Extraction of RDC/TMD Subscales from the Symptom Check List-90: Does Context Alter Respondent Behavior?

Richard Ohrbach, DDS, PhD
Associate Professor
Department of Oral Diagnostic Sciences
University at Buffalo
Buffalo, New York

Jeffrey Sherman, PhD
Assistant Professor
Department of Oral Medicine
University of Washington
Seattle, Washington

Carla Beneduce, RDH
Research Assistant
University at Buffalo
Buffalo, New York

Kimberly Zittel-Palamara, MSW, PhD
Research Assistant
University at Buffalo
Buffalo, New York

Youngju Pak, PhD
Graduate Student
Department of Biostatistics
University at Buffalo
Buffalo, New York

Correspondence to:
Dr Richard Ohrbach
Department of Oral Diagnostic Sciences
University at Buffalo
355 Squire
Buffalo, NY 14214
Fax: +716 829 3554
E-mail: ohrbach@buffalo.edu

Aims: *To test whether extraction of the 2 subscales in the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD) affected the subscale score reliability and whether scores from the RDC/TMD subscales are comparable to the same scales when the whole Symptom Check List-90 (SCL-90R) is administered.* ***Methods:*** *The full SCL90-R and a modified version containing only the depression and somatization scales were administered in counterbalanced order to 103 subjects. As another test of context, a subset of participants completed the modified and full versions as part of a larger battery of instruments relevant to facial pain. Statistical analyses included internal reliability for item analysis and intraclass correlation (ICC) and Lin's concordance correlation coefficient (CCC) for total scale score reliability.* ***Results:*** *Internal reliability was approximately 0.95 for depression and 0.87 for somatization, independent of test form. Total scale scores were reliable across test versions, with both ICC and CCC approximately 0.95 for depression and 0.91 for somatization. Permutation tests using the CCC indicated a mild influence on the somatization score but not the depression score due to order effects, but these effects were not significant when considering the 95% CIs based on resampling methods.* ***Conclusion:*** *Whether items from other subscales are present or not does not affect the internal reliability or parallel forms reliability of the total scores from either depression or somatization. Context of administration, via order of forms completion, does not alter total score or reliability of depressive items but may alter total scores for somatization.* J OROFAC PAIN 2008;22:331–339

Key words: psychometrics, RDC/TMD, reliability, validity

Psychological self-report instruments are used extensively in medical and dental research. Increasingly complex study designs often place high burdens on subject participation, and one method to reduce such burdens is to tailor the self-report assessments by extracting selected subscales from a validated parent instrument. Such a strategy was used in developing the Research Diagnostic Criteria for Temporomandibular Disorders (RDC/TMD). The RDC/TMD, which has been translated into more than 20 languages, is the most widely used research tool in the clinical setting for the diagnosis of TMD and for the assessment of psychosocial distress in a TMD population.[1] TMD are a group of musculoskeletal pain conditions associated with the muscles of mastication and/or the temporomandibular joint,[2] and they affect, for example, up to 18% of the US population.[3]

As with all chronic pain conditions, psychosocial distress and mental illness are quite common in TMD clinic populations.[4–8] The RDC/TMD uses a dual-axis diagnostic and classification system that includes 1 axis to record clinical physical findings and a second axis to record behavioral, psychological, and psychosocial status. The physical axis provides clinical researchers with a standardized system that can be evaluated for its use in examining, diagnosing, and classifying the most commonly appearing subtypes of temporomandibular disorders. The biobehavioral axis (Axis II) was intentionally designed as a brief screening tool to assess for 2 pain-relevant psychological constructs (depression and somatization) via 32 items extracted from the Symptom Check List 90 (SCL-90)[9] and to evaluate pain-related interference via the Graded Chronic Pain Index.[10] Depression and somatization, as 2 scales composing the SCL-90, were specifically selected for screening in this pain population because of the very strong theoretical and empirically demonstrated relationship of those constructs to the experience and progression of chronic pain. Importantly, the extraction of the 2 subscales from the SCL-90 for the RDC/TMD was accompanied by the administration of the isolated subscales in a random population design.[11] Independent norms were developed for the 2 subscales, which were shown to result in solid psychometric values in their extracted form. Consequently, the use of extracted subscales in the context of the RDC/TMD is psychometrically justified, but the larger question is whether the scale values and their interpretation can be generalized to information as obtained by the parent instrument (the SCL-90 or SCL-90R).

