A Critical Analysis of Multidimensional Pain Instruments for Researchers

[redacted]

University of Massachusetts Boston

Abstract

This paper seeks to discuss issues in validity and reliability for pain measurement, evaluate the quality of five multidimensional pain assessment tools, and provide recommendations for optimal use of these tools. There are certain validity and reliability issues inherent in the measurement of pain including: a lack of general theory of pain or gold standard, difficulty in distinguishing termporal variations in pain versus measurement error, cross-cultural differences in the experience of pain, and no standardized minimal clinically meaningful differences to evaluate therapeutic efficacy. Researchers should consider the underlying theory on which the instrument is based and its relationship to the research question; the instrument's features and psychometric properties; and cultural characteristics of the sample population before selecting an appropriate multidimensional pain instrument.

A Critical Analysis of Multidimensional Pain Instruments for Researchers

One hundred million American adults are affected by chronic pain totaling more than those affected by heart disease, cancer, and diabetes combined (Institute of Medicine of the National Academies Report (IOM), 2011). The United States also loses $635 billion each year in medical treatment and lost productivity attributed to chronic pain (IOM, 2011). Despite the prevalence and burden of chronic pain, controversy still exists regarding the appropriate criteria for pain assessment. Developing a reliable and valid pain measurement tool is complicated by the subjective nature of pain, and its affect on multiple domains of health, including physical, psychological, and social functioning. Self-report, as the sole method for measuring pain, is affected by individual variations in personality, cultural, linguistic and situational variables that could bias communication between client and clinician (Cleeland & Ryan, 1994). Although pain is widely accepted as a multidimensional experience, there is no consensus on which dimensions of pain should be prioritized in either prevalence or intervention studies. A comprehensive, though not exhaustive list of pain dimensions, includes: pain intensity, location, quality, onset/duration, functional ability, psychological and social functioning (Salaffi et al., 2011).

Unidimensional scales that measure pain intensity are the most commonly used instruments because they can be administered rapidly and with minimal effort (Salaffi, Sarzi-Puttini, Ciapetti, & Atzeni, 2011). Using intensity to approximate the experience of pain fails to capture the variability of responses to pain, and its effect on the different domains. Measuring any single dimension is not representative of the overall experience of pain, and selecting the appropriate multidimensional instrument for research purposes requires thoughtful consideration of its inherent features, psychometric properties, and relevance to the research question. The main objective for this paper is to discuss the complexities of selecting an appropriate

multidimensional pain instrument for research purposes. The secondary objectives are to discuss issues in validity and reliability in pain measurement, evaluate the quality of five multidimensional pain instruments, and provide recommendations for optimal use of these tools.

King's Conceptual System's Model

The nursing conceptual model developed by Imogene King (1981) allows integration of context, person, and interaction within a cogent framework. Her conceptual model is based on the premise that human beings are open systems interacting with the environment, specifically personal systems, interpersonal systems, and social systems. These three interacting systems define the physical and social environments in which human beings function (King, 1981). King's Conceptual System's Model is congruent with the idea that pain affects multiple dimensions of one's life over time. The following concepts from King's model are relevant in this paper: perception, self and time (personal system); interaction, communication, role, and stress (interpersonal system); and family systems and work systems (social system). For example, pain is measured through self-report which is subject to oneself, one's perception, and communication/interaction with a clinician. Several dimensions of pain relate to time including duration, onset, and frequency. Finally, because pain is a multidimensional stressor, it can affect one's ability to function in their social/work role, and interactions with family members.

Validity and Reliability Issues in Pain Measurement

There are certain validity and reliability issues inherent in the measurement of pain that complicate the evaluation of pain instruments (Gelinas et al., 2008). For example, there is no unified theoretical framework that defines pain and its dimensions . Therefore, the construct and content validity of any pain assessment tool is based on the author's ascribed theory of pain. For example the McGill Pain Questionnaire (MPQ) is based on the gate control theory of pain

(Melzack & Torgerson, 1971); while the Brief Pain Inventory (BPI) reflects the cognitive behavioral perspective of pain (Turk & Genest, 1987). Choosing an appropriate instrument requires that the research question be congruent with the instrument's underlying theory of pain. Criterion-related validity is also problematic in that there is no gold-standard for pain measurement (Gelinas et al., 2008). Concurrent, convergent, and discrminant validity is more appropriate because other validated instruments that measure similar or dissimilar constructs could be used for comparison.

Beyond validity, test-retest reliability is also complicated because pain exhibits temporal variation; and real changes in pain, as opposed to changes indicating random error, will lower the observed correlation between reliability samples over brief periods of time (Daut, Cleeland, & Flanery, 1983). When selecting a measure of chronic pain intensity, instead of current pain, average pain may be a more stable measure over time to better distinguish between real changes in pain from error in measurement (Daut et al., 1983). Reliability may differ among populations, as well (Jensen, 2003). Issues of cross-cultural relevance and equivalence for intended constructs are especially important in multidimensional measures. Population-specific validation studies are an ideal way to assure high quality measurement, but not always available. Responsiveness is often reported as a statistical significant difference in pain pre and post intervention but that may not represent a clinically meaningful result (Kumar, 2011). Thus, validity and reliability measures are specifically affected by the nature of pain, not just by the quality of the instrument. These issues complicate the evaluations of multidimensional instruments of pain.

Methods

A literature review was conducted to identify multidimensional pain instruments that met the following inclusion criteria: (a) instruments designed to measure more than 3 dimensions of

pain; (b) self-report instruments; (c) disease, condition, location neutral instruments; (d) targeted at adults >18yrs; (e) validated for chronic non-malignant pain; (f) had at least 2 validation studies published beyond the initial development publication. The exclusion criteria included: (a) instruments that measure acute pain, postoperative pain, or location-specific pain; (b) observational pain scales for non-verbal conditions; (c) subscales within other instruments that measure pain; (d) instruments that require both self-report and objective measures assessed by clinicians. The following databases were searched: CINAHL, MEDLINE, EBSCO, and OVID.

Of 54 identified pain instruments, five met the inclusion criteria: McGill Pain Questionnaire (MPQ) (Melzack, 1975), Short form McGill Pain Questionnaire (SF-MPQ) (Melzack, 1987), Brief Pain Inventory (BPI) (Daut et al., 1983; Cleeland & Ryan, 1994), Chronic Pain Grade Scale (CPGS) (Von Korff, Dworkin, & Le Resche, 1990) and the Multidimensional Pain Inventory (MPI) (Kerns, Turk, & Rudy, 1985). The features of each instrument were briefly summarized (Table 1); initial development publications along with additional validation studies for each instrument were evaluated (Tables 2, 3).

Quality Evaluation

The evaluation of health status questionnaires is more than psychometric analyses. Terwee et al. (2007) recommended specific quality criteria to evaluate the measurement properties of health status questionnaires adapted from the Scientific Advisory Committee, Medical Outcomes Trust (MOS) quality criteria (2002). The attributes and criteria adapted from Terwee et al. (2007) were measures for content validity (measurement aim, target population, constructs measured, item selection and interpretability), criterion-related validity, and construct validity. Clear description and adequate demonstration of each criterion is necessary for an instrument to demonstrate adequate validity (see Appendix A). The criteria for internal

consistency require either Cronbach's alpha (0.70-0.95) or factor analysis to confirm that items are measuring only the intended constructs (Terwee et al., 2007). If an item is represented in multiple factors, then that item should be considered for rejection, because it measures more than one construct and could conflate results (Smith et al., 1997; Terwee et al., 2007). Factor analysis has also been used as a technique to measure construct validity because it can determine how many distinct constructs are being measured (Jackobsson, 2009).

Reproducibility (test-retest reliability) and agreement are both measures of reliability that are included in the quality evaluation of an instrument. Agreement is measured for continuous variables using intraclass correlation coefficient (ICC) which is a ratio of variation in the population (inter-individual variation) divided by total variation which is inter-individual variation plus intra-individual variation (random error) (Cohen's kappa for ordinal variables). (Koepsell & Weiss, 2003). Studies using Pearson's correlation coefficient to measure agreement are only reporting linear association, not the consistency of the test and retest responses pair wise for each item, and each participant for the whole sample (as in the ICC) (Grafton, Foster & Wright, 2005). Responsiveness is an instrument's ability to distinguish between clinically important change and measurement error (Terwee et al., 2007). The final criterion is burden (administrative and participant).

The following is a brief description of each instrument's features, measurement properties, and validation studies summarized in Table 1 (below) and Table 2 (p.21).

