USASP
US ASSOCIATION FOR THE STUDY OF PAIN

PUBLISHED BY
ELSEVIER

The Journal of Pain, Vol 25, No. 7 (July), 2024: 104474
Available online at www.jpain.org and www.sciencedirect.com

Optimizing Pain Intensity Assessment in Clinical Trials: How Many Ratings are Needed to Best Balance the Need for Validity and to Minimize Assessment Burden?

Rocío de la Vega,[*,†] Prasert Sakulsriprasert,[‡] Jordi Miró,[§,¶] and Mark P. Jensen[‖]

[*]*Faculty of Psychology and Speech Therapy, University of Málaga, Málaga, Spain,* [†]*Biomedical Research Institute of Málaga (IBIMA – Plataforma BIONAND), Málaga, Spain,* [‡]*Faculty of Physical Therapy, Mahidol University, Salaya, Nakhon Pathom, Thailand,* [§]*Department of Psychology, Universitat Rovira i Virgili, Carretera de Valls, Tarragona, Spain,* [¶]*Unit for the Study and Treatment of Pain – ALGOS, Research Center for Behavior Assessment (CRAMC), Tarragona, Spain,* [‖]*Department of Rehabilitation Medicine, University of Washington, Seattle, Washington, United States*

Abstract: **Pain intensity is the most commonly used outcome domain in pain clinical trials. To minimize the chances of type II error (ie, concluding that a treatment does not have beneficial effects, when in fact it does), the measure of pain intensity used should be sensitive to changes produced by effective pain treatments. Here we sought to identify the combination of pain intensity ratings that would balance the need for reliability and validity against the need to minimize assessment burden. We conducted secondary analyses using data from a completed 4-arm clinical trial of psychological pain treatments (N = 164 adults). Current, worst, least, and average pain intensity in the past 24 hours were assessed 4 times before and after treatment using 0 to 10 numerical rating scale-11. We created a variety of composite scores using these ratings and evaluated their reliability (Cronbach's alphas) and validity (ie, associations with a gold standard score created by averaging 16 ratings and sensitivity for detecting between-group differences in treatment efficacy). We found that for each measure, reliability increased as the number of ratings used to create the measures increased and that ratings from 3 or more days were needed to have adequately strong associations with the gold standard. Regarding sensitivity, the findings suggest that composite scores made up of ratings from 4 days are needed to maximize the chances of detecting treatment effects, especially with smaller sample sizes. In conclusion, using data from 3 or 4 days of assessment may be the best practice.**

Perspective: ***Composite scores made up of at least 3 days of pain ratings appear to be needed to maximize reliability and validity while minimizing the assessment burden.***

Trial registration: ***clinicaltrials.gov NCT01800604.***

Key words: *Pain assessment, clinical trial, assay sensitivity, reliability, validity*

Pain intensity is the most commonly used and recommended outcome domain in pain clinical trials.[1] In order to maximize the success of those trials—that is, in order to maximize the chances of detecting real treatment effects when present (ie, "sensitivity")—the measure of pain intensity used should be as sensitive as possible to those effects. This is especially true when investigators have limited resources for enrolling large samples into the trial.[2]

One method to increase the sensitivity of pain intensity assessment is to maximize the reliability of the measure by averaging multiple pain intensity ratings into a single composite score. Based on psychometric theory,[3] as long as the individual ratings are valid for assessing pain intensity, the greater the number of

Received October 13, 2023; Received in revised form January 9, 2024; Accepted January 11, 2024
Address reprint requests to Rocío de la Vega, PhD, Departament of Personality, Assessment and Psychological Treatment, Faculty of Psychology and Speech Therapy, University of Málaga. C/ Doctor Ortiz Ramos n°12, 29010 Málaga, Spain. E-mail: rocio.delavega@uma.es

1526-5900/$36.00

ratings that are included in a composite score, the more reliable and sensitive the measure should be. However, increasing the number of ratings also results in an increase in assessment burden and study costs. In addition, it can also increase the risk of missing data, which can introduce bias into a study.[4]

Prior research addressing these issues[4–8] suggests that only 1 or 2 measures of 24-hour recalled average pain may be needed to achieve adequate levels of responsivity with minimal assessment burden. However, this research has only used ratings of average pain intensity. It is possible that composite scores created from ratings of multiple pain intensity domains (ie, current pain, as well as the average, least, and worst pain in the past 24 hours) assessed at 1 time only may be as or more valid for detecting significant treatment effects as composite scores made up of 2 or more ratings of 24-hour recalled average pain intensity assessed on different days. If so, researchers would only need to assess pain intensity at a single time point in order to obtain highly reliable and sensitive measures.