According to the psychometric literature,[12,13] subscale extraction for independent application has been considered inappropriate. Moreover, formal guidelines for instrument development explicitly indicate that if a subscale is extracted, absence of alteration in the scores needs to be demonstrated.[14] However, while the extraction of selected *items* from a parent instrument is associated with 2 established "sins"—the assumption that the reliability and validity of the parent items automatically apply to the extracted items and the belief that less validity evidence is consequently needed[13]—the claim of potential problems associated with the extraction of entire *subscales* is accompanied by little to no evidence. While the psychometric canon includes words of caution against extraction of subscales from parent instruments, assumptions within both classical test theory as well as item response theory regarding item invariance would suggest that the performance of individual items would remain consistent regardless of context.[15]

In a review of the literature on short-form use and methodology, Smith and colleagues suggested that the literature has been characterized by an overly optimistic view that validity will transfer from the parent form to the isolated subscales.[13] They suggested that methodologic and psychometric principles be equally applied to the isolated subscales to develop valid clinical assessment tools. Within the RDC/TMD, the depression and somatization scales have been normed on large samples[1] and used in a wide range of studies in the United States and internationally[16–20]; they have also demonstrated reliability and validity with other similar measures.[11] However, the relationship of these 2 scales to the original scales—ie, as used in the RDC/TMD versus as measured in the SCL-90—has never been assessed. This question has direct implications for the equivalence of the short version containing 32 items as used in the RDC/TMD and the long form of the SCL-90. It also has more general implications for other instruments in widespread use in the same manner. The aim of the present study was to test whether extraction of the 2 subscales in the RDC/TMD affected the subscale score reliability and whether scores from the RDC/TMD subscales are comparable to the same scales when the whole SCL-90R is administered.

Materials and Methods

Subjects

Subjects between 18 and 65 years of age who were proficient in the English language usage were recruited sequentially. The participants were either patients at a private facial pain practice (n = 51) or dental school patients identified as having special social and/or financial needs (n = 52). The latter group was known to exhibit life stress that interfered with their ability to fully participate in their dental treatment, and among that group, subjects were either patients in a specialty pain teaching clinic (n = 15) or were general dental school patients (n = 37). The indicated populations were sampled to minimize the number of low responses across the 2 testing situations, since that would upwardly bias any agreement. Among the subjects approached for the study, there were 2 refusals. One hundred fourteen subjects entered the study; 11 subjects returned incomplete data. The final sample of 103 subjects comprised 27 men (mean age 44.7, SD 12.3) and 76 women (mean age 41.5, SD 12.3). The preponderance of women was consistent with the gender distribution in each recruitment source. Subjects received monetary compensation of $10 for completing the study. The study was approved by an institutional review board, and informed consent was obtained from each subject.

Procedures

The 32 RDC/TMD items assessing depression and somatization as extracted from the original SCL-90 comprised the "modified form" of the targeted instrument. The full SCL-90 comprised the "full form." Item sequence in the modified form was as published in the original RDC/TMD, which followed, for the most part, the ordering of the corresponding items in the SCL-90; the item sequence in the full form was exactly as published in the SCL-90. In using "depression" and "somatization" as the 2 target constructs, 2 assessment domains were addressed in the present study: a set of items that relates to more psychological states (depression symptoms, per the *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition [DSM-IV]) and a set of items that relates to more physical symptom states (specifically, nonfunctional symptoms) such as would easily be found in a medical symptom checklist.

Both forms were administered to each subject, and the order of which form was administered first was counterbalanced as subjects entered the study. The counterbalancing of the modified and full forms of the target instrument created a context effect, in that those subjects randomly assigned to complete the full form first carried context effects of the other constructs that comprise the SCL-90 to the modified form, which was completed second. The subjects assigned to the reverse sequence carried to the second form administration a potentially stronger bias of depression and somatization.