Table 1. Instrument Features

Instrument/ Author (s)	Measurement aim*	Target Population	Constructs/Subscales (# of items)	Scoring	Burden†

Instrument/ Author (s)	Measurement aim*	Target Population	Constructs/Subscales (# of items)	Scoring	Burden†
West Haven Yale Pain Inventory (Kerns et al., 1985)	Discrimination - To assess individual differences among patients with chronic pain	Any persons with chronic pain	Part 1/6 scales: severity/suffering (3), interference (9), support (3), self control (2), negative mood (3); Part 2/3 scales: punishing (4), solicitous (6), distracting (4); Part 3/ 4 scales: household chores (5), outdoor work (5), activities away from home (4), social activities (4)	Respondents record their response to each item on a 7-graded scale. The response scale has fixed scores between 0 and 6, where 0 is "no, not at all" and 6 is "yes, very much". Higher scores equals more severe pain.	Can be self-administered 20 min to complete
Wisconsin Brief Pain Question-naire (BPQ) (Daut et al., 1983); later known as the Brief Pain Inventory (Cleeland and Ryan, 1983)	Discrimination and Evaluation - To measure pain severity, it's impact on the patient, and treatment responses	Persons with chronic malignant and non-malignant pain	Pain intensity in previous 1 wk (3); history (1); location (1) treatments/ medications (1), relief (1), beliefs about cause of pain (1); quality (1); pain interference with mood (1), relationships (1), walking (1), sleeping (1), normal work (1), enjoyment of life (1).	Two scores: Pain severity measured 0-10 NRS, mean score of now, worst, least and average pain over past week. Interference measured as 0-10 for each item, the total interference score is mean of all items.	Self or interviewer administra-tion; 5-15 min to complete
Graded Chronic Pain Scale (Von Korff et al., 1990)	Discrimination - To classify pain status according to recurrent, persistent and/or disabling states	Persons with varying degrees of persistent pain or non-persistent pain and pain-related disability	Pain severity subscale: interference (1), average intensity (1), daily duration (1), Persistence subscale (1); Pain-related disability subscale (1)	Scored into 1 of 4 categories Grade I, low disability-low intensity; II, low disability-high intensity; III high disability moderately limiting; and IV, high disability-severely limiting. Sub-scale scores for pain intensity, disability score and disability points.	10 min to complete, self-administered
McGill Pain Question-naire (MPQ) (Melzack, 1975)	Evaluation: To detect differences among different methods to relieve pain	General instrument to measure pain in any population	78 descriptors for pain categorized into 3 classes: sensory, affective, evaluative; (miscellaneous) and 20 subclasses.	Pain Rating Index: 2 numerical values; severity for each descriptor, and sum of rank values for each category; Present Pain Index: severity at time of administration	Instructions to be read aloud, and descriptors explained if necessary. Approx 5-15 min to complete.

Instrument/ Author (s)	Measurement aim*	Target Population	Constructs/Subscales (# of items)	Scoring	Burden†
Short Form McGill Pain Question-naire (SF-MPQ) (Melzack, 1987).	To develop a shortened version of the MPQ for when standard MPQ is too long, and qualitative information on pain is desired.	General instrument to measure pain in any population for studies that require rapid data collection	Present Pain Intensity Scale or visual analogue scale, and 11 sensory descriptors, 4 affective descriptors	Pain rating scores are the sum of intensity values for descriptors in each subclass. For MPQ, pain rating index scores are sum of rank values for each subclass.	Pain descriptors read aloud and possible explanation for each descriptor. 2-3 min to complete.

* According to Kirshner & Guyatt (1985), a discriminative instrument distinguishes between groups by a particular characteristic(s). An evaluative instrument measures longitudinal change of the intended construct.
†Burden is defined as the cost, effort and time required both by the participant and administrator to use the instrument.

Multidimensional Pain Inventory

The (West Haven–Yale) Multidimensional Pain Inventory (MPI) assesses not only pain intensity and how much pain interferes with activity, but also the way people cope with pain (Kerns et al., 1985). The MPI was based upon the cognitive-behavioral perspective of pain (Turk et al., 1987). This perspective highlights the effect of chronic pain on daily/social activities and mental health (Jakobsson, 2009). The MPI also measures family support as well as potential reinforcement of pain behaviors by an individual's significant other (Kerns et al., 1985). The MPI, is a 61 item questionnaire divided into one psychosocial section (Part 1: 6 scales), and two behavioral sections (Part 2: 3 scales; Part 3: 4 scales). A mean score is calculated for each of the 12 scales. The mean scores can be presented as above or below average when compared to data from a large heterogenous group of individuals with chronic pain (n=6,532) collected in 2004 available from the University of Pittsburgh Medical Center (UPMC) website (UPMC, 2013).

The MPI also categorizes individuals into three coping profiles: adaptive copers, the interpersonally distressed, and the dysfunctional group (Verra et al., 2012). The adaptive copers have higher perception of life control and higher activity level, less pain severity, less interference due to pain and less affective distress. The interpersonally distressed group has lower levels of perceived solicitous and distraction responses from spouses, and higher levels of

punishing responses than the other groups. The dysfunctional group reports high pain severity, high interference and activity distress, low life control, and low activity level. Individuals are grouped based on composite scores and presented as above or below average according to the normative data provided by UPMC (UPMC, 2013). Many individuals, however, cannot be classified into the standard MPI profiles, and therefore, its value as a monitoring tool for coping is limited (Harlacher, Persson, Rivano-Fischer, & Sjölund, 2011). In total, an individual can potentially have 12 scores (mean score on each scale) and one of three possible coping profiles.

Kerns et al. (1985) is the initial development publication for the MPI which contained the details of scale construction and item analysis, validity and reliability. The items for Part 1 were determined hypothetically by the authors, Part 2 was derived from interviews with significant others of people with chronic pain, and Part 3 was adopted from various activity lists and activity goals for treatment developed by chronic pain patients themselves. Expert panels were used in Part 1, but no pilot testing occurred before the instrument was administered to the sample. Pilot testing on a subsample of participants could aid in the item reduction process, especially for instruments like the MPI with a large number of items. Fewer items reduces participant burden.

The developers conducted a confirmatory factor analysis to test the a priori hypotheses and exploratory factor analysis where no priori hypotheses were generated. The items of each subscale had significant factor loading (>0.70) on one factor, representing good construct validity (Kerns et al., 1985). The developers also confirmed concurrent validity by testing the MPI's various subscales against 6 well validated instruments for correlation >0.70 with the Beck Depression Inventory (Beck, Mendelsen, Mock, & Erbaugh, 1961) and subscales of the Multidimensional Health Locus of Control Scale (Wallston & Wallston, 1981).

The scales showed good internal consistency with α ranging from 0.70-0.90 (0.70 as the minimum standard for adequate internal consistency) (Kerns et al., 1985; Terwee et al., 2007), meaning that the items within each subscale are measuring the same construct. A subsample of 60 were re-tested approximately two weeks after the baseline measure with the correlation coefficients ranging from 0.62-0.91 representing reliable variance over time (Gelinas et al., 2008). In a more recent study, ICC values were reported ranging from 0.57-0.87 after four weeks suggesting moderate reliability (Verra et al., 2012). It is possible that because the ICC is a more rigorous measure of agreement, that result more likely reflects the true error. Responsiveness of the MPI measured by effect size showed distinct differences between "affective distress" (0.37) and "locus of control" (0.36) scales, with a smaller size for "interference" (0.25) when measuring the efficacy of a rehabilitation program intervention. Thus, the MPI could be used to monitor these three specific outcomes to determine therapeutic efficacy, but the MPI generally is more relevant for discriminating between groups than evaluating the effect of an intervention.

Despite all the evidence for validity and reliability in the initial publication, the sample for the initial study was 81.5% male. The lack of diversity could reduce the overall variability and thus systematically lowering error and creating bias. Other validation studies were conducted with more diverse samples (Bergstrom et al., 1998; Jakobsson, 2009). Bergstrom et al. (1998) wasn't able to reproduce the 4 factor structure for Part 3 from Kerns et al. (1985) using a Swedish adaptation of the MPI (MPI-S), but rather produced a 3 factor structure. Bergstrom et al. (1998) posited that this may be the result of a more diverse sample (60% female); but also reflected on the relevance of that specific list of activities for a Swedish population. Thus, factor analysis of the initial publication does not necessarily reflect cultural variations.