Given these considerations, the current study aimed to identify the combination of pain intensity ratings that would provide the most reliability and validity while minimizing cost and assessment burden. To address this aim, we conducted a series of secondary analyses using data from a 4-arm clinical trial of psychological treatments for chronic pain.[9] We hypothesized that for each of 4 pain intensity domains (ie, current pain, worst pain, least pain, and average pain): 1) the reliability and validity of the measures would increase as the number of ratings included in a composite score increased and 2) composite scores made up of 2 ratings of average pain would be about as reliable and valid as composite scores made up of 3 or 4 ratings. In addition, based on the findings from prior research,[8] we hypothesized that the most valid composite measures would be those that included ratings of worst and average pain. Finally, we anticipated that, in general, composite scores using ratings from a greater number of days would evidence greater validity and reliability than measures made up of fewer days of ratings.

Methods

Procedures

In order to test the study hypotheses, and to test the replicability and extension of prior studies in new clinical samples, we conducted a secondary analysis using data from a published 4-arm randomized clinical trial of psychological treatments for chronic pain.[9] The trial received ethics approval by the University of Washington's Institutional Review Board. The participants in this trial received 1 of 4 treatments in 4 weekly 1-hour treatment sessions. The treatments that were compared were: 1) hypnotic cognitive therapy (ie, using hypnosis to change the meaning of pain; HYP-CT); 2) standard cognitive therapy (CT); 3) hypnosis focused on pain reduction (HYP), and 4) pain education (ED), which was the control condition. In the trial, participants were asked to complete measures of current pain intensity and 24-hour recall ratings of worst, least, and average pain on 4 occasions within a 1-week period at each assessment point, including before and after treatment. The primary outcome of the primary study was average pain intensity, computed as the arithmetic mean of 4 24-hour recall ratings of average pain. The primary endpoint was post-treatment.

In the original study, participants in all treatment conditions reported significant pretreatment to post-treatment improvements in the primary outcome, and one of the treatment conditions (HYP-CT) resulted in significantly larger pretreatment to post-treatment pain reductions than the control condition. In addition to the article presenting the findings regarding treatment efficacy,[9] 2 other articles have been published from this study. The first sought to identify treatment mediators,[10] and the second sought to identify treatment moderators.[11] None of these 3 papers focus on or address the hypotheses tested in the current secondary analyses.

Participants

One hundred and seventy-three adults with chronic pain participated in the clinical trial, and 164 of them provided complete data (ie, providing 4 ratings for each pain intensity domain 4 times before and after treatment). To be included in the study, they had to have low back pain or chronic pain secondary to one of the following chronic conditions: multiple sclerosis, spinal cord injury, acquired amputation, or muscular dystrophy. Full details about the study recruitment and assessment procedures can be found in the original article.[9] Only participants who provided complete data for the measures used in the current analyses at pretreatment and post-treatment are included in this study.

Measures

From the battery of measures administered in the original study, the following were selected from the pretreatment and post-treatment assessment points to describe the sample and address the hypotheses tested here.

Demographic Variables

Demographic information collected included age, self-identified sex, self-identified race, self-identified ethnicity, educational level, and employment status. These variables were assessed via telephone interview by trained research staff after obtaining informed consent.

Pain Intensity

Current pain intensity and worst, least, and average pain intensity in the past 24 hours were assessed 4 times in a 1-week period until either 1) ratings from 4 days were obtained or 2) the end of the 7-day window was reached. This was done at pretreatment and post-treatment via telephone interviews conducted by trained research staff using 0 ("No Pain") to 10 ("Pain as bad as you can imagine") numerical rating scale-11.[12] Sixteen ratings in all were obtained at each assessment point. In addition, we

used several procedures to minimize the chances of missing data. This included scheduling the interviews in advance, training the research staff in rapport-building strategies (eg, use of reflective listening), and providing multiple chances to obtain the 4 ratings (ie, 7 days).

Data Analyses

Descriptive Statistics

We first computed descriptive statistics for the demographic variables to describe the sample.

Composite Scores

We created a total of 60 pain intensity scores to evaluate in the current study from the 16 individual ratings obtained. The first 4 of these were the single ratings of current pain and 24-hour recall ratings of worst, least, and average pain intensity obtained on the first of the 4 assessment days. The next 4 were averages of the 2 ratings of each of these intensity domains obtained on assessment days 1 and 2. The next 4 were averages of 3 ratings of each intensity domain obtained on days 1, 2, and 3, and the next 4 were averages of 4 ratings of each intensity domain obtained on days 1, 2, 3, and 4. Next, we computed 24 mean intensity scores from averages of each possible pair of pain intensity domains. The first 6 consisted of the average of each possible intensity domain pair from the first assessment day (eg, an average of current pain and least pain from assessment day 1, an average of current pain and worst pain from the assessment day 1, etc). The next 6 consisted of an average of each possible intensity domain pair from the first and second day (eg, an average of current pain ratings and the least pain ratings from days 1 and 2, etc). The next 6 consisted of averages of each possible pain intensity domain pair from the first, second, and third assessment day, and the final 6 consisted of each possible pain intensity domain pair from all 4 assessment days. Next, 16 different composite scores were created by averaging the ratings of each possible trio of domains from the first day (eg, an average of the current, least, and worst pain ratings from assessment day 1), the first 2 days, the first 3 days, and all 4 days (eg, an average of all 4 current pain, least pain, and worst pain ratings). Finally, we created 4 composite scores from averages of all 4 pain intensity domains from assessment days 1, 2, 3, and 4. As noted in the introductory section, based on psychometric theory,[3] we anticipated that the last of these composite scores—a score created by averaging all 16 different pain intensity ratings—was likely to be the most reliable and responsive of the 60 measures created, and was therefore used as the "gold standard" validity criterion for the study.

