Research Paper

PAIN®

Reliability and responsivity of pain intensity scales in individuals with chronic pain

Pramote Euasobhon[a], Raviwon Atisook[a*], Kulsiri Bumrungchatudom[a], Nantthasorn Zinboonyahgoon[a], Nattha Saisavoey[b], Mark P. Jensen[c]

Abstract

Prior research supports the validity and short-term test–retest stability of 4 commonly used scales for assessing pain intensity (Visual Analogue Scale [VAS], 6-point Verbal Rating Scale [VRS-6], Numerical Rating Scale [NRS-11], and Face Pain Scale-Revised [FPS-R]). However, the relative stability and ability of these measures to detect changes in pain intensity over longer time periods have not yet been examined, although knowledge regarding these psychometric issues is important for selecting from among these measures. To address this knowledge gap, we administered these scales assessing worst and average pain intensity to 250 chronic pain outpatients on 2 occasions, a little over 6 weeks apart on average. All 4 scales were found to be valid for detecting decreases in pain, and the VAS, NRS-11, and FPS-R evidenced the most validity for detecting increases in pain. The NRS-11 and VAS evidenced better test–retest stability than the VRS-6 and FPS-R. Age affected the ability of the VRS-6 for detecting improvement in worst pain, as well as the ability of the VAS for detecting worsening in both worst and average pain. However, the psychometric properties of the scales were not influenced by education level. Overall, the NRS-11 emerged as showing the most sensitivity and stability. The FPS-R seems to be a good second choice to consider for samples of individuals who might have difficulty understanding or using the NRS-11.

Keywords: Face Pain Scale-Revised, Numerical Rating Scale, Pain assessment, Reliability, Verbal Rating Scale, Validity, Visual Analogue Scale

1. Introduction

Valid and reliable measures of pain intensity are essential for pain research and effective pain management.[14,21,22,29,32] Two critical properties of pain measures are as follows: (1) Their ability to provide the same estimate when no change in pain has occurred (ie, stability), and (2) their ability to detect change when a change has occurred (ie, sensitivity).[33] The most commonly used measures of pain intensity are the Visual Analogue Scale (VAS), the Verbal Rating Scale (VRS), the 11-point Numerical Rating Scale (NRS-11), and the Face Pain Scale-Revised (FPS-R).[20] Regarding stability, evidence supports the reliability of the VAS, NRS-11, and VRS over very brief periods of time, up to 24 hours.[1,4,13] However, to the best of our knowledge, the test–retest stability of the FPS-R has not yet been evaluated in samples of individuals who report that their pain has remained unchanged. In addition, the stability of the other 3 measures over periods of time longer than 24 hours has not yet been evaluated.

All of the research that has evaluated scale sensitivity—a key validity criterion—has examined this in individuals who have received pain treatment. In such studies, measures with more than 4 response options have all been shown to similarly responsive pain treatments.[3,5–7,10,19,22] However, little is known regarding the ability of these measures to detect increases in pain over time when such increases actually occur. In addition, although the ability of people to use different scales is sometimes shown to differ as a function of age and education level,[2,9,15,22,23,25] little research has examined the role of age and education level on measure stability or the ability of measures to detect changes in pain over time.

Given these considerations, the primary aim of this study was to conduct a head-to-head comparison of the VAS, VRS-5, NRS-11, and FPS-R with respect to their test–retest stability and sensitivity in a sample of individuals with chronic pain over the course of 4 weeks. Consistent with prior research, we hypothesized that all 4 measures would evidence at least adequate test–retest stability among those participants who reported that their pain was unchanged over time. We also hypothesized that the measures would evidence an ability to detect changes in pain both among those who reported that their pain became worse and those who reported that their pain improved. However, based on prior research showing that the measures with many response options (eg, the VAS) are more difficult for individuals to use,[12,22] we anticipated that the VAS would perform less well than the other measures with respect to stability and sensitivity. Finally, we examined the extent to which the measures of stability and sensitivity differed as a function of age and education level, anticipating that if any differences emerged, reliability and validity would be worse for the VAS than the other measures.

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

[a] Department of Anesthesiology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand, [b] Department of Psychiatry, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand, [c] Department of Rehabilitation Medicine, University of Washington, Seattle, WA, United States

**Corresponding author. Address: Department of Anesthesiology, Faculty of Medicine Siriraj Hospital, Mahidol University, 2 Wang Lang Rd, Bangkoknoi, Bangkok 10700, Thailand. Tel: (+66) 2-419-7995; fax: (+66) 2-411-3256. E-mail: raviwon.ati@mahidol.ac.th (R. Atisook).*

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.painjournalonline.com).

PAIN 163 (2022) e1184–e1191

http://dx.doi.org/10.1097/j.pain.0000000000002692

2. Methods

2.1. Participants

This single center longitudinal study was conducted in Siriraj hospital, Bangkok, Thailand. A written pamphlet was posted in the pain clinic, and clinic patients were encouraged to speak with the clinic staff if they were interested in participating. In addition, research staff were on site and approached patients to evaluate interest. Any patient who expressed an interest was then evaluated for eligibility. To participate in this study, a potential participant had to be 18 years old or older, report that they had chronic pain (ie, that the duration of their pain problem was ≥3 months[31]), and be able to read and speak Thai. Patients with cancer pain or any patient who was unable to communicate effectively in Thai were excluded. All potential participants who were eligible and who expressed an interest in participating were asked to read and sign a written study consent form and then asked to complete the initial set of measures (see below). All participants were then provided with hard copy versions of the measures (see **Fig. 1**), to ensure that they had both verbal instructions as well as visual information, to maximize their ability to respond to the measures. Anyone enrolled in the study who was unable to use or correctly respond to any 1 of the 4 measures (ie, unable to place a single mark on the VAS, see below), or who did not provide follow-up data 4 weeks after their initial assessment, was excluded from the analyses. A total of 250 consecutive outpatients with chronic pain and who provided data at both assessment points were enrolled into the study from August 2020 to June 2021.