A second factor was presence versus absence of other instruments (eg, disability, pain symptoms, limitation, stress experience items, demographics) administered at the same time as the target instruments. Because this study was conducted in a clinical setting, some subjects were recruited separate from their clinical process, whereas others were recruited as part of their standard clinical evaluation. While assignment to counterbalanced order was random per entry into the study, subjects who completed other forms with the target instruments did so on a quasi-experimental basis.[21]

Subjects from the respective settings were recruited, and the first study instrument was administered either in person or at home (depending on clinic flow procedures); the second form was always completed at home. All forms were self-administered. Forms from home were mailed in, and postal date was compared against declared date of completion in order to verify time interval between administrations. The requested time separation between administrations was 2 days, with an allowable window of 1 to 9 days. The ideal time interval for assessing parallel forms of an instrument is within the same day; however, endorsement by recall from the prior administration would potentially confound the study goals, while a long interval between administrations could result in low agreement due to the person's mood or bodily symptom states changing. Two days was selected as the target interval based on clinical experience that the constructs under examination do not typically change over that short interval.

Scoring

Consistent with recommended practice for the SCL-90,[22] a summary score was created for each scale within each instrument by computing the simple sum of the endorsed ordinal rank for each item. Following published scoring rules for the depression and somatization scales within the RDC/TMD assessment protocol, the depression score was based on 20 items (derived from the 13 original SCL-90 items for depression and the 7 items for vegetative symptoms) and the somatization score was based on 12 items. Scoring of the respective constructs in the long form was computed based on the same items. Per RDC/TMD guidelines, the sums were adjusted for missing values, as long as there were at least two-thirds valid responses within a construct. Ninety-four percent of the responses for the depression scales were complete (6% of the subjects had up to 2 missing items) and 96% of the somatization scale responses were complete (4% had 1 missing item). The summed score was then rescored on a 0 to 4 metric based on the number of items with valid responses.

Statistical Analyses

Internal reliability, via Cronbach's α, was computed as an index of individual item performance. Raw scores for the 2 versions (modified, full) of each scale were compared using percent difference,

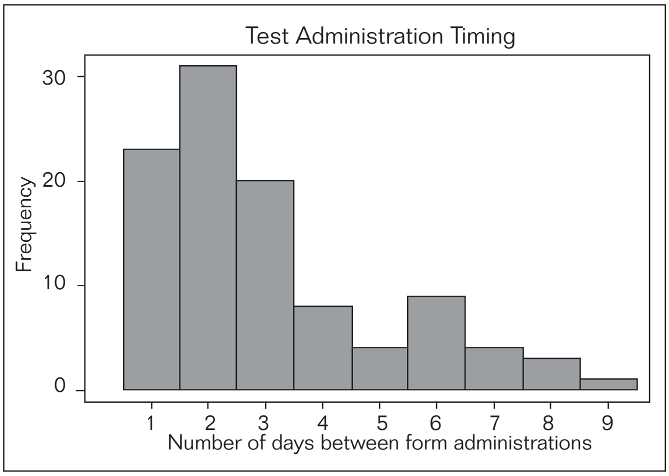

Fig 1 Histogram of number of days between self-administration of first study form and second study form. See Fig 2 for assessment of implications of this difference.

per the other 2 factors (counterbalanced order; isolation of instrument administration), to provide a descriptive summary of how the scales performed. In keeping with general methods of presenting test reliability, the modified form and the full form were compared using several approaches: Pearson correlation, intraclass correlation (ICC; fixed raters),[23] and an ICC computed from a full factorial model which included coding for the other 2 experimental factors present in the data collection methods. However, while the ICC is an often used statistic for instrument reliability, it is not without problems or critique.[24–28] The most notable problem is the large impact that sampling bias has on restriction of range, leading to a biased and underestimated statistic. Lin developed the concordance correlation coefficient (CCC) in order to resolve these criticisms,[29] and consequently the CCC is used for the primary statistic of reliability between the modified version and the full version for each construct. Note that a CCC value of 1.0 denotes perfect agreement, and a value of 0.0 denotes no agreement.

To evaluate the extent that the modified and full forms of depression and somatization agree, the CCC was computed for different subsets of the data. Bootstrap resampling methods were used to obtain 95% confidence intervals. Specifically, resampling with replacement of the data was simultaneously done from each group; for each resample, the CCC was calculated. This procedure was repeated 10,000 times. A confidence interval for the 2.5th and 97.5th percentiles from this simulated distribution was then obtained. In order to statistically evaluate differences observed between CCC values corresponding to subsets of interest, permutation testing methods were used. The data were permuted, ignoring subset assignment, after which the difference in CCC between the 2 groups was calculated. This was done 10,000 times to obtain the required null distribution from which the 2-sided *P* value was obtained.