Reliability

We computed Cronbach's alphas for each of the 56 composite scores (ie, all but the 4 measures made up of single ratings) to estimate the relative reliability of these scores.

Validity

Next, as the first indicator of validity, we computed correlation coefficients between each of the 60 measures and the gold standard score. As the second indicator of validity, we conducted a series of 60 1-way analysis of covariances (ANCOVAs) (using the same analysis strategy as used in the primary paper; see Jensen et al[9]), with the sex assigned at birth and baseline pain intensity as the covariates, group (ED, HYP, HYP-CT, and CT) as the independent variable, and change in pain intensity (ie, pretreatment score minus post-treatment score) as the dependent variable.

From this analysis, we computed the effect sizes (eta-squared) for the groups as well as the number of subjects that would be needed in each condition to be able to detect a significant group effect, given this effect size and assuming a power of .80 and a *P* value of .05.

Finally, given that the HYP-CT condition evidenced the largest improvement in pain intensity in the original study, and this improvement was significantly greater than the ED control condition (as had been hypothesized), we conducted a series of 60 ANCOVAs similar to those described above, except that the group effect only compared ED (ie, the active control condition) with HYP-CT. From this analysis, we computed the effect size (chi-squared) and also the number of subjects that would be needed in each condition to be able to detect a significant group effect, with the same assumptions as before. We used SPSS (IBM, Chicago Ilios)[13] to compute the descriptive statistics and conduct the ANCOVAs, and G*Power 3.1.9.7 version (Heinrich-Heine-Universität, Düsseldorf, Germany)[14] to compute the sample sizes needed to detect significant group differences given the different effect sizes for these analyses. Specifically, we calculated the sample sizes associated with 1) an omnibus test for between-group differences for all 4 treatment conditions and 2) a between-group test comparing the control condition (ED) and the treatment that had the largest effects (HYP-CT), taking into account the number of treatment conditions, the correlations between the pretreatment and post-treatment measures (coefficient = .5), and the number of covariates, and assuming a power of .80 and an alpha of .05.

Results

Sample Description

Only participants who provided all 32 of the pain ratings (ie, 16 ratings assessed at pretreatment and 16 ratings assessed at post-treatment) and responses to all of the validity criterion variables used in the current study were included in the analyses. One of the 173 participants enrolled in the study did not provide an average pain rating on 1 of the assessment days, and 2 of the 166 participants who provided post-treatment data did not provide an average pain rating on 1 of the assessment days. This left 164 individuals who provided complete data and who therefore provided the data for the current analyses. They were adults with chronic pain (58% women, mean age 55 years) living in the United States. See Table 1 for more details regarding the demographic and pain information for the study sample.

Table 1. Demographic Characteristics

VARIABLE	MEAN (SD) OR NUMBER (%)
Age in years	55.43 (12.64)
Sex assigned at birth	
Men	68 (42%)
Women	96 (58%)
Self-identified race	
American Indian/Alaskan Native	4 (2%)
Asian	7 (4%)
Black or African American	16 (10%)
Other Pacific Islander/Native Hawaiian	1 (1%)
White	130 (79%)
More than one race	6 (3%)
Other*	5 (3%)
Self-identified ethnicity	
Hispanic or Latino	4 (2%)
Not Hispanic or Latino	157 (96%)
Not reported	3 (2%)
Highest level of education	
Grade 9 or less	0 (0%)
Grade 10 to 11 (some high school)	3 (2%)
High school graduate or GED	21 (13%)
Vocational or technical school	7 (4%)
Some college	43 (26%)
College graduate	55 (34%)
Graduate school or professional school	35 (21%)
Employment status†	
Employed full-time	30 (18%)
Employed part-time	18 (11%)
Attending school or vocational training full-time	3 (2%)
Attending school or vocational training part-time	1 (1%)
Retired	57 (35%)
Homemaker	8 (5%)
Unemployed due to pain	8 (5%)
Unemployed due to disability	60 (37%)
Unemployed for other reasons	7 (4%)

Abbreviations: GED, General Equivalence Diploma; SD, standard deviation.
*Other races included European/Native American, French Cherokee, Hispanic, and Native Mexican.
†Employment: total percentage exceeds 100 because some subjects responded in more than 1 category.