2.2. Measures

2.2.1. Demographic and pain history variables

All participants were asked to complete a questionnaire assessing demographic data (ie, sex, age, education, and employment status) and pain history information (ie, diagnosis, pain duration, and cause of pain). For data analytic purposes (ie, to evaluate the effects of age and education level on the validity criteria evaluated in this study), participants were classified as being younger (≤60 year old) or older (>60 year old) and as having less (≤12 years) or more (>12 years) education.

2.2.2. Visual Analogue Scale

The VAS used here was a 10-cm straight line with "no pain at all" on 1 end and "pain as bad as it could be" on the other. Respondents are asked to make a mark on the line that represents their pain intensity level. The distance (in cm) between "no pain at all" end and the mark is the VAS rating of pain intensity.[14] In this study, the VAS (and all pain intensity measures) was used to assess both average and worst pain intensity.

2.2.3. 6-point verbal rating scale

The VRS-6 provides respondents with a list of which describes different levels of pain. The respondent is asked to mark the adjective which best represents their pain intensity.[14] The 6 descriptors in the VRS-6 used in this study translate into "no pain at all," "very mild," "mild," "moderate," "severe," and "pain as bad as it could be." Each of these descriptors is linked to a specific number (ie, 0, 1, 2, 3, 4, and 5), and the respondents' score is the number linked to the descriptor chosen.

2.2.4.11-point Numerical Rating Scale

The 11-point Numerical Rating Scale (NRS-11) asks respondents to select a number 0 to 10 that best represents their pain intensity, with 0 = "no pain at all" and 10 = "pain as bad as it could be."[11] The respondent's NRS score is the number they select.

2.2.5. Face Pain Scale-Revised

The FPS-R consists of line drawings of 6 faces with different expressions designed to represent a continuum of those associated with different levels of pain intensity, from "no pain at all" to "pain as bad as it could be." A numerical value from 0 to 10 (ie, 2, 4, 6, 8, and 10) is assigned to each face, and the respondent's pain intensity score is the number associated with the face that is chosen.[16]

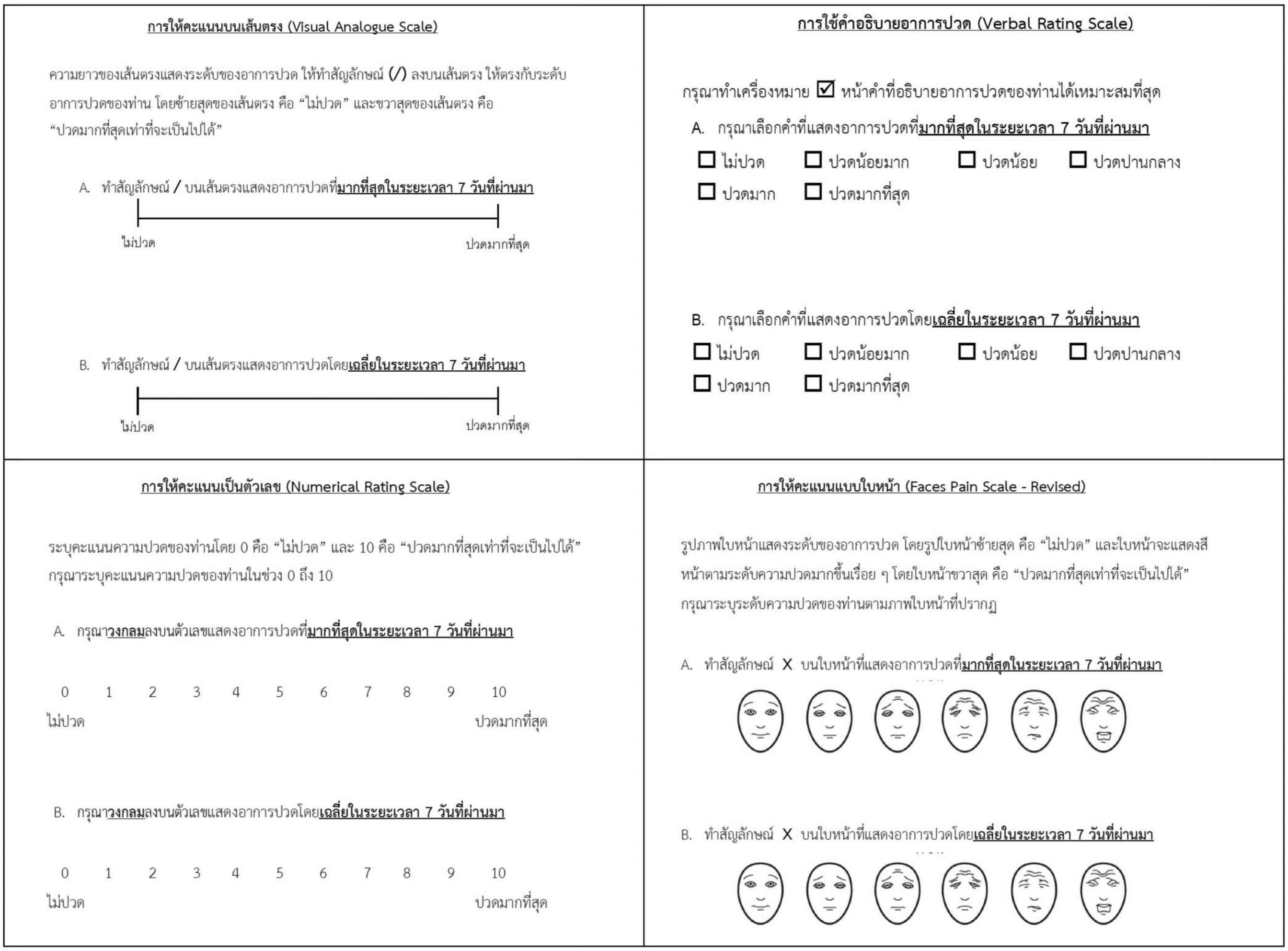
2.2.6. Patient global impression of change