In addition to the CCC, which incorporates magnitude of values, equivalence of summary scores was also assessed with factorial analysis of variance (ANOVA) using partial sums of squares to assess the factors of counterbalanced order and isolation of instrument administration on difference scores derived from the full form and modified form. In addition, as a secondary analysis to assess how the presence of other instruments affects respondent behavior, another set of ANOVAs was computed, testing the effects of these factors on each of the available raw scores from the target instruments. From the associated factorial cell means, differences between contrasts of interest were computed and converted to percentages to estimate any practical impact based on the various factors implemented in this study. Stata 8.0 and SAS v9 were used for statistical analyses.

Results

While a 2-day interval between administrations was deemed optimal, subjects completed the second instrument from 1 to 9 days after the first (Fig 1). Fifty-four percent of the sample completed the second form within 2 days of the first form. To assess whether a very short interval versus a long interval between time-1 and time-2 administrations had any impact on responding to the second instrument, the difference scores were plotted of the summed score responses from time-2 to time-1 as a dot-plot in order to determine whether there was a trend over the time interval between administrations.[25] As the dot-plots in Fig 2 demonstrate, the short interval of 1 to 2 days resulted in the same absence of effect as did the longer periods for both depression and somatization subscales. The data exhibited descriptive values sufficient for the present study: the mean value of depression was 1.1 (SD 0.91; min 0, max 3.5), the mean of somatization was 1.04 (SD 0.73; min 0.05, max 3.05), and the s mean scores for each measure did not differ across the subject groups based on recruitment source ($P > .14$).

Internal reliability, via Cronbach's α, was computed for only nonmissing data (Table 1). Internal

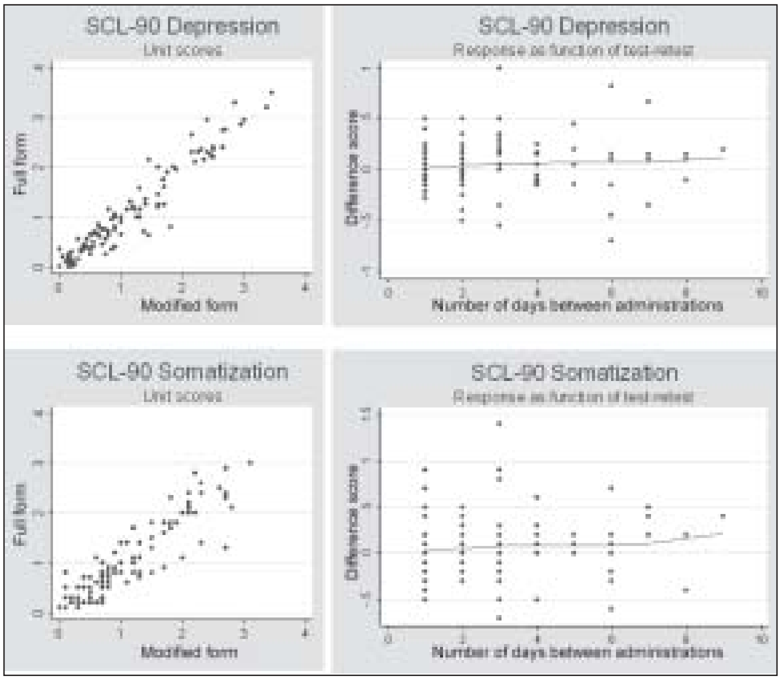

Fig 2 Scatter plots and dot-plots of raw data. Scatter plots display total scale score (adjusted for missing data) for each of modified and full form versions of depression and of somatization. Dot-plots display the difference between the score obtained from instrument administered at time-2 and the score from time-1 administration, according to the number of days between administrations. The Lowess regression line demonstrates absence of appreciable effect over time upon the difference in total score.

Table 1 Internal Reliability and Descriptive Statistics

	Overall α	Mean	SD	Range
Depression				
Full	0.953	1.08	0.90	0–3.5
Modified	0.948	1.13	0.87	0–3.5
Somatization				
Full	0.868	1.00	0.74	0.1–3.0
Modified	0.868	1.08	0.76	0–3.1

Cronbach's α estimates for each construct, according to whether full instrument (full) or subset of items (modified). The mean scores, comparing full versus modified, did not differ (paired *t* test, $P > .05$) within depression or somatization.