Reliability

Reliability of the Composite Measures of Pain Intensity Domains as the Number of Items Increases

Consistent with the study hypothesis regarding reliability based on psychometric theory, there was an increase in Cronbach's alpha values as the number of items of the composite scores increased for each pain intensity domain. For example, the internal consistency coefficients for composite scores made up of 2, 3, and 4 ratings of current pain intensity were alpha = .85, .87, and .88. These same internal consistency coefficients for least, worst, and average pain intensity were: 1) least pain: alpha = .88, .92, and .93; 2) worst pain: alpha = .81, .85, and .87; and 3) average pain: alpha = .86, .91, and .92.

However, as can be seen by observing the changes in the internal consistency of the scales made up from 2, 3, and 4 ratings, although there was an increase in internal consistency when the number of ratings increased from 2 to 3 and also from 3 to 4, there was sometimes a slight *decrease* in reliability when the number of rating *domains* increased from 2 to 3.

Reliability of Measures as a Function of the Pain Intensity Domain Being Rated

The pain intensity domain that evidenced the most reliability was the least pain, followed by average pain, current pain, and worst pain. Thus, composite scores made up of current and least pain intensity ratings (from days 1 to 4, internal consistencies of alpha = .84, .91, .94, and .95) were larger than those made up of least and worst pain intensity ratings (alpha = .56, .82, .88, and .90). Given the standard ranges for determining that internal consistency is good (.80–.89) or excellent (.90–.99), we found that only 6 of the 11 composite scores (54%) created from the ratings from just 1 day met the criteria for being good and none met the criteria for being excellent. Reliability increased for the composite scores created using the ratings from 2 days, with 8 (53%) being good and 7 (47%) being excellent. Four (26%) and 11 (74%) of the composite scores created using the ratings from 3 days were good and excellent, respectively, and 2 (13%) and 13 (87%) of the composite scores created using the ratings from 4 days were good and excellent, respectively. See Table 2 for details.

Validity

Association With the Study Gold Standard as the Number of Items Increases

Similar to the findings with respect to reliability, and consistent with both psychometric theory and the study hypothesis, as the number of items assessing any single pain domain increased, the association with the gold standard measure of pain intensity used for this study increased. For example, the Pearson *r* correlation coefficients for the 1-day rating and 2-day, 3-day, and 4-day composite scores for current pain were .58, .68, .74, and .84. These same coefficients for average pain, least pain, and worst pain were .60, .65, .70, and .80, .56, .62, .69, and .78, and 51, .57, .61, and .72, respectively. All of the composite scores made up of an average of the ratings of *all four* pain intensity domains, including the composite score made up of an average of these ratings from just 1 day (.89), were $>.80$. Otherwise, for the individual pain intensity domains and composite scores made up of 2 or 3 pain intensity domains, the correlation coefficients between the composite measures and the gold standard was not $\geq.80$ for any other composite scores made up of 2 or 3 pain intensity days from 2 or 3 days; most (12 or 80%) were $\geq.80$ for the composite measures made up of ratings from 4 days. The strongest associations (i., range, .98–.99) were found for the composite scores created by averaging ≥3 of the 4 pain intensity domains from 4 days of ratings. See Table 3 for details.

Table 2. Internal Consistency (Cronbach's Alpha) as a Function of Number of Days of Ratings and Pain Intensity Domains

MEASURE ASSESSING	NUMBER OF DAYS 1	2	3	4
	Single pain domains			
Current pain	-	.85	.87	.88
Least pain	-	.88	**.92**	**.93**
Worst pain	-	.81	.85	.87
Average pain	-	.86	**.91**	**.92**
	Two domain composites			
Current/Least	.84	**.91**	**.94**	**.95**
Current/Worst	.68	.85	.89	**.91**
Current/Average	.73	.89	**.92**	**.94**
Least/Worst	.56	.82	.88	**.90**
Least/Average	.73	.88	**.93**	**.95**
Worst/Average	.87	**.91**	**.93**	**.94**
	Three domain composites			
Current/Least/Worst	.79	**.90**	**.93**	**.94**
Current/Least/Average	.84	**.92**	**.95**	**.96**
Current/Worst/Average	.82	**.91**	**.94**	**.95**
Least/Worst/Average	.80	**.90**	**.93**	**.95**
	Four domain composites			
Current/Least/Worst/Average	.85	**.93**	**.95**	**.96**

NOTE. Internal consistency coefficients in the excellent range (ie, ≥.90) are in *boldface text*.