Brief Pain Inventory

The Pain Research Group of the World Health Organization **(**WHO) Collaborating Centre for Symptom Evaluation in Cancer Care adapted the BPI, to assess malignant pain, originally called the Wisconsin Brief Pain Questionnaire (Cleeland & Ryan, 1994; Daut et al., 1983). The BPI was developed with the intention of being a brief, clear, and self-administered alternative to the MPQ. The BPI measures both the intensity of pain and interference of pain in the one's activities. It consists of front and back body diagrams where individuals can indicate location of pain, four pain severity items, seven pain interference items, and one item on pain relief by analgesics. Items are rated on a 0-10 numeric rating scale (NRS), 0 as "no pain" and 10 "worst pain imaginable" (Cleeland & Ryan, 1994).

The items measuring either "worst pain" or "average pain" can represent pain severity, or a mean composite score of the four severity items (Cleeland, 2009). BPI pain interference is typically scored as the mean of the seven interference items. This mean interference score can be used if more than 50% or four of seven, of the total items have been completed. The BPI was initially validated for cancer pain, but has since been validated for chronic non-malignant pain (Kapstad, Rokne, & Stavem, 2010; Tan et al., 2004). Validity was demonstrated in the selected studies in several ways. First, there was a significant association between increased medication use and higher pain ratings (Daut et al., 1983; Shin, Kim, Kim, Chee, & Im, 2007). Repeated factor analysis has consistently produced 2 factors (severity and interference) (Shin et al., 2007; Tan et al., 2004). The interference scale was tested against the Roland-Morris Disability Questionnaire (RMDQ) (Roland & Morris, 1983) which had a moderate correlation with BPI interference (r=0.50-0.70) and a low correlation with BPI pain severity (r=0.30-0.50) (Tan et al., 2004). Because the RMDQ and BPI interference are not identical concepts; and the RMDQ is an

alternative method of measuring interference; a moderate correlation is evidence for good convergent validity (Gelinas et al., 2008). The low correlation with BPI pain severity and RMDQ confirms that severity and interference are two distinct concepts (discriminant validity) (Tan et al., 2004). In terms of reliability, good internal consistency (α>0.80) was demonstrated among the all the selected BPI validation studies (Ger, Ho, Sun, Wang, & Cleeland, 1999; Shin et al., 2007; Tan et al., 2004). Ger et al. (1999) reported ICC ratios of 0.79 for severity and 0.81 for interference in a Taiwanese population of cancer patients providing evidence for good agreement. Tan et al. (2004) reported statistically significant improvement in a male population of veterans with chronic non-malignant pain after treatment in both intensity and interference scales from visit 1 to visit 3, with an average of 27.73 days between visits to demonstrate responsiveness of the BPI. Unfortunately, only 97 out of the original 440 participants completed all three visits, and it is possible that those who were left in the study were those most likely to improve.

Chronic Pain Grade Scale

The goal of the CPGS is "to classify a population sample into a set of mutually exclusive and exhaustive states defining clinically meaningful and empirically valid stages in the natural history of a chronic pain condition" (Von Korff et al., 1990, p.280). Von Korff et al. (1990) proposed the measurement of chronic pain severity in three dimensions: persistence (duration), intensity and disability. Persistent pain was defined as pain present on more than half the days in the prior 6 mos. Disability was measured according to the number of days in the prior 6 mos that an individual was unable to carry out usual activities (i.e. work, school, housework) due to pain. This seven item instrument has 3 subscales which are combined to calculate a chronic pain grade that enables classification of individuals into 5 hierarchical categories: grades 0 (no pain) to IV

(high disability–severely limiting) (Elliott, Smith, Smith, & Chambers, 2000; Hawker, Mian, Kendzerska, & French, 2011). All items are scored on an 11–point Likert scale, with responses ranging from 0–10, 0 as "no pain and 10 as "pain as bad as could be." The scores are then manipulated and re-coded for classification into the appropriate pain grade. Grades I and II are based on pain intensity and low disability, while Grades III and IV are based on disability regardless of pain severity (Smith et al., 1997). Similar to the BPI and MPI, the development of the CPGS was derived from the cognitive-behavioral perspective (Turk et al., 1987).

The initial validation and development study addressed a variety of conditions including back pain, headache, temporomandibular joint disorder (TMJ), abdominal pain, and chest pain; on a sample of 1,016 (Von Korff et al., 1990). Internal consistency for the severity item amongst the aforementioned conditions was $\alpha>0.73$, except for TMJ ($\alpha=0.62$), meaning that the item was measuring the same construct in each condition except for TMJ. In a later study, Cronbach's alpha was $\alpha=0.91$ demonstrating that each item is measuring the same construct, pain. The item-total correlations ranged from 0.69-0.83 (all items correlate well) when tested on a general sample of the population through a postal self-completion questionnaire without specific diagnostic categories of pain (Smith et al., 1997). Construct validity was demonstrated with CPG associations to increased pain medication use and health care utilization (Von Korff et al., 1990). Smith et al. (1997) confirmed construct validity and internal consistency using factor analysis with all items heavily loading (>0.75) on a single factor (pain). Discriminant validity was demonstrated by negative correlations with each category of the SF-36 (Ware & Sherbourne, 1992), a measure of overall well-being (Elliott et al., 2000; Smith et al., 1997). When considering the construct of disability, Dixon, Pollard, & Johnston (2006) conducted a study specifically investigating the CPGS and its ability to distinctly measure pain-related disability

according to the International Classification of Disability (ICF). (World Health Organization (WHO), 2001). The ICF is a taxonomy for disability which defines any given health condition, such as chronic pain, with three main outcomes: impairment, activity limitations, and participation restrictions within an environmental context, which is based on the functional limitation paradigm of disability (Nagi, 1965; WHO, 2001). According to Dixon et al. (2006), five individual items within the CPG were able to measure a single outcome (3 items for impairment, 1 item for activity limitations, and 1 item for participation restrictions), but 2 items measured both activity limitations and participation restrictions based on the opinions of 12 expert panelists (ICC= 0.93, 95% C.I. 0.87–0.97). Items of the CPGS that represent more than one category on the ICF could conflate disability scores. Thus, the CPGS may not be relevant for research questions based on the functional limitation paradigm or ICF model for disability.

McGill Pain Questionnaire (MPQ)

The MPQ (long or short version) is the most widely used multidimensional instrument for measuring the quality and intensity of pain (Ngham et al., 2012). It is based on the gate control theory of pain which was introduced by Melzack and Torgerson (1971). Gate control theory suggests that pain is a complex neurological phenomenon that includes a sensory, affective, and evaluative component. The MPQ was developed to measure the pain experience from multiple dimensions using quality pain descriptors represented by 4 categories: (a) sensory (pain location, intensity, quality, and pattern), (b) affective (fear, depression, and pain-related anxiety); (c) evaluative (overall pain appraisal), (d) miscellaneous (Melzack, 1975). In its original (long) form, the MPQ consists of 78 pain descriptors out of a possible 102 derived from a clinical literature search on pain. Groups of doctors, participants, and students assigned an intensity value to each word, and then put them in rank order. The rank value for each descriptor

is based on its position in the word set. The sum of the rank values is the pain rating index (PRI) (Melzack, 1975). The Present Pain Intensity (PPI) is pain at the time of administration of the questionnaire and measured according to a 6-point scale: (0) none; (1) mild; (2) discomforting; (3) distressing; (4) horrible; and (5) excruciating.

In the initial validation study a sample of 297 participants with a variety of pain-related conditions, i.e. arthritis, cancer, dental, dermatological, gastrointestinal, low back pain, menstrual, musculoskeletal, neurological, obstetric, phantom limb and post surgical pain were assessed using the MPQ (Melzack, 1975). In order to demonstrate construct validity, taped recordings of participants' comments before and after intervention regarding pain level, drug intake and activity levels were collected to compare to PPI (Melzack, 1975). The PRI scores, however, are not standardized nor have any reference values, limiting cross-sectional comparisons across different samples and generalizability. Also there may not be a relationship between the PPI and PRI due to the variability in interpretations of the pain descriptors (Melzack, 1975).

Other data that can be collected with the MPQ is number of words endorsed, severity rating of each pain descriptor chosen, and location/distribution of pain from shaded areas on a human body outline. The severity ratings per descriptor and PRI (sum of rank values) were highly correlated in each of the three categories (sensory, affective, evaluative) in the initial validation study ($r>0.9$) (Melzack, 1975). Turk et al. (1985) used confirmatory factor analysis to show that the average correlations for the PRI between the three categories (0.71) is larger than the average correlation within the categories (0.58), thus the categories are not distinct constructs, and measuring them separately can produce misclassification. Therefore, the total score of the PRI is a more appropriate measure than the individual PRI scores for each category.