The PGIC was used to assess the perceived amount of change in pain over time, from the initial assessment to the 4-week follow-up (see Procedures, below).[17] With the PGIC, respondents are asked to indicate whether (and how much) their pain improved, stayed the same, or got worse, using a 7-point categorical scale, with 1 = "very much improved," 2 = "much improved," 3 = "minimally improved," 4 = "no change," 5 = "minimally worse," 6 = "much worse," and 7 = "very much worse."[11,24] Based on the participant's response to this measure at the 4-week follow-up assessment point, we classified the participants into 1 of 3 categories: 1 = improved group (ie, those who responded with "very much improved," "much improved," or "minimally improved"), 2 = no change group (ie, those who responded with "no change"), or 3 = worsened group (ie, those who responded with "minimally worse," "much worse," or "very much worse"). In addition, to facilitate sensitivity analyses (described below), we also classified participants into 3 different groups, with an improved group made up of those who reported that their pain was "very much improved" or "much improved," a stable made up of those who reported that their pain was "minimally improved" or "minimally worse," or who reported "no change" in their pain, and a worsened group made up of those who reported that their pain was "much worse" or "very much worse."

2.3. Procedures

After study enrollment, participants were asked to complete the demographic and pain history questionnaire. They were then asked to read the instructions for each pain intensity scale, and rate their worst and average pain intensity during the past week of each of the 4 scales, using paper-and-pencil versions of all 4 pain scales (see **Fig. 1**). Participants were asked to rate their worst and average pain using 1 scale before moving on to the next. To minimize the potential biasing effects of the scale order, the scales were presented on separate pages and in 4 different orders, using a Latin square design. All 4 of the intensity measures used have been shown to be valid in individuals from Thailand with chronic pain.[2]

All participants were then scheduled for a follow-up visit. This follow-up appointment is usually scheduled for 4 weeks or later after the initial visit, depending on the patient's and clinician's schedules. However, because of the COVID situation during the time of data collection, many participants (52%) chose not to come to the hospital for their follow-up visit. Those participants who returned to the hospital visit in person were asked to again



การให้คะแนนบนเส้นตรง (Visual Analogue Scale)

ความยาวของเส้นตรงแสดงระดับของอาการปวด ให้ทำสัญลักษณ์ (/) ลงบนเส้นตรง ให้ตรงกับระดับอาการปวดของท่าน โดยซ้ายสุดของเส้นตรง คือ "ไม่ปวด" และขวาสุดของเส้นตรง คือ "ปวดมากที่สุดเท่าที่จะเป็นไปได้"

A. ทำสัญลักษณ์ / บนเส้นตรงแสดงอาการปวดที่**มากที่สุดในระยะเวลา 7 วันที่ผ่านมา**

ไม่ปวด ปวดมากที่สุด

B. ทำสัญลักษณ์ / บนเส้นตรงแสดงอาการปวดโดย**เฉลี่ยในระยะเวลา 7 วันที่ผ่านมา**

ไม่ปวด ปวดมากที่สุด

การใช้คำอธิบายอาการปวด (Verbal Rating Scale)

กรุณาทำเครื่องหมาย ☑ หน้าคำที่อธิบายอาการปวดของท่านได้เหมาะสมที่สุด

A. กรุณาเลือกคำที่แสดงอาการปวดที่**มากที่สุดในระยะเวลา 7 วันที่ผ่านมา**

□ ไม่ปวด □ ปวดน้อยมาก □ ปวดน้อย □ ปวดปานกลาง
□ ปวดมาก □ ปวดมากที่สุด

B. กรุณาเลือกคำที่แสดงอาการปวดโดย**เฉลี่ยในระยะเวลา 7 วันที่ผ่านมา**

□ ไม่ปวด □ ปวดน้อยมาก □ ปวดน้อย □ ปวดปานกลาง
□ ปวดมาก □ ปวดมากที่สุด

การให้คะแนนเป็นตัวเลข (Numerical Rating Scale)

ระบุคะแนนความปวดของท่านโดย 0 คือ "ไม่ปวด" และ 10 คือ "ปวดมากที่สุดเท่าที่จะเป็นไปได้"
กรุณาระบุคะแนนความปวดของท่านในช่วง 0 ถึง 10

A. กรุณา**วงกลม**ลงบนตัวเลขแสดงอาการปวดที่**มากที่สุดในระยะเวลา 7 วันที่ผ่านมา**

0 1 2 3 4 5 6 7 8 9 10
ไม่ปวด ปวดมากที่สุด

B. กรุณา**วงกลม**ลงบนตัวเลขแสดงอาการปวดโดย**เฉลี่ยในระยะเวลา 7 วันที่ผ่านมา**

0 1 2 3 4 5 6 7 8 9 10
ไม่ปวด ปวดมากที่สุด

การให้คะแนนแบบใบหน้า (Faces Pain Scale - Revised)

รูปภาพใบหน้าแสดงระดับของอาการปวด โดยรูปใบหน้าซ้ายสุด คือ "ไม่ปวด" และใบหน้าจะแสดงสีหน้าตามระดับความปวดมากขึ้นเรื่อย ๆ โดยใบหน้าขวาสุด คือ "ปวดมากที่สุดเท่าที่จะเป็นไปได้" กรุณาระบุระดับความปวดของท่านตามภาพใบหน้าที่ปรากฏ

A. ทำสัญลักษณ์ X บนใบหน้าที่แสดงอาการปวดที่**มากที่สุดในระยะเวลา 7 วันที่ผ่านมา**

B. ทำสัญลักษณ์ X บนใบหน้าที่แสดงอาการปวดโดย**เฉลี่ยในระยะเวลา 7 วันที่ผ่านมา**

Figure 1. The Thai versions of the VAS, VRS-6, NRS-11, and FPS-R used in this study. FPS-R, Face Pain Scale-Revised; NRS-11, 11-point Numerical Rating Scale; VAS, Visual Analogue Scale; VRS-6, 6-point Verbal Rating Scale.