Table 2 Summary of Reliability Statistics

	Depression	Somatization
Pearson correlation	0.961	0.908
Simple ICC* (fixed raters)	0.960	0.907
Complex ICC† (fixed raters)	0.959	0.905

Comparison of total scale scores from modified versus full forms of the respective SCL-90 subscales.

*Based on 1-way ANOVA, comparing modified to full instrument total scores.

†Based on full ANOVA factorial model: modified versus full, form sequence from counter-balancing, and whether administered with other instruments

reliability was excellent for both constructs of depression and somatization, with that of depression slightly higher than that of somatization. There was no difference in overall internal reliability according to whether the modified or full form version was used for either the depression or somatization subscales.

Standard reliability statistics are shown in Table 2, indicating highly comparable responses between Pearson and ICC statistics; given the adherence of the raw data to the line of unity as shown in the scatter plots in Fig 2, the equivalence of the statistics is not surprising. While the items for depression exhibit a higher level of reliability between modified and full versions of the instrument, the reliability statistics for both depression and somatization are excellent.

Table 3 Estimated CCC for Each Group

Pair of comparing variable	Isolques	Formseq	Sample size	Estimated CCC	95% Bootstrap CI of CCC (simulation = 10,000)
DEP	Overall		97	.958	[.932, .975]
	0	.	24	.959	[.867, .984]
	1	.	73	.956	[.931, .972]
	.	1	47	.939	[.886, .969]
	.	2	50	.978	[.956, .990]
	0	1	12	.926	[.520, .974]
	0	2	12	.987	[.849, .974]
	1	1	35	.940	[.872, .974]
	1	2	38	.974	[.943, .990]
SOM	Overall		97	.894	[.837, .928]
	0	.	24	.696	[.458, .816]
	1	.	73	.920	[.875, .950]
	.	1	47	.777	[.648, .871]
	.	2	50	.955	[.875, .950]
	0	1	12	.597	[.189, .770]
	0	2	12	.797	[.423, .903]
	1	1	35	.812	[.672, .922]
	1	2	38	.970	[.947, .984]

Isolques 0 = no, and 1 = yes for whether the form was administered in isolation of other instruments (cf., with other instruments). Formseq 1 = modified full, while Formseq 2 = full modified. Sample size for total sample in this analysis (n = 97) differs from the n = 103; see text for explanation. DEP = depression; SOM = somatization.

Table 4 Estimated Difference in CCC Between Different Groups

Pair of comparing variable	Comparing	Sample size	Estimated CCC (1) to CCC (2)	*P*
DEP	Isolques = 0	N1 = 24	.00319	.985
	Isolques = 1	N2 = 73		
	Formseq = 1	N1 = 47	–.0387	.791
	Formseq = 2	N2 = 50		
SOM	Isolques = 0	N1 = 24	–.224	.118
	Isolques = 1	N2 = 73		
	Formseq = 1	N1 = 47	–.178	.152
	Formseq = 2	N2 = 50		

See Table 3 for explanation of codes.
DEP = depression; SOM = somatization.

The CCC and 95% confidence interval for each construct, according to each of the different experimental factors in this study, are shown in Table 3; for these analyses 6 subjects were dropped because of missing data related to whether they completed study forms with other forms or not. The experimental factors of counterbalancing and modified versus full instruments made little difference in the reliability of scores for either depression or somatization; in contrast, whether one of the target instruments (modified or full) was completed in isolation of other instruments or not resulted in a substantial shift in the CCC for somatization but not depression. In order to assess whether that shift in CCC reliability was significant, the difference between the respective CCC statistics was computed and tested against a distribution created via permutation tests. As shown in Table 4, none of these differences were significant, demonstrating that neither of the 2 observed experimental factors (whether modified versus full instrument was administered first; completion of the instrument in isolation or with other instruments) had any appreciable effect on the scale reliability.