Menezes Costa, Maher, McAuley, & Costa (2009) conducted a systematic review of the cross-cultural adaptations of the MPQ in order to critique their methods of adaptation. The process of cross-cultural adaptation include: initial translation, synthesis, back translation, expert committee review, and pilot test (Menezes Costa et al., 2009). Of the 29 versions of the adapted MPQ only 53%, 47%, 27%, 27%, and 20% of these steps were followed, respectively. If selecting an adapted version, gaps in the adaptation process must be reported as part of any publication or report.

Short-Form McGill Pain Questionnaire (SF-MPQ)

The SF-MPQ, a shorter version of the MPQ, is composed of 15 pain descriptors (11 sensory and 4 affective) from the long form's 78 descriptors most commonly used by 33% of the validation sample (Melzack, 1987). The five-point intensity (Present Pain Index) and visual analog scales (VAS) are included to provide indices of overall pain intensity (Hawker et al., 2011). Individuals are asked to select a descriptor for their pain and then rate its intensity from 0 (none) to 3 (severe). The Pain Rating scores are calculated by summing the item scores (range 0-45). There is no composite measure rather the scale produces three separate scores, sensory, affective and total pain rating scores (Melzack, 1987). Higher scores represent worse pain. The initial development publication showed significant correlations between the sensory, affective and total scores for both short and long MPQ form. The SF-MPQ could also detect statistically significant changes pre and post pain treatment, specifically analgesic drugs, epidural blocks, and transcutaneous electrical nerve stimulation.

Reliability of the SF-MPQ has been demonstrated by Grafton et al. (2005) who recruited 57 persons awaiting hip/knee replacement surgery to complete the SF-MPQ at baseline and 5 days (again at 10 days if no change in pain from baseline to 5 days). Strand, Ljunggren, Bogen,

Ask, & Johnsen (2008) also demonstrated reliability with a sample of 137 persons with musculoskeletal pain, and rheumatic pain (inflammatory and osteoarthritic) assessed at 1 and 3 days apart. Both studies reported high ICC values for total and sensory scores, representing low measurement error; lower ICC values were observed in both studies for the affective dimension. Absolute reliability, measured as smallest detectable difference (SDD) or minimal clinically important difference (MCID), was smaller in Grafton (2005) most likely due to the fact that the sample was limited to those whose pain hadn't changed, excluding 12.7% of the sample (n=9). However Strand et al. (2008) used an external criterion to confirm the change in pain after treatment (Patient's Global Impression of Change (PGIC) scale) (Buchbinder, Bombardier, Yeung, & Tugwell, 1995). In completing the questionnaires, 75% of the sample completed by mail had errors, versus 27% completed in clinic, reflecting possibility of individuals needing clarification of descriptors (Grafton et al., 2005).

Shin et al. (2007) compared the SF-MPQ in a sample of predominantly female Asian Americans with different types of cancer-related pain. Factor analysis confirmed 2 factors for the SF-MPQ as hypothesized. In the MPQ-SF, the item "tender" did not contribute to the measure and four items: (1) punishing/cruel, (2) splitting, (3) aching, (4) sickening were considered redundant, which is most likely related to ethnic differences in pain descriptions. Thus ethnic differences in pain measurement can occur in English-speaking populations as well.

Discussion

All of the multidimensional instruments selected in this paper could potentially be used to collect high quality data to address a variety of research questions. The information presented in this paper is meant to guide researchers toward the most appropriate choice for a given study. The most important aspect of selecting an instrument is to assure that the research question is

congruent with the underlying theory of pain on which the instrument was based. The MPI, BPI, and CPGS are based on the cognitive behavioral theory of pain, while the MPQ is based on the gate control theory (Cleeland & Ryan, 1994; Kerns et al., 1985; Melzack, 1975; Von Korff et al., 1990). The MPQ doesn't address how pain interferes with one's physical or social activities. The MPI, BPI and CPGS don't consider the affective domains of pain, such as fear or anxiety. For example, in conducting a study of the relationship between fear of falling and pain, the MPQ could provide information to compare the mean severity rating of pain descriptors such as "fearful, terrifying, or frightful" with the Falls Efficacy Scale (Tinnetti, Richman, & Powell, 1990). The BPI, on the other hand, would be a better choice to test the hypothesis that fear of falling is associated with pain severity and pain interference.

One of the main findings of this paper is the nuances in interpreting and evaluating factor analysis for construct validity, especially in non-diverse samples. Bergstrom et al. (1998) found a different factor structure for the MPI when adapting it for Swedish population, than the initial publication. He posited that this may be the result of a more diverse sample (60% female); but could represent the how the specific list of activities translate for Swedish population. For example, "work in the garden" or "mow the lawn" may not be relevant when gardening season is only 4 months out of the year in Sweden (Bergstrom et al., 1998). After excluding 5 items from Part 3, the model fit was sufficient for 3 factors, instead of the original 4 factor solution. Cultural adaptation is also not the same as language translation as demonstrated by Shin et al. (2007) where analysis of a population of English-speaking Asian Americans resulted in a different factor structure. One reason could've been that the pain descriptor "tender" was interpreted as soft (Shin et al., 2007). When conducting population-specific research using culturally adapted versions of the instruments, a thorough investigation of the cross-cultural adaptation process that

was used to develop and validate that version is necessary. It is important to assure that any common characteristic of a non-diverse sample does not systematically bias results, including cultural differences beyond language.

Evaluating individual characteristics of each instrument are the next consideration. The MPI is unique in that it collects an extensive amount of information, not captured by the other selected instruments for this paper, such as perceived life control (including perceived ability to solve problems and feelings of personal mastery and competence) and appraisal of support received from a spouse or significant other and family (Kerns et al., 1985). It assesses an individual's perception of how others respond to displays of pain and suffering, by using item responses such as "expresses sympathy" or "ignores me." Therefore, this instrument may be more suitable for a mixed methods approach because it is congruent with studies seek to understand phenomena in terms of the meanings people bring to them, and quantifying the effect of that phenomenon on different aspects of health. The major drawback is that scores are not readily calculated or interpreted due to the number of scales, and comparisons to normative data. The MPI takes approximately 20 minutes to complete and can be self-administered, keeping resource costs lower, but has a large number of items increasing participant burden. It has been validated in English and 6 other languages (Verra et al., 2012).

The BPI has the advantage of being highly cited, and validated for many different chronic diseases and conditions (Dworkin et al., 2005, 2008). It is easy to score, interpret, and the results are amenable to a variety of statistical tests. There is evidence that single items of the pain severity scale can be used separately, thus making it easier to measure responsiveness over longer periods, such as "pain in the past month" which is a more stable measure than "pain in the last 24 hours" (Dworkin et al., 2005). The BPI is easy to administer, quick, and easy to score

(Cleeland, 2009). The score is easily interpretable and based on a simple NRS (0-10) which is very familiar to most people, and quantifies qualitative data on level of activity limitations due to pain. The burden on participant is low for effort and time spent, and it can be self-administered rather than interviewer administered (depending on the sample population) decreasing administrative burden. It is the most economical choice, and provides information on the most dimensions with the least amount of time and administrative cost. The BPI, however, only measures four dimensions of pain: location, severity, interference, and pain relief by analgesics (Cleeland & Ryan, 1994). If a more comprehensive measure of pain dimensions is necessary, then the BPI would not be an appropriate choice, but is otherwise appropriate to a variety of study designs. The BPI has been translated and validated in 12 languages (Cleeland, 2009).

The CPGS is classification system that allows for one overall rating for global pain severity and activity limitation, as opposed to separate ratings for activity limitation and severity as in BPI and MPI. As noted earlier, CPGS may not be relevant for research questions based on the functional limitation paradigm or ICF model for disability (Nagi, 1965; WHO, 2001). The CPGS is a taxonomy for distinguishing between groups and not a tool to measure therapeutic efficacy. The developers of this scale proposed its use in epidemiological studies to report on prevalence, relative risks, and natural history of pain (Von Korff et al., 1990). It is easy to administer, but the scoring is very complex, increasing administrative burden and cost of resources. The CPGS has not been validated in languages beyond English and Italian, so it is less appropriate for culturally diverse samples, than the other selected instruments (Sallaffi, Stancati, & Grassi, 2006). It takes approximately 10 min to complete the CPGS.