rate their worst and average pain intensity in the past week. They were also asked to indicate the extent to which their pain changed since the initial visit using the PGIC. Those who did not return for an in-person follow-up visit were asked to respond to the follow-up questionnaires via email. The average time between the initial and follow-up assessments was 6.25 ± 1.77 weeks. As noted earlier, participants needed to have used all 4 pain intensity measures correctly to be included in the study analyses. In order for them to be deemed to use the measures correctly, they had to have (1) provided only a single response to each rating scale, (2) provide a response that was always within the possible ranges of a (eg, rating pain intensity from 0 to 10 only on the NRS-11; a "12" on the NRS-11 would be classified as an incorrect response), (3) avoided responding with a range of responses (eg, "3-5," in response to the NRS-11), and (4) to rate their worst pain intensity as equal to or larger than their average pain intensity. However, respondents were allowed to provide a response between 2 viable response options for any scale, if they chose to. For example, if they rated their pain intensity as "4.5" on the NRS-11 or as being something between a pair of faces or between a pair of verbal descriptors. For the FACES and VRS-6 scales, in this case, they were given the scores that would be the midpoint between the 2 response options (ie, a "3" if they indicated their pain was somewhere between the face associated with a "2" and the face associated with the "4").

The study was approved by the Siriraj Institutional Review Board, Faculty of medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand (SI607/2020), and was conducted in accordance with the Declaration of Helsinki.

2.4. Data analysis

Descriptive statistics for the demographic variables, pain history variables, and pain ratings were computed to describe the sample. Next, we examined the distributions of the study measures to determine if parametric or nonparametric analyses should be conducted to address the study aims. These analyses indicated nonnormal distributions, supporting the need to use nonparametric analyses. Primary analyses to evaluate the relative stability of the measures consisted of 2 sets of analyses for those participants who reported no change in pain over the 4-week period. Specifically, we (1) computed Spearman rank correlation coefficients (r_S) between the ratings obtained at the 2 assessment points and (2) tested for significant differences in pain intensity from the initial to the follow-up assessment, using a series of 8 Wilcoxon signed-rank tests (1 for each measure assessing worst and average pain). Although the specific cut-offs to determine whether or not a test–retest reliability coefficient is adequate depends on many factors (including the time period between assessments), there is general agreement that a coefficient of

0.70 or higher is an indication of adequate reliability.[8,28] Next, for the primary analyses to evaluate the relative sensitivity of the 4 measures for detecting change in pain over time, we conducted a series of Wilcoxon signed-rank tests comparing pain ratings from the initial to the follow-up assessment separately for participants who reported that their pain intensity increased (ie, rating their pain as being worse at the second time point) and who reported that their pain intensity decreased (ie, rating their pain as being improved at the second time point) over time. We also repeated all of these analyses in a series of sensitivity analyses, using the secondary group classifications, described above (ie, improved, stable, and worsened groups).

We then evaluated the effects of age and education level on the reliability and ability of the measures to detect changes in pain over time by conducting a series of 48 Mann–Whitney *U* tests (separate sets of analyses for the ratings of average and worst pain for participants reporting that their pain was better, did not change, or got worse and examining the rating of both worst and average pain for each of the 4 measures; ie, 2 × 3 × 2 × 4 = 48 analyses). In these analyses, time (initial vs follow-up assessments) and group (older vs younger, or more vs less educated) were the independent variables. The effects of interest with these analyses were the time × age and time × age interaction effects. Significant interactions, if they emerged, would suggest differences in change in pain over time for the particular scale used as a function of the moderator (ie, age or education level). In the event that a significant interaction emerged, we planned to examine the change scores to determine how the groups differed about the moderation variable. Data analyses were performed using PASW Statistics version 18 (SPSS, Inc, Chicago, IL).

3. Results

3.1. Demographic data and pain characteristics of the participants

The demographic and pain history information for the study participants are presented in **Table 1**. Sixty percent (N = 150) of the participants were women, and a majority of the study sample were in the younger age group. The mean age was 54 years. Most of the participants had neuropathic pain (71%, N = 178), and the second most common pain problem was muscle pain (43%, N = 108). Most of the participants (55%, N = 137) lived in an urban area. More than half of the study participants (69%, N = 173) were in the higher education group, with 41% (N = 102) having finished a bachelor degree or above. Seventy (28%) participants reported no change in average pain intensity over time, and 41 (16%) participants reported an increase in average pain intensity since their initial assessment. Regarding worst pain intensity, 76 (30%) participants reported no change and 41 (16%) participants reported an increase in worst pain intensity since their initial evaluation.

3.2. Relative reliability of the 4 measures of pain intensity

The results of the correlation analyses between the initial and follow-up ratings for the participants who reported no change in worst or average pain over 4 weeks are presented in **Table 2**. As can be seen, all of the scales were fairly stable, with coefficients being either higher than or very close to the cut-off of 0.70. However, the NRS-11 was more stable than the other scales, with test–retest coefficients of 0.73 and 0.78 for worst and average pain, respectively. The replication of these analyses with the improved, stable, and worsened sensitivity analysis groups are presented in Supplemental Tables S1–S4 (available at http://links.lww.com/PAIN/B650). As can be seen, these findings are generally consistent with the primary analyses, although the test–retest reliability coefficients of the measures were somewhat smaller in the sensitivity analyses (ie, 0.48-0.69) than the primary analyses (ie, 0.63-0.78).

The results of the analyses examining change in worst (76 participants) and average pain (70 participants) over time among the participants who reported no change in pain are presented in **Table 3**. As can be seen, the VAS, VRS-6, and NRS-11 evidenced stability for both pain intensity domains, as indicated by a lack of statistically significant differences between the ratings obtained at the different assessment points (*P*'s range, 0.057-0.801). The FPS-R of average pain, on the other hand, evidenced a statistically significant (*P* = 0.024) increase over this same time period in the no pain change sample.

Table 1

Descriptive information about the study participants (N = 250).