Scale scores for each construct are shown in Table 1. Factorial ANOVA for the difference score between full form and modified form for depression revealed no significant main or interaction effects ($P > .14$); the same was true for the somatization instruments ($P > .17$). These results indicate that scale scores for the full form and the modified form were not different from one another for depression or somatization within the 2 factors assessed in this study. In contrast, administration of other instruments simultaneous with the target instrument significantly decreased scale scores, regardless of forms sequence, for the full form (*P* = .045) and marginally for the modified form (*P* = .053) for depression; there was no impact by administration of other instruments on either the full form score (*P* = .35) or the modified form score (*P* = .35) of somatization. Inspection of raw means, partitioned by the 2 study factors, disclosed that individuals consistently exhibited a pattern of endorsing a lower level of depression symptoms when the target instrument was administered with other instruments versus when administered alone. Importantly, this was equally true for each of the full and modified test forms. When the mean total values were re-expressed as a percentage difference of the factorial mean values, the simple difference in scale scores between the full instrument and the modified instrument for each of depression and somatization was 5% and 7%, respectively; in contrast, the presence of other instruments resulted in differences of up to 35% in the scale scores for depression instruments.

Discussion

This study was undertaken in order to answer 2 questions. The more specific and answerable question is whether the 2 subscales of depression and somatization, as used in the RDC/TMD, can be extracted from the SCL-90 and retain their validity and reliability with respect to retaining the same interpretation of the construct. The importance of this specific question lies in the daily clinical use of these items worldwide in TMD assessment. The second question is one of psychometric principle, and asks if the general caution against subscale extraction is indeed necessarily warranted. From a validity perspective, these data indicate that item extraction from the parent instrument was successful. Scores were comparable regardless whether the full instrument or the subset of scales was used, and the resulting instrument is shorter and more appropriately tailored for the specific research task. This finding, by itself, suggests that the use of the SCL-90 subscales of depression and somatization in the RDC/TMD protocol is indeed valid not only in terms of the separate validity data published for the RDC/TMD protocol, but now also for comparability of scores obtained using the RDC/TMD protocol to scores obtained in settings where the full SCL-90 is used.

Counterbalancing in this study resulted in 100% of the subjects completing the same items again, with 50% of the subjects doing so a second time with only the 2 scales comprising the modified form, and with 50% of the subjects doing so a second time with the other 58 items from the SCL-90 interwoven into the study's primary target items. Counterbalancing thus also forms an experimental variable of tightly controlled context, and this more local context did not appreciably alter subject responses. Similarly, reduction in the number of items for the Center for Epidemiologic Studies–Depression (CES-D), a tool comparable to the SCL-90 for assessing depression,[11] did not result in any appreciable alteration in its core psychometric properties.[30] In contrast, studies examining serial order of items report consistent changes in response patterning due to changing item order (as would occur in the present study in the full instrument form); however, the studies examining serial order effects have been limited to personality assessment.[31] Overall, these findings suggest that changes in internal instrument structure, depending on the underlying construct, can occur without affecting scoring properties.

A different type of context effect occurred in this study when other instruments were administered at the same time as the target instruments. In the assessment of personality with self-report instruments, it is hypothesized that a self-reflective focus[32] and consequent engagement of the self[33] is created by the context of the instrument, and that results in the individual endorsing more rather than fewer characteristics about themselves due to better access to memory. The present data suggest the opposite pattern for depression symptoms, at least in a dental setting, in that other instruments administered with the target instrument resulted in the individuals reporting less symptomatology compared to when they completed the test instrument alone. It is tempting to interpret this as due to time effects: a longer instrument results in the individual perhaps allocating less time pondering individual items and hence underreporting relative to when they complete the target instrument alone. However, that the somatization data did not yield such a pattern suggests that at least 1 other factor may be operating. Since the bulk of the other items comprising the other instruments was focused on pain and functioning, perhaps the individuals reframed their depressive symptoms as part of the pain disorder (hence, underreporting the depression), and the somatization symptoms (ie, physical body symptoms) reporting did not decline because they are congruent with the pain disorder. Other evidence suggests that the content of the early items within an instrument appears to clarify the meaning of later items in the instrument, with the consequence of improved overall reliability.[33] While these are clearly important considerations for future research addressing self-report–based information, for the present study the important conclusion is that these other influences affected not only the modified form but also, to the same extent, the full form.