Table 3. Association (Pearson Correlations) Between the Pain Intensity Measures and the Study Gold Standard

MEASURE ASSESSING	NUMBER OF DAYS 1	2	3	4
	Single pain domains			
Current pain	.58	.68	.74	.84
Least pain	.56	.62	.69	.78
Worst pain	.51	.57	.61	.72
Average pain	.60	.65	.70	.80
	Two domain composites			
Current/Least	.61	.69	.75	.84
Current/Worst	.63	.70	.75	.86
Current/Average	.66	.72	.77	.86
Least/Worst	.64	.70	.75	.85
Least/Average	.65	.69	.75	.85
Worst/Average	.59	.64	.68	.79
	Three domain composites			
Current/Least/Worst	.66	.72	.77	**.99**
Current/Least/Average	.66	.72	.77	**.98**
Current/Worst/Average	.65	.70	.76	**.99**
Least/Worst/Average	.65	.70	.75	**.99**
	Four domain composite			
Current/Least/Worst/Average	.89	**.95**	**.98**	1.00

NOTE. Validity coefficients ≥.90 are in *boldface text*.

Table 4. Effect Sizes (Eta^2) and Sample Sizes (N) Needed to Detect Significant Group Effects for the Omnibus ANOVAs Testing for Overall Group Differences

MEASURE ASSESSING	NUMBER OF DAYS 1	2	3	4
	Single pain domains (Eta^2/N)			
Current pain	.28/144	.26/166	.23/211	.20/277
Least pain	.19/306	.19/306	.16/430	.15/489
Worst pain	.20/277	.20/277	.18/341	.16/430
Average pain	.19/306	.18/341	.16/430	.12/762
	Two domain composites (Eta^2/N)			
Current/Least	.20/277	.29/134	.25/179	.27/154
Current/Worst	.16/430	.30/126	.20/277	.23/211
Current/Average	.14/561	.27/154	.19/306	.20/277
Least/Worst	.13/650	.25/179	.19/306	.22/230
Least/Average	.13/650	.22/230	.15/489	.20/277
Worst/Average	.16/430	.29/134	.21/252	.20/277
	Three domain composites (Eta^2/N)			
Current/Least/Worst	.14/561	.27/154	.20/277	.23/211
Current/Least/Average	.13/650	.25/179	.19/306	.22/230
Current/Worst/Average	.13/650	.28/144	.18/341	.20/277
Least/Worst/Average	.12/762	.24/194	.17/382	.20/277
	Four domain composite (Eta^2/N)			
Current/Least/ Worst/ Average	.12/762	.25/179	.18/341	.21/252

Ability to Detect a Significant Treatment Effect as the Number of Items Increases

In general, and consistent with the study hypotheses, the ability to detect significant effects with an omnibus test comparing the 4 active treatments (see Table 4) and significant between-group effects between ED and HYP-CT (see Table 5) improved as the number of items making up the pain outcome scale increased. Thus, the effect sizes and number of subjects needed to detect significant effects for the omnibus (Eta^2) and between-group (Cohen's *d*) analysis of variance (ANOVA) analyses ranged from .19 to .28 (N range, 144–306) and .24 to .74 (N range, 17–139) for the single item measures of the 4 domains assessed on just 1 day. These same statistics were .18 to .26 (N range, 166–341) and .23 to .71 (N range, 19–151) using the composite scores averaged over 2 days, .16 to .23 (N = 211–430) and .28 to .58 (N = 26–103) using the composite scores averaged over 3 days, and .12 to .20 (N = 277–762) and .39 to .67 (N = 20–54) using the composite scores averaged over 4 days.

Relatedly, and using a cutoff of 50 subjects per group (ie, 200 subjects for the omnibus 4-group comparison and 100 subjects for the 2-group between-group comparison), the number of times significant effects would have been detected using the set of 15 scores computed from just 1 day of ratings was 1 (7%) for the test of an overall omnibus group effect among the 4 treatment conditions and 14 (93%) for the 2-group comparison. These same statistics for scores computed from 2, 3, and 4 days of ratings were: 1) 2 days—11 (73%) and 14 (93%); 2) 3 days—1 (7%) and 14 (93%), and 3) 4 days—1 (7%) and 15 (100%), respectively.

Discussion

This study aimed to identify the number of pain ratings that are needed to balance the goals of maximizing

Table 5. Effect Sizes (Cohen's d) and Sample Sizes (N) Needed to Detect Significant Group Effects Comparing the ED and HYP-CT Groups

	NUMBER OF DAYS			
MEASURE ASSESSING	1	2	3	4
	Single pain domains (Cohen's d/N)			
Current pain	.35/67	.38/57	.40/52	.54/30
Least pain	.74/17	.71/19	.58/26	.67/20
Worst pain	.24/139	.23/151	.28/103	.39/54
Average pain	.44/43	.39/54	.42/47	.42/47
	Two domain composites (Cohen's d/N)			
Current/Least	.56/28	.56/28	.51/33	.65/21
Current/Worst	.34/70	.33/75	.37/60	.50/34
Current/Average	.47/38	.43/45	.46/40	.53/31
Least/Worst	.50/34	.47/38	.46/40	.57/27
Least/Average	.63/23	.59/25	.54/30	.59/25
Worst/Average	.36/63	.32/79	.35/67	.39/54
	Three domain composites (Cohen's d/N)			
Current/Least/Worst	.49/35	.47/38	.46/40	.60/25
Current/Least/Average	.58/26	.54/30	.53/31	.60/25
Current/Worst/Average	.40/52	.37/60	.40/52	.47/38
Least/Worst/Average	.50/34	.46/40	.46/40	.51/33
	Four domain composite (Cohen's d/N)			
Current/Least/Worst/Average	.51/33	.47/38	.47/38	.54/30