The MPQ is the oldest and most widely used multidimensional pain instrument worldwide (Ngham et al., 2012). It is the only instrument in this selection that collects

information on quality of pain and severity per quality. The number of descriptors endorsed can also be calculated to investigate individuals who endorse fewer or more qualities. Because the MPQ is not based on any reference values, it may be more relevant for measuring longitudinal changes, therapeutic efficacy or repeated measures analysis where individuals' answers can be matched and compared at different time points. The MPQ has been adapted into 26 different languages (Menezes Costa et al., 2009). Participants can take 25-30 minutes to complete the MPQ long version (Ngham, 2012), the longest out of the selected instruments for this paper. It may require an interviewer if there are any questions about the definitions of the pain descriptors. The differences between the MPQ and the SF-MPQ are that the MPQ collects more data and is based on three dimension model, as opposed to only two in the SF-MPQ. The SF-MPQ takes 2-5 min to complete, thus the more economical choice if expanded data is not necessary.

Conclusion

Pain is a multidimensional experience that can be best measured by questionnaires that consider its effect on multiple dimensions of physical, psychological and social functioning. There is no consensus on a general theory of pain, so research questions must be congruent with the theory of pain from which the instrument was generated. In selecting an appropriate instrument for research purposes, a thorough investigation of the instrument's features and psychometric properties is required. If using a culturally adapted version of an instrument the cultural adaption process must be scrutinized; or if conducting a study on a non-diverse sample, researchers must be wary of culturally dependent differences that could affect the instrument's properties to effectively measure distinct constructs. Further population-specific validation studies are needed, especially for English-speaking minority populations.

Table 2. Quality Evaluation of Initial Development and Validation Studies

Instrument/ Author (s)	Sample Population for validation study	Concurrent validity	Construct validity	Internal consistency	Reproducibility/ Agreement	Responsiveness	Item Selection
West Haven-Yale Multidimensional Pain Inventory (WHYMPI) (Kerns et al., 1985)	120 (81.5% male) patients referred consecutively from 2 major VA medical centers in West Haven, CT. Mean age 50.8 yrs ±14.5; mean duration of pain 10.2 yrs, predominately low back pain (36.4%)	No criteria were used to compare the overall instrument's validity; though validated questionnaires were used to compare subscales with their intended constructs	Factor analysis indicated a 4 factor solution for the intercorrelations of 12 subscales, also compared with 9 validated instruments for discriminant/convergent validity: correlations were as hypothesized.	Cronbach's alpha ranged from 0.70-0.90 among the 12 scales	60 participants were re-tested 2 weeks later with stability coefficients ranging from 0.62-0.91.	No data available	61 total items; Items for Part I were based on the cognitive-behavioral perspective of pain; Part II - derived from interviews of significant others; Part III - adopted from various lists of activities and activity goals
Wisconsin Brief Pain Questionnaire (BPQ) (Daut et al., 1983); later known as the Brief Pain Inventory (Cleeland and Ryan, 1994)	1200 inpatients and outpatients (85% female) with breast, prostate, colorectal, gynecological cancer (ages 57-67); 34 outpatients (84% female) with rheumatoid arthritis (ages 48-58); recruited from the Wisconsin Clinical Cancer Center and Univ of Wisconsin Rheumatology Clinic	Authors note that no gold standard criteria exist for comparison	Factor analysis revealed 2 factors (severity: 4 items; and interference: 7 items). Increased medication use for higher pain ratings; high correlation between usual pain ratings and interference r=0.624, p<0.001; Patterns of correlations among pain and interference measures were different for different diseases	No data available	Pain history was retested after a brief period (1-7d) (r=0.59-0.93) and extended period (4-224d) (r=0.22-0.34)	No data available	McGill Pain Questionnaire was used as a model for development

Instrument/ Author (s)	Sample Population for validation study	Concurrent validity	Construct validity	Internal consistency	Reproducibility/ Agreement	Responsiveness	Item Selection
Graded Chronic Pain Scale (Von Korff et al., 1990)	1016 persons enrolled the Group Health Cooperative (HMO) with back pain, headache, temporomandibular disorder (TMD); abdominal pain, and chest pain; later sample with 242 persons specifically with TMD	No comparable instruments	Graded pain status was found to be consistently associated with increased psychological impairment, frequent use of health care and pain medications.	Internal consistency was measured for each pain condition: 0.75 for back pain; 0.79 for headache; 0.73 for abdominal pain; 0.79 for chest pain; and 0.62 for TMD	Results for the initial sample were replicated for the second sample for frequency of medication use and health care, self-rated health, and psychological impairment.	No data available	Noted a multidimensional conceptual model as framework, but no specific protocol for how items were selected.
McGill Pain Questionaire (MPQ) (Melzack, 1975)	297 persons with a variety of diagnostic pain categories arthritis, cancer, dental, dermatological, gastrointestinal, low back pain, menstrual, musculoskeletal, neurological, obstetric, phantom limb and post surgical.	Taped recordings of patient's comments before and after intervention regarding pain level, drug intake and activity levels were collected to compare with rank order variable and present pain index.	The descriptors were rank ordered consistently by the doctors, patients and students consulted who have different cultural, socio-economic, and educational background.	For 4 categories of descriptors: sensory, affective, evaluative, miscellaneous and total inter-correlations > 0.9. Number of words chosen and scale value: r=0.97; number of words chosen and rank value: r=0.89	After 3 to 7 days there was mean consistency of 70.3% among choice of subclasses and the same Present Pain Index score.	Although rated the same numerically, the rank order variable showed a 10% decrease in pain post treatment, showing that the rank order is a more valid index of change.	Authors identified 102 descriptors from clinical literature on pain, then groups of doctors, patients and students assigned intensity value to each word, and then put in rank order

Instrument/ Author (s)	Sample Population for validation study	Concurrent validity	Construct validity	Internal consistency	Reproducibility/ Agreement	Responsiveness	Item Selection
Short Form McGill Pain Questionaire (SF-MPQ) (Melzack, 1987).	70 patients in postop and obstetrical floors, and musculoskeletal pain (physical therapy) from Montreal General Hospital	The SF-MPQ was tested against the original MPQ (long form)	Based on MPQ	The scores for the SF-MPQ and MPQ were significantly correlated for the affective, sensory and total scores for all types of pain studied and for French and English forms.	Relationships were consistent between English and French translations of the SF-MPQ and MPQ, No other data available.	Long and short form scores were significantly correlated for before and 30 min after pain treatment for each group: analgesic drugs, and epidural blocks, with a statistically significant reduction in pain for each group.	The descriptors chosen were the most commonly used by 33% or more of patients with different types of pain (listed in MPQ)

Table 3. Quality Evaluation of Validation Studies for Selected Instruments

Author (s)	Purpose	Sample Population	Construct validity	Concurrent Validity	Internal consistency	Reproducibility/ Agreement	Responsiveness
Multidimensional Pain Inventory (MPI)							
Bergstrom et al., 1998	To test the reliability and factor structure of a Swedish translation of the (WHY)MPI, the MPI-S, and the generalizability of the factor structure of WHY(MPI)	682 patients suffering from chronic musculoskeletal pain consecutive, referrals to three pain clinics in Sweden, specializing in the rehabilitation of chronic musculoskeletal pain, mean age 41.8 ±9.4, mean duration of pain 67.9 mos ±70.4	Three psychologists separately translated the (WHY)MPI into Swedish, then pilot tested. It was also compared both on a semantic and an empirical basis using factor loadings extracted from exploratory factor analyses.	(WHY)MPI, English German and Dutch translations provided criteria for comparison to Swedish version MPI (MPI-S), scale intercorrelations for other versions were in accordance with MPI-S.	3 sections: part 1 $\alpha \geq 0.80$ with exclusion of items 13, 16; part 2 $\alpha = 0.76$-0.86 excluding items 1, 3; part 3 $\alpha = 0.67$–0.81 excluding items 2, 3, 6, 7 and 16	test-retest coefficients 0.73–0.89, re-tested within 2 weeks	No data available
Jakobsson, 2009	To psychometric-cally evaluate the brief version of the MPI-S	384 people aged 18–102 years with chronic pain most commonly caused by osteoarthritis, cardiovascular diseases and back injuries, selected from a Swedish population register; Chronic pain defined as pain for at least 3 months.	Four factors were extrapolated for the brief version from the original 11 factors in the long form: F1 pain severity, F2 interference, F3 life control, and F4 affective distress	MPI-S compared to brief version of MPI-S, correlation coefficients were all significant	α for the four factors (F1–F4) in the total sample were F1: 0.90; F2: 0.93; F3: 0.81; and F4: 0.90.	No data available	No data available