Variable	Mean ± SD	N (%)
Sex: female		150 (60)
Age (y)	54.27 ± 15.78	
Marital status		
Married		144 (58)
Unmarried		65 (26)
Separated or divorced or widowed		41 (16)
Employed (full or part time)		189 (76)
Place of residence		
Urban area		137 (55)
Provincial area		46 (18)
Rural area		67 (27)
Highest level of education		
No formal education		2 (1)
Primary school		43 (17)
Junior high school		32 (13)
Senior high school		32 (13)
Vocational certificate		39 (15)
Bachelor degree or above		102 (41)
Cause of pain or pain type*		
Neuropathic pain		178 (71)
Muscle pain		108 (43)
Bone and joint pain		115 (46)
Visceral pain		15 (6)
Other diagnosis of pain type		21 (8)
Unknown		44 (18)
PGIC of worst pain intensity		
1 very much improved		35 (14)
2 much improved		53 (21)
3 minimally improved		45 (18)
4 no change		76 (30)
5 minimally worse		15 (6)
6 much worse		16 (6)
7 very much worse		10 (4)
PGIC of average pain intensity		
1 very much improved		40 (16)
2 much improved		50 (20)
3 minimally improved		49 (20)
4 no change		70 (28)
5 minimally worse		17 (7)
6 much worse		16 (6)
7 very much worse		8 (3)

* Percents sum to more than 100% because participants could have more than 1 cause of pain or pain type.
PGIC, patient global impression of change.

Table 2

Correlation coefficients between the initial and follow-up assessments of the 4 scales assessing worst and average pain for the sample reporting no change in pain.

	Spearman rank correlation coefficients (r_s)	
Scale change	Worst pain (n = 76)	Average pain (n = 70)
VAS (0-10)	0.71*	0.72*
VRS-6 (0-5)	0.63*	0.76*
NRS-11 (0-10)	0.73*	0.78*
FPS-R (0-10, by 2 s)	0.67*	0.66*

* $P < .001$.

FPS-R, Face Pain Scale-Revised; NRS-11, 11-point Numerical Rating Scale; VAS, Visual Analogue Scale; VRS-6, 6-point Verbal Rating Scale.

3.3. Relative sensitivity of the 4 measures for detecting change in pain intensity over time

The Wilcoxon signed-ranks' results for the participants reporting improved pain (ie, less pain) and worse pain (ie, more pain) over time are presented in **Tables 4 and 5**. In the improved pain group (**Table 4**), all of the pain measures were valid for detecting decreases in pain intensity for both worst and average pain intensity (*P*'s < 0.001). In the worse pain group (**Table 5**), 3 of the measures (VAS, NRS-11, and FPS-R) were valid for detecting an increase in pain over time for both worst and average pain intensity (*P*'s < 0.05); the VRS-6, on the other hand, did not evidence a statistically significant difference over time for detecting an increase in either worst or average pain intensity.

3.4. Effects of age and education in the reliability and ability of the ratings scales to detect changes in pain over time

Among the participants who reported that their pain got worse, 3 significant moderating effects for age emerged. One was for the VAS assessing average pain ($U = 67.5$, $z = -3.05$, $P = 0.002$), and a second was for the VAS assessing worst pain ($U = 98.5$, $z = -2.34$, $P = 0.018$). In both of these cases, the interaction effects were explained by the ability of the VAS to detect change in pain was better for the older group than the younger group (median and interquartile range for average pain for the younger group were −0.60 [−1.65, 0.55], and these statistics for the older group were −3.40 [−4.58, −1.73]; these statistics for worst pain were −0.25 [−1.15, 0.45] for the younger group and −1.20 [−3.60, −0.55] for the older group). The third interaction effect was associated with the NRS-11 assessing worst pain group ($U = 103.5$, $z = -2.06$, $P = 0.042$). The NRS-11 was better able to detect an increase in pain among the older participants (change in pain on the NRS-11 = −2.00; interquartile range: −4.00, 0.00) than the younger participants (−1.00; 12.00, 0.00). Age did not moderate the effects of time on change in pain in the no pain change group or in improvement in pain group, and education level did not moderate ability to detect change in pain for any of the groups.

Table 3

Median, IQR, and *P*-values associated with change in pain intensity over time for the 4 scales assessing worst and average pain for the sample reporting no change in pain.

Variable	Median (IQR)		*P*
	Baseline	At 4 wk	
Worst pain intensity (n = 76)			
VAS (0-10)	6.80 (4.60, 8.05)	5.80 (4.33, 7.60)	0.066
VRS-6 (0-5)	3.00 (3.00, 4.00)	3.00 (3.00, 4.00)	0.057
NRS-11 (0-10)	7.00 (5.00, 8,00)	6.50 (5.00, 8.00)	0.552
FPS-R (0-10, by 2 s)	6.00 (4.00, 8.00)	6.00 (4.00, 8.00)	0.677
Average pain intensity (n = 70)			
VAS (0-10)	5.10 (3.75, 6.50)	5.00 (3.30, 6.43)	0.755
VRS-6 (0-5)	3.00 (2.00, 3.00)	3.00 (2.00, 3.00)	0.801
NRS-11 (0-10)	5.00 (3.75, 6.25)	5.00 (4.00, 7.00)	0.501
FPS-R (0-10, by 2 s)	4.00 (4.00, 6.00)	6.00 (4.00, 6.00)	**0.024**

FPS-R, Face Pain Scale-Revised; IQR, interquartile range; NRS-11, 11-point Numerical Rating Scale; VAS, Visual Analogue Scale; VRS-6, 6-point Verbal Rating Scale.

P values that are in bold face are all statistically significant at $P < 0.05$ or less.

Table 4

Median, IQR, and *P*-values associated with change in pain intensity over time for the 4 scales assessing worst and average pain for the sample reporting decreases in pain intensity.