Overall, the present findings suggest that the general caution against use of extracted subscales must be considered in context: different kinds of items can be expected to either be sensitive to, or not sensitive to, context effects, and while that context may be embedded within the administered target items, that context is certainly also created by the proximity of other instruments in addition to the items comprising a parent instrument. Finally, that context affects items differentially. In the case of the present study, items assessing bodily symptoms were more sensitive to the items in other instruments assessing pain-relevant symptoms, while pain-relevant mood items (ie, depression) were not sensitive to that context.

The context of the present study itself should be considered as yet 1 more layer in this investigation. Over 50% of the subjects were currently being either evaluated or treated in a dental (not psychological) facility for a chronic pain condition, and the remainder of the subjects, while being evaluated by a social work unit, were available for recruitment in this setting due to their primary complaint of dental problems. Hence, it would be expected that this subject sample had a high likelihood of substantial priming of physical (versus mental health) symptoms and beliefs.

The present study has several limitations. Foremost is that the study used only 2 subscales from 1 multidimensional instrument; this hardly answers the larger question of whether any subscale from all multidimensional instruments can be extracted. Indeed, if extracted from a parent instrument, subscales should be validated independently without presumption that the psychometric properties of the parent carry over to the extracted subscales.[13] A second limitation is that the administration of other instruments was performed in only a subset of subjects and although that sample was adequate in size for the bootstrap resampling procedure, potential bias in sampling of those subjects (since they came from a different clinical population) cannot be ruled out. A third limitation is that this study only examined the impact of other self-report instruments and items on performance of subscales. Future research in the clinical setting should also examine the impact of clinical examinations on performance of instruments. Finally, this study was conducted in a dental school, among patients who, despite an approximate 50% rate of diagnosable mental disorders, were nevertheless seeking somatic help for somatically oriented problems (at least, based on the nature of the chief complaints among these 2 populations). Subjects recruited from other areas may behave differently with respect to the tension between constructs such as depression versus somatization.

As measured with SCL-90 items using 2 questionnaire versions, depression was noted to exhibit extremely robust reliability and to not be influenced by adjacent items or other instruments. Somatization exhibited very good reliability, but was also more influenced by context of immediately adjacent items and other instruments. Consequently, extraction of subscales from other instruments should be empirically validated for retaining the expected reliability. In sum, current wisdom says that subscales cannot be extracted, yet we turn a “blind eye” to other context influences potentially affecting response behaviors. Because of increasing pressures to administer shorter tests (either by reducing the number of items, or by administering only the necessary subscales of a multidimensional instrument), the question of whether subscales can be extracted from a parent instrument is even more relevant clinically in current settings in addition to its importance for psychometric theory. These data suggest that extraction and administration of a subscale need not necessarily compromise the reliability and validity of that scale.

Acknowledgments

The authors thank Dr Yoly Gonzalez, Dr Lance Ortman, and the School of Dental Medicine staff at the University at Buffalo for their help during the recruitment phase. This research was presented at the International Association for Dental Research, Göteborg, Sweden, July 2003. Research supported by NIH-DE-13331 from NIDCR/NIH, Bethesda, MD, US. The RDC/TMD Validation Study Group comprises:

- University of Minnesota: Eric Schiffman (Study Primary Investigator), Mansur Ahmad, Gary Anderson, Quintin Anderson, Pat Carlson, Mary Haugan, Amanda Jackson, John Look, Wei Pan
- University at Buffalo: Richard Ohrbach (Study Primary Investigator), Leslie Garfinkel, Yoly Gonzalez, Krishnan Kartha, Sharon Michalovic, Betsy Seyler-Piccolo, and Theresa Speers
- University of Washington: Edmond Truelove (Study Primary Investigator), Lars Hollender, Kimberly Huggins, Kathy Scott, and Earl Sommers