Author (s)	Purpose	Sample Population	Construct validity	Concurrent Validity	Internal consistency	Reproducibility/ Agreement	Responsiveness
Vera et al., 2012	To examine the test-retest stability of the MPI Taxonomy Classification; and internal consistency and test-restest reliability at the scale level for the MPI German version in patients with persistent musculoskeletal pain	204 persons with chronic musculoskeletal pain (82% chronic non-specific back pain) and had pain for at least 6 months recruited from a rehab clinic in Switzerland	No data available.	No data available.	Cronbach alpha range 0.76-0.86 for scales; except for negative mood alpha=0.60 and distracting responses alpha=0.69	Test-retest reliability at an average 4- week time interval, for the mean MPI scores ICC = 0.72 and 0.87; except for MPI scale life control, ICC = 0.57; 159 patients (78%) had a stable MPI subgroup classification at 4 wks: kappa values of (0.58-0.70)	No data available
Brief Pain Inventory (BPI)							
Tan et al., 2004	To determine the reliability and validity of the BPI Intensity and Interference scales for assessing pain in persons referred to a multidisciplinary pain center with the primary complaint of pain	440 patients with chronic pain referred to the chronic pain center at a metropolitan Veteran Affairs Medical Center; male (91.8%), mean age 54.9 yrs (range 21-85 yrs); 72.3% white, 21.2% black, and 5.4% other	A factor analysis was performed, resulting in 2 factors that accounted for 67.8% of the variance as predicted.	BPI interference correlated moderately with pain-related disability measured by RMDQ (r=0.57) because they measure similar but not identical concepts	Cronbach alpha coefficient .85 for the Intensity scale and .88 for the Interference scale	No data available	BPI Intensity and Interference scales showed significant changes in the expected direction from visit 1 to visit 3, with 27.73 avg number of days between visits; 97 out of 440 completed visit 3

Author (s)	Purpose	Sample Population	Construct validity	Concurrent Validity	Internal consistency	Reproducibility/ Agreement	Responsiveness
Kapstad et al., 2010	To evaluate the psychometric properties of the BPI in patients with OA of the hip undergoing total hip replacement surgery (THR)	250 persons from 6 hospitals in Norwegian counties on the waiting list for THR; 70% female, mean age 69 ±10; mean duration of pain prior to surgery 6.3 yrs ±6.7	Correlation coefficient for BPI pain severity and WOMAC pain r=0.66, pain interference and WOMAC pain r=0.57; SF-36 pain and BPI severity r= -0.58; SF-36 pain and BPI interference r= -0.65	BPI was compared with WOMAC, and SF-36 to assess discriminant and convergent validity, all hypothesized relationships were significant and in the correct direction.	α >0.80 for the BPI pain severity index and function interference index	No data available	The responsiveness indices (ES, SRM and RI) for change from baseline to 1 year after THR minimum values of 1.57 for BPI severity and 1.52 for BPI interference
Keller et al., 2004	To determine the validity of the BPI for measuring non-cancer pain by evaluating its relationship to generic and condition-specific pain measures.	Sample of 250 persons recruited from primary care clinics with osteoarthritis, rheumatoid arthritis, and low back pain with or without worker's compensation	2 factor solution: (severity and interference) was confirmed, accounting for 67% of variance.	Highly correlated with SF-36 Bodily Pain, RMDQ, and HAQ. BPI able to differentiate between categories of CPG.	α=0.82-0.95, similar for arthritis and low back pain groups	No data available	SRM for change from baseline to follow up visit (no interval noted), Persons who had different levels of change on the criterion scales also had significantly different scores on the BPI severity and interference scales.

Author (s)	Purpose	Sample Population	Construct validity	Concurrent Validity	Internal consistency	Reproducibility/ Agreement	Responsiveness
Chronic Pain Grade Scale (CPG)							
Smith et al., 1997	To evaluate the CPG as a self completion questionnaire administered to a general sample drawn from a UK sample with no diagnostic distinction.	293 patients aged over 18, stratified for age, gender and receipt or non-receipt of regular prescriptions for pain-relieving medication.	Factor analysis identified 1 factor. All 7 items had a factor loading greater than 0.75	CPG was compared to each category within the SF-36; all significantly negatively correlated (SF-36 is a measure of positive well-being)	α=0.9132, and the item-total correlations 0.6885-0.8285	No data available	No data available
Elliott et al., 2000	To test the responsiveness and validity of the CPG over time in a group of chronic pain patients from a general population	A random sample of 450 chronic pain patients, ≥25 yrs from an existing cohort and stratified for age, gender and chronic pain severity >3, pain for longer than 3 months from North East Scotland, surveyed at baseline and at 18 months;	No data available.	SF-36 scores decreased as CPG score became more severe, increased as CPG score became less severe, and remain constant when there was no change in the CPG score.	No data available.	No data available.	No data available

Author (s)	Purpose	Sample Population	Construct validity	Concurrent Validity	Internal consistency	Reproducibility/ Agreement	Responsiveness
Dixon et al., 2006	To explore the ability of the CPG to operationalize the ICF definition for disability.	12 health professionals were recruited from the Institute of Applied Health Sciences at the University of Aberdeen and asked to categorize each of the 7 items from the CPG into the 3 categories of ICF: impairment, activity limitations and participation restrictions	Individual items within the CPG were able to measure a single outcome (3 items for impairment, 1 item for activity limitations, and 1 item for participation restrictions) but 2 items measured both activity limitations and participation restrictions.	Items of the CPG that represent more than one category on the ICF could conflate disability scores by confounding more than one outcome on the ICF.	No data available.	The ICC for all judgements across all seven items was 0.93 (95% C.I. 0.87–0.97). The ICC for each construct was as follows, 0.95 (95% C.I. 0.88–0.99) for Impairment, 0.94 (95% C.I. 0.85–0.99) for Activity limitations 0.95 (95% C.I. 0.86–0.99) for Participation restrictions	No data available
McGill Pain Questionnaire (MPQ)							
Menezes Costa et al., 2009	To identify the available cross-cultural versions of the MPQ, to describe the clinimetric testing for each version and to evaluate the quality of the translation procedures.	Systematic review of 53 studies, representing 29 different culturally adapted versions of the MPQ	Assessed by comparing the MPQ scores with many different constructs ranging from self-report measures such as pain intensity, depression, anger, and anxiety scales to functional impairment tests and laboratory markers of disease status.	No data available	15 out of 53 studies reported internal consistency (most subscales $\alpha>0.7$)	Reproducibility was assessed in 11 studies; with ICC of the Pain Rating Index Total score ranging from 0.72-0.97.	11 studies tested responsiveness, 9 of 11 tested internal responsiveness (e.g., effect sizes and t-tests) rather than external responsiveness

Author (s)	Purpose	Sample Population	Construct validity	Concurrent Validity	Internal consistency	Reproducibility/ Agreement	Responsiveness
Lazaro et al., 2001	To investigate the psychometric properties and comprehension of a Spanish version of the MPQ in several Spanish-speaking Latin American countries.	Collaborating researchers recruited participants from Argentina (n=40), Costa Rica (n=24), Mexico (n=96) and Panama (n=45) to use the MPQ-SV in their usual clinical practice. Total n=282.	Values from Latin American countries were compared to those from Spain were the MPQ-SV was developed and validated. All correlation coefficients were >0.9 for PRI and severity of descriptor scale.	MPQ-SV PRI scores were compared to Visual Analogue and Verbal rating scales, correlation coefficient for Spain and Panama >0.6; Argentina, Costa Rica and Mexico coefficients ranged from 0.38-0.22.	No data available	No data available	No data available
Turk et al., 1985	To test the 3 factor structure of the Pain Rating Index of the MPQ using a confirmatory factor analysis.	2 samples were tested for cross-validation. 70 persons referred to VA pain clinic mean age 50.4, mean duration of pain 10.7 yrs, 81% male with range of conditions; 98 persons referred from orthopedic back pain clinic mean age 45.8, mean duration of pain 6.9yrs, 64% female	The 3 components (sensory, affective, and evaluative) of the PRI do not display adequate discriminant validity, thus the total PRI score is more appropriate than the use of 3 subscales independently.	No data available	Sensory α=0.78; affective α=0.71; evaluative α=0.46; total scale α=0.84; Average correlation between factors is 0.71 and average correlations within factors is 0.58.	No data available	No data available

Author (s)	Purpose	Sample Population	Construct validity	Concurrent Validity	Internal consistency	Reproducibility/ Agreement	Responsiveness
Short-Form McGill Pain Questionnaire (SF-MPQ)							
Grafton et al., 2005	To evaluate the test-retest reliability of SF-MPQ in patients with osteoarthritis	Serial evaluation of 57 persons (mean age 64.8 ± 10.4) awaiting hip or knee replacement surgery, consecutively recruited in an orthopedic clinic or via mail in England whose pain was unchanged at 10 and 15 days from baseline.	Cites other studies of validity	Cites other studies of validity	No data available	Total pain score ICC= 0.96; sensory ICC=0.95; affective ICC=0.88; average pain score ICC=0.89 Current pain score ICC=0.75	For a change in SFMPQ to be a clinical change, it must be > 5.2 for the total score (possible 0-45) or >4.5 for sensory component; any change less than those parameters is measurement error.
Strand et al., 2008	To examine the relative and absolute test-retest reliability and responsive-ness to clinically important change of the SF-MPQ scales among different conditions	Consecutive recruitment of 137 persons with MSK, rheumatic, and osteoarthritic pain from 2 Norwegian hospitals followed 1-3 days apart for test-retest, and longitudinally before and after treatment for responsiveness.	Cited validation study of Norwegian version of SF-MPQ conducted by Ljunggren et al. (2007).	Used Patient Global Impression of Change (PGIC) scale as external criterion of clinically important change, ≤3 defined as important change	No data available	ICC ranged from 0.95-0.79 for rheumatic pain; and 0.76-0.63 for MSK pain	SDD for MSK pain =11.9 points on a 0-45 scale; SDD for rheumatic pain=7.5 (for two repeated measures of the same individual) 1-3 days apart; sensitivity 0.81; specificity 0.45.