Variable	Median (IQR)		*P*
	Baseline	At 4 wk	
Worst pain intensity (n = 133)			
VAS (0-10)	5.70 (4.20, 7.50)	4.30 (2.25, 5.85)	**<0.001**
VRS-6 (0-5)	3.00 (3.00, 4.00)	3.00 (2.00, 3.00)	**<0.001**
NRS-11 (0-10)	6.00 (5.00, 8,00)	5.00 (3.00, 6.00)	**<0.001**
FPS-R (0-10, by 2 s)	6.00 (4.00, 8.00)	4.00 (2.00, 6.00)	**<0.001**
Average pain intensity (n = 139)			
VAS (0-10)	4.80 (3.00, 6.30)	3.60 (1.80, 5.20)	**<0.001**
VRS-6 (0-5)	3.00 (2.00, 3.00)	3.00 (1.50, 3.00)	**<0.001**
NRS-11 (0-10)	5.00 (3.00, 7.00)	4.00 (2.00, 5.00)	**<0.001**
FPS-R (0-10, by 2 s)	4.00 (2.00, 6.00)	4.00 (2.00, 6.00)	**<0.001**

FPS-R, Face Pain Scale-Revised; IQR, interquartile range; NRS-11, 11-point Numerical Rating Scale; VAS, Visual Analogue Scale; VRS-6, 6-point Verbal Rating Scale.

P values that are in bold face are all statistically significant at $P < 0.05$ or less.

4. Discussion

To the best of our knowledge, this is the first study that has examined the relative validity of the 4 most commonly used pain intensity measures for detecting both increases and decreases in pain over time, and for showing stability in pain over time, using the same sample of participants. Information regarding these aspects of the scales' validity and reliability are essential for determining which measure to use and with which populations. We found that

Table 5

Median, IQR, and *P*-values associated with change in pain intensity over time for the 4 scales assessing worst and average pain for the sample reporting increases in pain intensity.

Variable	Median (IQR)		*P*
	Baseline	At 4 wk	
Worst pain intensity (n = 41)			
VAS (0-10)	6.70 (5.20, 8.25)	7.50 (6.60, 8.20)	**0.022**
VRS-6 (0-5)	4.00 (3.00, 4.00)	4.00 (3.00, 4.00)	0.734
NRS-11 (0-10)	7.00 (5.50, 8,25)	8.00 (7.00, 8.50)	**0.021**
FPS-R (0-10, by 2 s)	6.00 (5.00, 8.00)	8.00 (6.00, 8.00)	**0.035**
Average pain intensity (n = 41)			
VAS (0-10)	5.40 (3.30, 7.05)	6.60 (5.50, 7.80)	**0.001**
VRS-6 (0-5)	3.00 (3.00, 4.00)	3.00 (3.00, 4.00)	0.573
NRS-11 (0-10)	6.00 (4.00, 7.00)	7.00 (6.00, 8.00)	**<0.001**
FPS-R (0-10, by 2 s)	6.00 (4.00, 6.00)	6.00 (5.00, 8.00)	**0.007**

FPS-R, Face Pain Scale-Revised; IQR, interquartile range; NRS-11, 11-point Numerical Rating Scale; VAS, Visual Analogue Scale; VRS-6, 6-point Verbal Rating Scale.

P values that are in bold face are all statistically significant at $P < 0.05$ or less.

all 4 scales were similarly valid for detecting improvements in pain. However, only 3 (VAS, NRS-11, and FPS-R) seemed to be valid for detecting increases in pain over time. A different set of 3 (VAS, NRS-11, and FPS-R) evidenced adequate stability, in the subsample reporting no changes in pain intensity. In addition, we found that participant age moderated the ability of the VAS and NRS-11 to detect worsening in pain over time. These findings have important implications for decisions regarding which measure(s) to use for different purposes.

4.1. Selecting pain intensity measures when the goal is to be able to detect meaningful improvement

The findings indicate that all 4 of the scales examined are valid for being able to detect improvements in pain over time, at least for those individuals who are able to use each measure. That said, prior research in developed countries has shown that the VAS tends to evidence higher incorrect response rates than other scales in many populations, and the NRS-11 tends to have the lowest incorrect response rates across different samples.[2,15,18] However, in a developing country such as Nepal, Pathak et al.[23] found the NRS-11 had the highest incorrect response rate (64%), followed by VAS (33%). As a whole, it would seem reasonable to recommend that researchers and clinicians use the NRS-11 if the patient or study sample is able to use this scale and the goal is to be able to detect improvements in pain over time.[26] For those samples that include individuals who may not be able to use the NRS-11,[23] the findings indicate that the FPS-R would be an appropriate second choice, again, when the aim is to determine the efficacy of a pain treatment for reducing pain. The FPS-R should be considered as a first choice measure if a goal is to be able to compare the results of pain clinical trials across samples from different countries.[2]

4.2. Selecting pain intensity measures when the goal is to track both increases and decreases in pain

When the clinical or research goal is to track both increases and decreases in pain, the study findings suggest that the VRS-6 is less valid than the other scales for this purpose, given its inability to detect significant increases in pain in the subsample that reported increased pain over time. Although it is possible that the lower levels of responsiveness for the VRS-6 may have been due to the fewer number of response options for this measure, relative to the NRS-11 and VAS, this explanation seems unlikely, because the FPS-R has the same number of response options (ie, 6) as the VRS-6 and yet was also able to detect an increase in pain intensity in the group reporting worse pain.

In addition, the findings indicate that age moderates the effects of the VAS and NRS-11 for detecting pain getting worse over time; although interestingly, the direction of this moderating effect was inconsistent with the study hypothesis. Specifically, we found that the VAS and NRS-11 were better able to detect pain getting worse in the older subsample than the younger subsample. Either way, the findings point to the FPS-R as being better able than the NRS-11, VAS, or VRS-6 to detect both or either increases or decreases in pain over time, at least in the current sample.