References

1. Dworkin SF, LeResche L. Research Diagnostic Criteria for Temporomandibular Disorders: Review, criteria, examinations and specifications, critique. J Craniomandib Disord 1992;6:301–355.
2. McNeill C, Mohl ND, Rugh JD, et al. Temporomandibular disorders: Diagnosis, management, education, and research. J Am Dent Assoc 1990;120:253–263.
3. Von Korff M, Dworkin SF, Le Resche L, Kruger A. An epidemiologic comparison of pain complaints. Pain 1988; 32:173–183.
4. Kight M, Gatchel RJ, Wesley L. Temporomandibular disorders: Evidence for significant overlap with psychopathology. Health Psychol 1999;18:177–182.
5. Ohrbach R, LeResche L, Dworkin SF. Longitudinal changes in TMD: Influence of baseline findings and treatment-seeking. J Dent Res 1997;76:389.
6. Dworkin SF. Behavioral, emotional, and social aspects of orofacial pain. In: Stohler CS, Carlson DS (eds). Biological and Psychological Aspects of Orofacial Pain. Ann Arbor, MI: Center for Human Growth and Development, 1994: 93–112.
7. Wilson L, Dworkin SF, Whitney C, LeResche L. Somatization and pain dispersion in chronic temporomandibular disorder pain. Pain 1994;57:55–61.
8. Turner JA, Dworkin SF. Screening for psychosocial risk factors in patients with chronic orofacial pain: Recent advances. J Am Dent Assoc 2004;135:1119–1125.
9. Derogatis LR, Lipman RS, Covi L. SCL-90: An outpatient psychiatric rating scale—Preliminary report. Psychopharmacology 1973;9:13–28.
10. Von Korff M, Ormel J, Keefe FJ, et al. Grading the severity of chronic pain. Pain 1992;50:133–149.
11. Dworkin SF, Sherman J, Mancl L, Ohrbach R, LeResche L, Truelove E. Reliability, validity, and clinical utility of RDC/TMD Axis II scales: Depression, non-specific physical symptoms, and graded chronic pain. J Orofac Pain 2002;16:207–220.
12. Anastasi A. Psychological Testing. New York: Macmillan, 1988.
13. Smith GT, McCarthy DM, Anderson KG. On the sins of short-form development. Psychol Assess 2000;12: 102–111.
14. American Educational Research Association. Standards for Educational and Psychological Testing. Washington, DC: Author, 1999.
15. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum, 2000.
16. List T, Dworkin SF. Comparing TMD diagnoses and clinical findings at Swedish and U.S. TMD centers using Research Diagnostic Criteria for Temporomandibular Disorders. J Orofac Pain 1996;10:240–253.
17. Lobbezoo F, van Selms MKA, John MT, et al. Use of the Research Diagnostic Criteria for Temporomandibular Disorders for multinational research. Translation efforts and reliability assessments in The Netherlands. J Orofac Pain 2004;19:301–308.
18. Wahlund K, List T, Dworkin SF. Temporomandibular disorders in children and adolescents: Reliability of a questionnaire, clinical examination, and diagnosis. J Orofac Pain 1998;12:42–52.
19. Yap AU, Dworkin SF, Chua EK, List T, Tan KB, Tan HH. Prevalence of temporomandibular disorder subtypes, psychologic distress, and psychosocial dysfunction in Asian patients. J Orofac Pain 2003;17:21–28.
20. John MT, Hirsch C, Reiber T, Dworkin S. Translating the Research Diagnostic Criteria for Temporomandibular Disorders into German: Evaluation of content and process. J Orofac Pain 2006;20:43–52.
21. Campbell DT, Stanley JC. Experimental and Quasi-Experimental Designs for Research. Boston: Houghton Mifflin, 1963.
22. Derogatis LR. SCL-90-R: Administration, Scoring and Procedures Manual-II, for the Revised Version. Towson, MD: Clinical Psychometric Research, 1983.
23. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:420–428.
24. Bartko JJ, Carpenter WT Jr. On the methods and theory of reliability. J Nerv Ment Dis 1976;163:307–317.
25. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–310.
26. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med 1990;20:337–340.
27. Chinn S. The assessment of methods of measurement. Stat Med 1990;9:351–362.
28. Ludbrook J. Comparing methods of measurement. Clin Exp Pharmacol Physiol 1997;24:193–203.
29. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–268.
30. Cole JC, Rabin AS, Smith TL, Kaufman AS. Development validation of a Rasch-derived CES-D short form. Psychol Assess 2004;16:360–372.
31. Steinberg L. Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. J Person Soc Psychol 1994;66: 341–349.
32. Hamilton JC, Shuminsky TR. Self-awareness mediates the relationship between serial position and item reliability. J Person Soc Psychol 1990;59:1301–1307.
33. Knowles ES, Byers B. Reliability shifts in measurement reactivity: Driven by content engagement or self-engagement? J Person Soc Psychol 1996;70:1080–1090.