reliability and validity while minimizing assessment burden. The hypothesis that "more [ratings] are better" derived from psychometric theory was supported. However, the hypothesis that composite scores made from 2 days would provide similar reliability and validity as those made from 3 or 4 days was not supported, nor was the hypothesis that composite scores made up of ratings of worst and least pain would provide greater validity than composite scores made up of ratings of least or current pain. Instead, we found that 1) composite scores made up of ratings from 4 days were more sensitive to between-group differences than composite scores made up of ratings from 3 days or fewer and 2) least pain ratings were most sensitive to between-group differences.

Composite Scores From 4 Days of Data Were Most Sensitive

The findings here contrast with prior research suggesting that assessing pain on 3 or more days does not add benefits to the validity of a composite score.[4–8] The reasons for the discrepancy in findings are not readily apparent. They may be due to the possibility that ratings from more days are needed to adequately obtain high levels of reliability and validity for some pain populations and not others. Until more is known about these issues, researchers would be wise to continue to use ratings from at least 4 days in order to maximize reliability as well as the ability to detect real differences.

Least Pain was the Most Sensitive Pain Domain

Contrary to our hypothesis, "least pain" was the domain with the highest validity for detecting significant between-group differences. Studies comparing the reliability and validity of least pain ratings to ratings of other pain intensity domains are rare and have produced mixed results.[8] One study with adults with spinal cord injury found that a composite score made up of the least pain ratings evidenced more internal consistency than composite scores made up of worst, average, or current pain ratings.[15] Another study focusing on the validity of recalled pain scores in patients with cancer[16] found that the least pain ratings contributed the most to a composite score. Interestingly, a study with older adults found that a combination of usual/least pain had the best predictive validity for adults without cognitive impairment, but for those with impairment, a usual/worst pain composite score evidenced the best validity.[17]

In published clinical trials, the most commonly assessed pain intensity outcome domains are average and worst pain, but there are several studies that included ratings of least pain as an outcome measure. Some of them found significant between-group differences in this domain along with others, such as worst or average pain,[18–23] and others did not find between-group differences in the least pain while finding significant effects for other pain intensity domains.[24,25] In summary, no single pain intensity domain appears to be consistently more reliable or sensitive to the effects of pain treatment than others. The current findings, as well as the lack of consistency on this issue, suggest that the common exclusion of least pain intensity as an outcome should be reconsidered; it would appear most reasonable to assess *all* 4 intensity domains and use a composite measure made up of these as a primary outcome. Clinicians should also consider assessing least pain (in addition to assessing average and worst pain) when monitoring the benefits of the treatments they prescribe, as it is possible that treatment could result in reductions in this intensity domain but not others for some patients; if least pain intensity is not assessed, then the beneficial effects of treatment might go undetected.

Sample Size Considerations

Focusing on the sample sizes needed to detect between-treatment effects, we found that relatively large sample sizes would have been needed to detect significant effects when there are multiple arms that all produce benefits. For example, the trial for this study[9] would have needed 4.5 times the number of participants it had to be able to detect a significant omnibus group effect, even when between-group differences actually existed in the sample.

The latest Cochrane collaboration reviews and meta-analyses for psychological interventions for individuals with chronic pain show that for adults[26] sample sizes range from 24 to 232 per arm (with an average of 71 at baseline). However, for children, the latest World Health Organization (WHO) review and meta-analysis[27] showed that for psychological interventions, sample sizes usually range from 26 to 100, with only 10% of the studies having

samples > 100 per arm across all treatment conditions. As a result, and even when significant between group differences likely exist, only a small percent of the published studies would appear to have enough power to detect significant group effects. This suggests that several design features may be needed in order to avoid type II errors. Firstly, researchers should consider including a no-treatment condition as the primary control condition, rather than an active control condition, as was used in the trial that provided data for the current analyses. Secondly, it would seem inappropriate to hypothesize a significant group effect for an omnibus group effect, unless there are 100 of participants per treatment condition. Instead, researchers should consider making a priori hypotheses regarding pairwise comparisons (eg, each active treatment condition compared to the control condition) as their primary study hypotheses. These considerations may be particularly important when recruiting large sample sizes is challenging, such as studies involving individuals with rare pain conditions, children, or minoritized individuals.