4.3. Test–retest stability

In addition to being able to detect improvements or worsening in pain when these occur, a valid measure of pain intensity should be able to demonstrate stability over time when pain intensity stays the same. Here, we used 2 criteria to evaluate stability in the 4 measures among the group that reported no change in pain: (1) a lack of a significant time effect and (2) relative high test–retest stability coefficients. As far as we know, this is the first study to evaluate and compare the stability of the 4 pain intensity measures evaluated in a direct head-to-head comparison in the same sample of individuals with chronic pain. We found that the FPS-R failed the first criterion for average pain, as it detected a significant increase in average pain over time in the no change group. Regarding the second criterion, only 2 measures—the NRS-11 and VAS—met the a priori cut-off of showing a >0.70 test–retest stability coefficient in the subsample of individuals who reported no change in pain. Thus, the findings again point to the NRS-11 (r_s's $= 0.73$ and 0.78) as the best measure overall, as long, of course, that the sample is able to understand and use the NRS-11. Regarding this criterion alone, the VAS (r_s's $= 0.71$ and 0.72) would seem to be a reasonable second choice, again as long as the sample is able to use the measure (ie, are relatively young and educated[2,25]).

The findings regarding the reliability and validity of the measures were generally replicated in sensitivity analyses that included participants who rated their pain as "minimally improved" and "minimally worse" in the no change group, instead of the improved and worsened groups, respectively. The primary difference in results was a slight reduction in the reliability coefficients and validity results, consistent with what would be expected due to (1) including participants in the stable group whose reported their pain changed, even only a minimal amount, and (2) reducing the power to be able to detect change over time in the worsened and improvement groups.

4.4. Study limitations

The study has a number of limitations that should be considered when interpreting the results. First, all of the participants came from a single clinic in Bangkok, Thailand. We were unable to determine the extent to which the study participants' pain intensity levels are representative of other individuals with pain in Thailand because of the lack of Thai norms for the 4 measures. Relatedly, based on data from Thailand's Department of Interior, 21% of the Thai population lived in urban areas in 2019, and based on data from Thailand's Office of Education, 46% of the population would be classified as being in the higher education level group using the cut-offs we used here; our sample had more education and was more likely to live in an urban area than the Thai population in general. Thus, the generalizability of the study findings to other individuals with pain or living in other parts of Thailand (or even other parts of the world) is not known. It would be important to replicate the study using individuals with chronic pain sampled from other populations to determine the reliability of the results. Second, we did not include a measure of cognitive function in the study. Thus, we were unable to determine if cognitive function had a moderating impact on the study findings; for example, if the moderating effects of age found in the study were due to the effects of age on cognitive function. Future research in this area should include a measure of cognitive function if possible. Third, we did not assess analgesic medication use at either time point, so were unable to determine the effects of analgesic use on the psychometric properties of the measures. Forth, we used "average pain" as the label for the global pain domain assessed. Although this is the most commonly used descriptor for this pain intensity domain in pain research, one might reasonably question the ability of individuals to compute actual "average" pain over the previous 7 days. Future research should examine the possibility that alternative descriptors for this domain—such as "usual pain" or "overall pain"[27,30]—might improve the reliability and validity of the measures. Finally, we did not evaluate the participants' previous experience with, and knowledge about, the scales evaluated in this study. Li et al.[22] found that the reliability of pain rating scales improves with exposure to, and practice with, pain rating scales. Thus, it is possible that one of the reasons the NRS-11 performed so well in this study is that this is the pain intensity measure most often used in the Siriraj Hospital, where this study was conducted. It is possible that the other measures may have performed better if the participants had more experience with them or even if they were given more detailed instructions and practice with the measures than were provided in this study. Future research to evaluate the impact of instructions and practice on the psychometric qualities of pain measures is warranted. Moreover, we did not recruit every patient in the pain clinic but only patients who were interested to participate in this study, potentially resulting in selection bias.

4.5. Summary and conclusions

Despite the study's limitations, the findings provide new information regarding the relative validity and reliability of the 4 most commonly used measures of pain intensity across clinical and research settings. Overall, among those individuals who are able to use all 4 measures, the NRS-11 seems to be the most sensitive and stable measure. The FPS-R seems to have the second best psychometric qualities overall, except that it evidenced a lack of reliability for assessing average pain in the sample of participants reporting no change in pain. Additional research in other populations, including those from other countries and from other (ie, more rural) parts of Thailand, is needed, to determine the generalizability of the findings.

Conflict of interest statement

None of the authors have any conflicts of interest related to the topic of this paper.

Acknowledgements

The authors gratefully acknowledge the patients who generously agreed to participate in this study, Ms Julaporn Pooliam for assistance with the statistical analysis, and Ms Nattaya Bunwatsana for general research assistance. The authors also acknowledge financial support from the Siriraj research fund, Faculty of Medicine Siriraj Hospital, Mahidol University, grant number (R016331070), for the conduct of this study.

Appendix A. Supplemental digital content

Supplemental digital content associated with this article can be found online at http://links.lww.com/PAIN/B650.

Supplemental video content

A video abstract associated with this article can be found at http://links.lww.com/PAIN/B651.

Article history:
Received 19 December 2021
Received in revised form 27 April 2022
Accepted 28 April 2022
Available online 18 May 2022

References

[1] Alghadir AH, Anwer S, Iqbal A, Iqbal ZA. Test-retest reliability, validity, and minimum detectable change of visual analog, numerical rating, and verbal rating scales for measurement of osteoarthritic knee pain. J Pain Res 2018;11:851–6.

[2] Atisook R, Euasobhon P, Saengsanon A, Jensen MP. Validity and utility of four pain intensity measures for use in international research. J Pain Res 2021;14:1129–39.

[3] Bellamy N, Campbell J, Syrotuik J. Comparative study of self-rating pain scales in rheumatoid arthritis patients. Curr Med Res Opin 1999;15: 121–7.

[4] Bergh I, Sjostrom B, Oden A, Steen B. An application of pain rating scales in geriatric patients. Aging (Milano) 2000;12:380–7.

[5] Bolognese JA, Schnitzer TJ, Ehrich EW. Response relationship of VAS and Likert scales in osteoarthritis efficacy measurement. Osteoarthritis Cartilage 2003;11:499–507.