Author (s)	Purpose	Sample Population	Construct validity	Concurrent Validity	Internal consistency	Reproducibility/ Agreement	Responsiveness
Shin et al., 2007	To evaluate and compare psychometric properties of the SF-MPQ and BPI among Asian American cancer patients	119 self-identified Asian Americans convenient sample recruited through a facility and internet; mean age 52.2 ±10.9, and 82.4% female	2 factors each were extracted from the SF-MPQ and BPI; pain scores were positively correlated with the usage of pain medications for the SF-MPQ (r= 0.23-0.33); for the BPI-SF (r=0.40-0.42); "punishing/cruel, splitting, aching, and sickening" were considered redundant	No data available	Cronbach's alpha coefficients ranged from 0.85-0.94 for the SF-MPQ, and from 0.91-0.97 for the BPI; total item correlation (0.69-0.90)	No data available	No data available

References

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*(6), 561.

Bergström, G., Jensen, I. B., Bodin, L., Linton, S. J., Nygren, A. L., & Carlsson, S. G. (1998). Reliability and factor structure of the multidimensional pain inventory--swedish language version (MPI-S). *Pain, 75*(1), 101-110.

Buchbinder, R., Bombardier, C., Yeung, M., & Tugwell, P. (1995). Which outcome measures should be used in rheumatoid arthritis clinical trials?. *Arthritis & Rheumatism, 38*(11), 1568-1580.

Cleeland, C. S. (2009). *The Brief Pain Inventory user guide.* Houston, TX: The University of Texas MD Anderson Cancer Center.

Cleeland, C. S., & Ryan, K. M. (1994). Pain assessment: global use of the Brief Pain Inventory. *Annals of the Academy of Medicine, Singapore, 23*(2), 129-138.

Daut, R. L., Cleeland, C. S., & Flanery, R. C. (1983). Development of the wisconsin brief pain questionnaire to assess pain in cancer and other diseases. *Pain, 17*(2), 197-210.

Dixon, D., Pollard, B., & Johnston, M. (2007). What does the chronic pain grade questionnaire measure?. *Pain, 130*(3), 249-253.

Dworkin, R. H., Turk, D. C., Wyrwich, K. W., Beaton, D., Cleeland, C. S., Farrar, J. T., ... & Zavisic, S. (2008). Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *The Journal of Pain, 9*(2), 105-121.

Dworkin, R. H., Turk, D. C., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Katz, N. P., ... & Witter, J. (2005). Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain, 113*(1), 9-19.

Elliott, A. M., Smith, B. H., Smith, W. C., & Chambers, W. A. (2000). Changes in chronic pain severity over time: the Chronic Pain Grade as a valid measure. *Pain, 88*(3), 303-308.

Gélinas, C., Loiselle, C. G., LeMay, S., Ranger, M., Bouchard, E., & McCormack, D. (2008). Theoretical, psychometric, and pragmatic issues in pain measurement. *Pain Management Nursing: Official Journal of the American Society of Pain Management Nurses*, *9*(3), 120-130.

Ger, L. P., Ho, S. T., Sun, W. Z., Wang, M. S., & Cleeland, C. S. (1999). Validation of the Brief Pain Inventory in a Taiwanese population. *Journal of Pain and Symptom Management, 18*(5), 316-322.

Grafton, K. V., Foster, N. E., & Wright, C. C. (2005). Test-retest reliability of the Short-Form McGill Pain Questionnaire: Assessment of intraclass correlation coefficients and limits of agreement in patients with osteoarthritis. *The Clinical Journal of Pain, 21*(1), 73-82.

Harlacher, U., Persson, A. L., Rivano-Fischer, M., & Sjölund, B. H. (2011). Using data from Multidimensional Pain Inventory subscales to assess functioning in pain rehabilitation. *International Journal of Rehabilitation Research*, 34(1), 14-21.

Hawker, G. A., Mian, S., Kendzerska, T., & French, M. (2011). Measures of adult pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), Short Form-36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP) *Arthritis Care & Research*, *63*(S11), S240-S252.

Institute of Medicine of the National Academies Report. (2011). *Relieving pain in America: A blueprint for transforming prevention, care, education, and research.* Washington DC.: The National Academies Press,.

Jakobsson, U. (2009). Psychometric testing of the brief screening version of multidimensional pain inventory (swedish version). *Scandinavian Journal of Caring Sciences, 23*(1), 171-179.

Jensen, M. P. (2003). Questionnaire validation: a brief guide for readers of the research literature. *The Clinical Journal of Pain, 19*(6), 345-352.

Jensen, M. P., Molton, I. R., Groah, S. L., Campbell, M. L., Charlifue, S., Chiodo, A., . . . Tate, D. (2012). Secondary health conditions in individuals aging with SCI: Terminology, concepts and analytic approaches. *Spinal Cord, 50*(5), 373-378.

Kapstad, H., Rokne, B., & Stavem, K. (2010). Psychometric properties of the Brief Pain Inventory among patients with osteoarthritis undergoing total hip replacement surgery. *Health and Quality of Life Outcomes, 8*(1), 148.

King, I. M. (1981). A theory for nursing: Systems, concepts, process. New York: Wiley.

Keller, S., Bann, C. M., Dodd, S. L., Schein, J., Mendoza, T. R., & Cleeland, C. S. (2004). Validity of the brief pain inventory for use in documenting the outcomes of patients with noncancer pain. *The Clinical Journal of Pain, 20*(5), 309-318.

Kerns, R. D., Turk, D. C., & Rudy, T. E. (1985). The west haven-yale multidimensional pain inventory (WHYMPI). *Pain, 23*(4), 345-356.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases, 38*(1), 27-36.

Klepstad, P., Loge, J. H., Borchgrevink, P. C., Mendoza, T. R., Cleeland, C. S., & Kaasa, S. (2002). The norwegian brief pain inventory questionnaire: Translation and validation in cancer pain patients. *Journal of Pain & Symptom Management, 24*(5), 517-525.

Koepsell, T. D., & Weiss, N. S. (2003). Epidemiologic methods: studying the occurrence of illness. New York: Oxford University Press.

Kumar, S. P. (2011). Utilization of brief pain inventory as an assessment tool for pain in patients with cancer: A focused review. *Indian Journal of Palliative Care, 17*(2), 108.

Melzack, R. (1987). The short-form McGill pain questionnaire. *Pain, 30*(2), 191-197.

Melzack, R. (1975). The McGill pain questionnaire: Major properties and scoring methods. *Pain, 1*(3), 277-299.

Melzack, R., & Torgerson, W. (1971). On the language of pain. *Anesthesiology, 34*(1), 50-59.

Menezes Costa, L. D. C., Maher, C. G., McAuley, J. H., & Costa, L. O. P. (2009). Systematic review of cross-cultural adaptations of McGill Pain Questionnaire reveals a paucity of clinimetric testing. *Journal of Clinical Epidemiology, 62*(9), 934-943.

Ngamkham, S., Vincent, C., Finnegan, L., Holden, J. E., Wang, Z. J., & Wilkie, D. J. (2012). The McGill Pain Questionnaire as a multidimensional measure in people with cancer: an integrative review. *Pain Management Nursing, 13*(1), 27-51.

Roland, M., & Morris, R. (1983). A study of the natural history of back pain: part I: development of a reliable and sensitive measure of disability in low-back pain. *Spine, 8*(2), 141-144.