Study Limitations

This study has a number of limitations. Perhaps most importantly, it represents one of the very few studies evaluating the role that the number of pain intensity ratings plays on the reliability and validity of pain intensity outcome assessment. However, given that the participants in this study might not be representative of other populations of people with pain, the questions addressed by the current analyses should be studied in additional populations in order to help determine which findings are reliable. A second limitation is that we compared pain intensity scores made up of only a maximum of 4 days' worth of ratings. Given that we found the greatest sensitivity for those composite scores that used data from all 4 days, it is possible that sensitivity could have been increased further had we been able to include ratings from 5 or even 6 days. Future studies should include ratings from more days if possible. A third limitation is the time frame used for the recall ratings in the clinical trial that provided the data for the current analyses. Many years ago Dworkin and Siegfried[28] noted that if one were to use a large enough recall period and ask study participants to rate their average pain over that period, individual ratings of briefer periods may not be necessary. We were not able to test this possibility with the data that were available. It is therefore not known how well composite scores made up of ratings assessing pain over relatively brief periods of time (eg, 4 or more ratings of 24-hour recalled pain) might compare to single ratings assessing pain over a longer period of time (eg, a single 7-day recall rating). However, as Darnall[29] has noted, some caution about the use of recall ratings that extend for excessively large periods of time seems warranted, as the use of recall ratings can potentially add bias for very long recall periods. Finally, the focus of this article is on the ability of the measures to detect *statistically significant* treatment-related differences in pain intensity. The findings do not speak to the clinical meaningfulness of such differences.

Conclusions

Despite the study's limitations, the findings provide new information regarding the role of the number of ratings on the reliability and validity of pain intensity assessment. They suggest that—contrary to our original hypothesis—fewer than 3 days of ratings may not be enough to create adequately reliable measures of pain intensity, and fewer than 4 days of ratings may not be enough to ensure an adequately high chance of detecting real between-group differences, especially in situations where the sample size is 50 individuals or fewer per condition. Moreover, if the findings from the current study replicate in trials with other populations testing other interventions, assessing all 4 intensity domains (current and also least, worst, and average pain) on multiple occasions and combining all of these, a single global composite score, may result in the most consistently sensitive measure. Additional research testing these ideas using data from completed clinical trials is needed to evaluate the generalizability of the findings.

Disclosures

This work was partly supported by a Research Grant from the National Institutes of Health/National Institute of Child Health & Human Development/National Center for Rehabilitation Research (Grant number R01 HD070973). J.M.'s work is supported by the Government of Catalonia (AGAUR; 2021SGR-730) and ICREA-Acadèmia; in addition, the Chair in Pediatric Pain is partially funded by Fundación Grünenthal. R.V.'s work is supported by the Spanish Ministry of Science and Innovation with a Ramon y Cajal contract (RYC2018-024722-I). Funding for open access charge: Universidad de Málaga / CBUA. The authors declare no conflicts of interest.

Data Availability

Data from this article are available upon reasonable request.

References

1. Dworkin RH, Turk DC, Farrar JT, *et al.* Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 113(1-2):9-19, 2005. https://doi.org/10.1016/j.pain.2004.09.012

2. Smith SM, Fava M, Jensen MP, *et al.* John D. Loeser Award Lecture: size does matter, but it isn't everything: the challenge of modest treatment effects in chronic pain clinical trials. Pain 161(1):S3, 2020. https://doi.org/10.1097/J.PAIN.0000000000001849

3. Nunnally JC: Psychometric Theory. McGraw-Hill; 1979. https://doi.org/10.1109/PROC.1975.9792

4. Jensen MP, Hu X, Potts SL, Gould EM: Single vs composite measures of pain intensity: relative sensitivity for

detecting treatment effects. Pain 154(4):534-538, 2013. https://doi.org/10.1016/j.pain.2012.12.017

5. Jensen MP, Turner JA, Romano JM, Fisher LD: Comparative reliability and validity of chronic pain intensity measures. Pain 83(2):157-162, 1999. https://doi.org/10.1016/S0304-3959(99)00101-3

6. Stone AA, Schneider S, Broderick JE, Schwartz JE: Single-day pain assessments as clinical outcomes: not so fast. Clin J Pain 30(9):739-743, 2014. https://doi.org/10.1097/AJP.0000000000000030

7. Jensen MP, Hu X, Potts SL, Gould EM: Measuring outcomes in pain clinical trials: the importance of empirical support for measure selection. Clin J Pain 30(9):744-748, 2014. https://doi.org/10.1097/AJP.0000000000000046

8. Jensen MP, Mcfarland CA: Increasing the reliability and validity of pain intensity measurement in chronic pain patients. Pain 55:195-203, 1993.