[6] Bolton JE, Wilkinson RC. Responsiveness of pain scales: a comparison of three pain intensity measures in chiropractic patients. J Manipulative Physiol Ther 1998;21:1–7.

[7] Breivik EK, Bjornsson GA, Skovlund E. A comparison of pain rating scales by sampling from clinical trial data. Clin J Pain 2000;16:22–8.

[8] Carlozzi NE, Boileau NR, Roche MW, Ready RE, Perlmutter JS, Chou KL, Barton SK, McCormack MK, Stout JC, Cella D, Miner JA, Paulsen JS. Responsiveness to change over time and test-retest reliability of the PROMIS and Neuro-QoL mental health measures in persons with Huntington disease (HD). Qual Life Res 2020;29:3419–39.

[9] Clark P, Lavielle P, Martinez H. Learning from pain scales: patient perspective. J Rheumatol 2003;30:1584–8.

[10] Conti PC, de Azevedo LR, de Souza NV, Ferreira FV. Pain measurement in TMD patients: evaluation of precision and sensitivity of different scales. J Oral Rehabil 2001;28:534–9.

[11] Farrar JT, Young JP Jr, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. PAIN 2001;94:149–58.

[12] Gagliese L, Weizblit N, Ellis W, Chan VWS. The measurement of postoperative pain: a comparison of intensity scales in younger and older surgical patients. PAIN 2005;117:412–20.

[13] Grossman SA, Sheidler VR, McGuire DB, Geer C, Santor D, Piantadosi S. A comparison of the Hopkins Pain Rating Instrument with standard visual analogue and verbal descriptor scales in patients with cancer pain. J Pain Symptom Manage 1992;7:196–203.

[14] Haefeli M, Elfering A. Pain assessment. Eur Spine J 2006;15(suppl 1): S17–24.

[15] Herr K, Spratt KF, Garand L, Li L. Evaluation of the Iowa pain thermometer and other selected pain intensity scales in younger and older adult cohorts using controlled clinical pain: a preliminary study. Pain Med 2007; 8:585–600.

[16] Hicks CL, von Baeyer CL, Spafford PA, van Korlaar I, Goodenough B. The Faces Pain Scale-Revised: toward a common metric in pediatric pain measurement. PAIN 2001;93:173–83.

[17] Hurst H, Bolton J. Assessing the clinical significance of change scores recorded on subjective outcome measures. J Manipulative Physiol Ther 2004;27:26–35.

[18] Jensen MP, Castarlenas E, Roy R, Tome Pires C, Racine M, Pathak A, Miro J. The utility and construct validity of four measures of pain intensity: results from a university-based study in Spain. Pain Med 2019;20: 2411–20.

[19] Jensen MP, Chen C, Brugger AM. Postsurgical pain outcome assessment. PAIN 2002;99:101–9.

[20] Jensen MP, Karoly P. Self-report scales and procedures for assessing pain in adults. In: Turk DC, Melzack R, eds. Handbook of pain assessment. New York: Guilford Press, 2011; 19–44.

[21] Kumar P, Tripathi L. Challenges in pain assessment: pain intensity scales. Indian J Pain 2014;28:61.

[22] Li L, Liu X, Herr K. Postoperative pain intensity assessment: a comparison of four scales in Chinese adults. Pain Med (Malden, MA) 2007;8:223–34.

[23] Pathak A, Sharma S, Jensen MP. The utility and validity of pain intensity rating scales for use in developing countries. Pain Rep 2018;3:e672.

[24] Perrot S, Lanteri-Minet M. Patients' Global Impression of Change in the management of peripheral neuropathic pain: clinical relevance and correlations in daily practice. Eur J Pain 2019;23:1117–28.

[25] Peters ML, Patijn J, Lame I. Pain assessment in younger and older pain patients: psychometric properties and patient preference of five commonly used measures of pain intensity. Pain Med 2007;8:601–10.

[26] Safikhani S, Gries KS, Trudeau JJ, Reasner D, Rudell K, Coons SJ, Bush EN, Hanlon J, Abraham L, Vernon M. Response scale selection in adult pain measures: results from a literature review. J Patient Rep Outcomes 2017;2:40.

[27] Silverberg JI. Validity and reliability of a novel numeric rating scale to measure skin-pain in adults with atopic dermatitis. Arch Dermatol Res 2021;313:855–61.

[28] Smit EB, Bouwstra H, Roorda LD, van der Wouden JHC, Wattel ELM, Hertogh CMPM, Terwee CB. A patient-reported outcomes measurement information system short form for measuring physical function during geriatric rehabilitation: test-retest reliability, construct validity, responsiveness, and interpretability. J Am Med Dir Assoc 2021;22:1627–32.e1.

[29] Sullivan MD, Ballantyne JC. Must we reduce pain intensity to treat chronic pain?. PAIN 2016;157:65–9.

[30] Tait RC, Chibnall JT, Krause S. The pain disability index: psychometric properties. PAIN 1990;40:171–82.

[31] Treede RD, Rief W, Barke A, Aziz Q, Bennett MI, Benoliel R, Cohen M, Evers S, Finnerup NB, First MB, Giamberardino MA, Kaasa S, Korwisi B, Kosek E, Lavand'homme P, Nicholas M, Perrot S, Scholz J, Schug S, Smith BH, Svensson P, Vlaeyen JWS, Wang SJ. Chronic pain as a symptom or a disease: the IASP classification of chronic pain for the International Classification of Diseases (ICD-11). PAIN 2019;160: 19–27.

[32] van Dijk JFM, van Wijck AJM, Kappen TH, Peelen LM, Kalkman CJ, Schuurmans MJ. Postoperative pain assessment based on numeric ratings is not the same for patients and professionals: a cross-sectional study. Int J Nurs Stud 2012;49:65–71.

[33] Williamson A, Hoggart B. Pain: a review of three commonly used pain rating scales. J Clin Nurs 2005;14:798–804.