Salaffi, F., Sarzi-Puttini, P., Ciapetti, A., & Atzeni, F. (2011). Clinimetric evaluations of patients with chronic widespread pain. *Best Practice & Research Clinical Rheumatology, 25*(2), 249-270.

Salaffi, F., Stancati, A., & Grassi, W. (2006). Reliability and validity of the Italian version of the Chronic Pain Grade questionnaire in patients with musculoskeletal disorders. *Clinical Rheumatology, 25*(5), 619-631.

Scientific Advisory Committee of the Medical Outcomes Trust (2002). Assessing health status and quality-of-life instruments: attributes and review criteria. *Quality Of Life Research, 11*(3), 193-205.

Shin, H., Kim, K., Kim, Y. H., Chee, W., & Im, E. O. (2007). A comparison of two pain measures for Asian American cancer patients. *Western Journal of Nursing Research, 29*(5), 545-560.

Smith, B. H., Penny, K. I., Purves, A. M., Munro, C., Wilson, B., Grimshaw, J., ... & Smith, W. C. (1997). The Chronic Pain Grade questionnaire: validation and reliability in postal research. *Pain, 71*(2), 141-147.

Strand, L. I., Ljunggren, A. E., Bogen, B., Ask, T., & Johnsen, T. B. (2008). The Short-Form McGill Pain Questionnaire as an outcome measure: Test–retest reliability and responsiveness to change. *European Journal of Pain, 12*(7), 917-925.

Tan, G., Jensen, M. P., Thornby, J. I., & Shanti, B. F. (2004). Validation of the Brief Pain Inventory for chronic nonmalignant pain. *The Journal of Pain, 5*(2), 133-137.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., ... & de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34-42.

Tinetti, M. E., Richman, D., & Powell, L. (1990). Falls efficacy as a measure of fear of falling. *Journal of Gerontology, 45*(6), P239-P243.

Turk, D. C., Meichenbaum, D., & Genest, M. (1987). *Pain and behavioral medicine: A cognitive behavioral perspective (Vol. 1)*. New York, NY: Guilford Press.

Turk, D. C., Rudy, T. E., & Salovey, P. (1985). The McGill Pain Questionnaire reconsidered: confirming the factor structure and examining appropriate uses. *Pain, 21*(4), 385-397.

University of Pittsburgh, UPMC Pain Medicine Program (2013). *Multidimensional Pain Inventory (MPI) Computer Program, Version 3.0.* Retrieved from http://www.pain.pitt.edu/mpi/

Verra, M., Angst, F., Staal, J., Brioschi, R., Lehmann, S., Aeschlimann, A., & de Bie, R. (2012). Reliability of the Multidimensional Pain Inventory and stability of the MPI classification system in chronic back pain. *BMC Musculoskeletal Disorders*, 13(1), 155.

Von Korff, M., Dworkin, S. F., & Le Resche, L. (1990). Graded chronic pain status: an epidemiologic evaluation. *Pain, 40*(3), 279-291.

Wallston, K. A., & Wallston, B. S. (1981). Health locus of control scales. *Research with the Locus of Control Construct, 1*, 189-243.

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical Care,* 473-483.

World Health Organization (2001). *International classification of functioning, disability, and health: ICF*. Geneva: World Health Organization.

Wright, K. D., Asmundson, G. J., & McCreary, D. R. (2001). Factorial validity of the short-form McGill pain questionnaire (SF-MPQ). *European Journal of Pain*, *5*(3), 279-284.

Appendix A. Quality Criteria for Measurement Properties of Health Status Questionnaires

Property	Definition	Adequate quality criteria
Internal consistency	Internal consistency is a measure of the homogeneity of a (sub) scale. It indicates the extent to which items in a (sub) scale are inter-correlated, thus measuring the same construct. Factor analysis should be applied to determine the dimensionality of the item this is, to determine whether or not they formed only one overall dimension or more than one.	Factor analyses performed on adequate sample size AND Cronbach's alpha(s) calculated per dimension AND Cronbach's alpha(s) between 0.70 and 0.95;
Construct validity	Content validity examines the extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured.	Specific hypotheses were formulated AND at least 75% of the results are in accordance with these hypotheses
Reproducibility/ agreement	The degree to which repeated measurements in stable persons (test retest) provide similar answers. The extent to which the scores on repeated measures are close to each other (absolute measurement error)	ICC or Kappa >0.70; Test-Retest correlation coefficients >0.70
Responsiveness	The ability of a questionnaire to detect clinically important change over time in the concept being measured. Predefined hypotheses about the relation of change in the instrument to corresponding changes in reference measures should be postulated.	Smallest detectable change per individual or Smallest detectable change per group <Minimal important change OR Minimal important change outside the limits of agreement OR Responsiveness ratio > 1.96 OR Area under the curve >0.70;
Property	Definition	Adequate quality criteria

Content validity	Content validity examines the extent to which the concepts of interest are comprehensively represented by the items in the questionnaire	Clear description of measurement aim; target population; concepts; item selection and item reduction methods; interpretability of the items.
Concurrent validity	Concurrent validity is a measure of correlation between the instrument and a criterion.	Correlation coefficients between 0.40 and 0.60
Burden	The time, effort, and other demands placed on those to whom the instrument is administered (participant burden) or on those who administer the instrument (administrative burden).	Information on: average and range of the time needed to complete the instrument; any resources required for administration of the instrument

Research Proposal

Chronic pain is a highly prevalent condition affecting at least half of the older population, often contributing to disability and other poor health outcomes. Few rigorous studies have examined the older adult's pain experience and its functional outcomes for many reasons, such as limited research funding in the area of pain, the reluctance of elders to report pain, the assumption that pain is a normal part of aging, and a historical disease orientation among researchers and clinicians resulting in little focus on the problem of pain itself.

Although quality of pain is often collected as part of the health history for elders with chronic pain, its clinical value and relevance in the assessment and management of pain is poorly understood. The purpose of this study is to examine the significance of the quality of pain descriptors reported by older adults and their relation to adverse functional outcomes. The aims are summarized in Table 1. The first aim of the study is to explore quality of pain descriptors including prevalence and their correlates in older adults. In the second aim, I will examine the persistence of quality of pain descriptors over 18 months and their relation to other pain characteristics such as location, severity, and pain interference with activity. I hypothesize that the quality of chronic pain in older adults is a persistent characteristic of the pain experience. In addition, I hypothesize that the relation between quality of pain and other pain characteristics will continue over the 18 month follow-up. The third aim of this study is to examine the relation between quality of pain in older adults and physical function over the 18 month follow-up. I hypothesize that selected quality of pain descriptors will be associated cross-sectionally and longitudinally with decline in physical function.

The population-based MOBILIZE Boston Study is one of the first longitudinal studies to examine multiple pain domains and disability in a population of older adults in the community.

Pain was assessed using a variety of standardized instruments, including the Brief Pain Inventory and McGill Pain Questionnaire, and a joint pain questionnaire assessing pain in major joint areas. These instruments are unique in that they incorporate sensory, affective and evaluative components of the pain experience and include assessment of 20 pain quality descriptors (aching, stiffness, sharp, shooting, throbbing, etc.). Using SAS statistical packages, I will assess the factor structure of pain quality measures using confirmatory factor analysis based on the Short-Form McGill Pain Questionnaire (Cleeland & Ryan, 1983) in terms of model fit and hierarchical factor loadings on specified factors and latent global pain factor. I will also use latent variable analysis to estimate associations between the various pain quality constructs and disability.

Based on the multi-dimensional aspects of pain, along with the special characteristics of the aging population, quality of pain is an often ignored characteristic in the profile of the pain experience. A careful examination of quality of pain descriptors will help us to understand chronic pain and its functional consequences in the older population. This work will open a new avenue for research to better understand the problems of pain and disability in older adults.

Table 1
Research Study Aims

Specific Aims	Objectives	Hypotheses
Aim 1	To explore quality of pain descriptors including their prevalence and correlates in the older population	Specific quality of pain descriptors will be associated with health and sociodemographic characteristics and with other characteristics of chronic pain including severity, location, and pain interference in the older population.
Aim 2	To derive constructs based on aggregates of pain quality descriptors and examine these constructs in relation to other pain characteristics (location, severity, and pain interference) and chronic conditions.	A defined set of pain quality constructs will be associated with other pain characteristics and pain-related chronic conditions in older adults.
Aim 3	To determine whether quality of pain constructs are associated cross-sectionally and longitudinally with physical function in the older population.	Pain quality constructs will be predictive of decline in physical functioning over 18 months, as measured by self-report and by physical performance testing.