9. Jensen MP, Mendoza ME, Ehde DM, *et al.* Effects of hypnosis, cognitive therapy, hypnotic cognitive therapy, and pain education in adults with chronic pain: a randomized clinical trial. Pain 161(10):2284-2298, 2020. https://doi.org/10.1097/j.pain.0000000000001943

10. Jensen MP, Hakimian S, Ehde DM, *et al.* Pain-related beliefs, cognitive processes, and electroencephalography band power as predictors and mediators of the effects of psychological chronic pain interventions. Pain 162(7):2036-2050, 2021. https://doi.org/10.1097/j.pain.0000000000002201

11. Jensen MP, Ehde DM, Hakimian S, Pettet MW, Day MA, Ciol MA: Who benefits the most from different psychological chronic pain treatments? An exploratory analysis of treatment moderators. J Pain 24(11):2024-2039, 2023. https://doi.org/10.1016/j.jpain.2023.06.011

12. Jensen MP, Karoly P: Self-report scales and procedures for assessing pain in adults. In: Turk DC, Melzack R, editors. Handbook of Pain Assessment. 3rd ed.Guilford Press; 2001. pp 15-34.

13. IBM Corp: IBM SPSS Statistics for Windows. IBM Corp; 2021

14. Faul F, Erdfelder E, Lang AG, Buchner AG: Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39(2):175-191, 2007. https://doi.org/10.3758/BF03193146

15. Jensen MP, Tomé-Pires C, Solé E, *et al.* Assessment of pain intensity in clinical trials: individual ratings vs composite scores. Pain Medicine 16(1):141-148, 2015. https://doi.org/10.1111/pme.12588

16. Jensen MP, Wang W, Potts SL, Gould EM: Reliability and validity of individual and composite recall pain measures in patients with cancer. Pain Medicine 13(10):1284-1291, 2012. https://doi.org/10.1111/J.1526-4637.2012.01470.X

17. Chibnall JT, Tait RC: Pain assessment in cognitively impaired and unimpaired older adults: a comparison of four scales. Pain 92(1-2):173-186, 2001. https://doi.org/10.1016/S0304-3959(00)00485-1

18. Konno SI, Alev L, Oda N, Ochiai T, Enomoto H: An open-label, 52-week, phase III trial of duloxetine in japanese patients with chronic low back pain. Pain Med 20(8):1479-1488, 2019. https://doi.org/10.1093/PM/PNZ027

19. Konno S, Oda N, Ochiai T, Alev L: Randomized, double-blind, placebo-controlled phase III trial of duloxetine monotherapy in Japanese patients with chronic low back pain. Spine 41(22):1709-1717, 2016. https://doi.org/10.1097/BRS.0000000000001707

20. Groenveld TD, Smits MLM, Knoop J, *et al.* Effect of a behavioral therapy-based virtual reality application on quality of life in chronic low back pain. Clin J Pain 39(6):278-285, 2023. https://doi.org/10.1097/AJP.0000000000001110

21. Yildirim YK, Cicek F, Uyar M: Effects of pain education program on pain intensity, pain treatment satisfaction, and barriers in Turkish cancer patients. Pain Manag Nurs 10(4):220-228, 2009. https://doi.org/10.1016/J.PMN.2007.09.004

22. Dalton JA, Keefe FJ, Carlson J, Youngblood R: Tailoring cognitive-behavioral treatment for cancer pain. Pain Manag Nurs 5(1):3-18, 2004. https://doi.org/10.1016/S1524-9042(03)00027-4

23. Lin SY, Neoh CA, Huang YT, Wang KY, Ng HF, Shi HY: Educational program for myofascial pain syndrome. J Altern Complement Med 16(6):633-640, 2010. https://doi.org/10.1089/ACM.2009.0378

24. Ross JR, Goller K, Hardy J, *et al.* Gabapentin is effective in the treatment of cancer-related neuropathic pain: a prospective, open-label study. J Palliat Med 8(6):1118-1126, 2005. https://doi.org/10.1089/JPM.2005.8.1118

25. Nygaard AS, Rydningen MB, Stedenfeldt M, *et al.* Group-based multimodal physical therapy in women with chronic pelvic pain: a randomized controlled trial. Acta Obstet Gynecol Scand 99(10):1320-1329, 2020. https://doi.org/10.1111/AOGS.13896

26. de C Williams AC, Fisher E, Hearn L, Eccleston C: Psychological therapies for the management of chronic pain (excluding headache) in adults. Cochrane Database Syst Rev 2020(8):CD007407, 2020. https://doi.org/10.1002/14651858.CD007407.PUB4/MEDIA/CDSR/CD007407/IMAGE_N/NCD007407-CMP-008.06.SVG

27. Fisher E, Villanueva G, Henschke N, *et al.* Efficacy and safety of pharmacological, physical, and psychological interventions for the management of chronic pain in children: a WHO systematic review and meta-analysis. Pain 163(1):E1-E19, 2022. https://doi.org/10.1097/J.PAIN.0000000000002297

28. Dworkin RH, Siegfried RN: Are all those pain ratings necessary? Pain 58(2):279, 1994. https://doi.org/10.1016/0304-3959(94)90213-5

29. Darnall BD: On the importance of baseline pain intensity and measurement methods. Pain 164(9):1887-1888, 2023. https://doi.org/10.1097/j.pain.0000000000002931