heb ik wel eens pijn in mijn **schouder** of **nek** gehad.	☐	☐
3	Vanwege mijn rugpijn **liep** ik alleen **korte afstanden.**	☐	☐
4	In de laatste 2 weken **kleedde ik me trager** dan gewoonlijk **aan** vanwege mijn rugpijn.	☐	☐
5	Voor iemand in mijn toestand is het echt niet veilig om lichamelijk actief te zijn.	☐	☐
6	**Ongeruste gedachten** gingen vaak door mijn hoofd.	☐	☐
7	Ik vind dat mijn **rugpijn verschrikkelijk** is en ik geloof dat **het nooit meer beter zal worden.**	☐	☐
8	Over het geheel genomen heb ik **niet genoten** van alle dingen waar ik vroeger wel van genoot.	☐	☐

9. Over het geheel genomen, hoe hinderlijk was uw rugpijn in de laatste 2 weken?

In het geheel niet	Een beetje	Matig	Erg	Extreem
☐ 0	☐ 0	☐ 0	☐ 1	☐ 1

Totale uitslag (alle 9) : _____ **Sub Uitslag (Q5-9):**_____

B; the translated Dutch version of the STarT Back Tool