Pain 125 (2006) 208–215

PAIN

www.elsevier.com/locate/pain

Review and recommendations

Developing patient-reported outcome measures for pain clinical trials: IMMPACT recommendations

Dennis C. Turk [a,*], Robert H. Dworkin [b], Laurie B. Burke [c], Richard Gershon [d], Margaret Rothman [e], Jane Scott [c], Robert R. Allen [f,1], J. Hampton Atkinson [g], Julie Chandler [h], Charles Cleeland [i], Penny Cowan [j], Rozalina Dimitrova [k], Raymond Dionne [l,2], John T. Farrar [m], Jennifer A. Haythornthwaite [n], Sharon Hertz [c], Alejandro R. Jadad [o], Mark P. Jensen [a], David Kellstein [p,3], Robert D. Kerns [q,r], Donald C. Manning [s], Susan Martin [t], Mitchell B. Max [l,2], Michael P. McDermott [b], Patrick McGrath [u], Dwight E. Moulin [v], Turo Nurmikko [w], Steve Quessy [x], Srinivasa Raja [n], Bob A. Rappaport [c], Christine Rauschkolb [y,4], James P. Robinson [a], Mike A. Royal [z,5], Lee Simon [aa], Joseph W. Stauffer [ab,6], Gerold Stucki [ac], Jane Tollett [ad], Thorsten von Stein [ae,7], Mark S. Wallace [g], Joachim Wernicke [af], Richard E. White [ag], Amanda C. Williams [ah], James Witter [c], Kathleen W. Wyrwich [ai]

[a] *University of Washington, Seattle, WA 98195, USA*
[b] *University of Rochester School of Medicine and Dentistry, Rochester, NY, USA*
[c] *United States Food and Drug Administration, Rockville, MD, USA*
[d] *Northwestern University, Chicago, IL, USA*
[e] *Johnson and Johnson, Raritan, NY, USA*
[f] *AstraZeneca, Wilmington, DE, USA*
[g] *University of California San Diego, La Jolla, CA, USA*
[h] *Merck and Company, Blue Bell, PA, USA*
[i] *University of Texas, M.D. Anderson Cancer Center, USA*
[j] *American Chronic Pain Association, Rocklin, CA, USA*
[k] *Allergan, Inc, Irvine, CA, USA*
[l] *National Institute of Dental and Craniofacial Research, Bethesda, MD, USA*
[m] *University of Pennsylvania, Philadelphia, PA, USA*
[n] *Johns Hopkins University, Baltimore, MD, USA*
[o] *University Health Network and University of Toronto, Toronto, Canada*
[p] *Novartis Pharmaceuticals, East Hanover, NJ, USA*
[q] *VA Connecticut Healthcare System, West Haven, CT, USA*
[r] *Yale University, New Haven, CT, USA*
[s] *Celgene Corporation, Warren, NJ, USA*
[t] *Pfizer Global Research and Development, Ann Arbor, MI, USA*
[u] *Dalhousie University, Halifax, Nova Scotia, Canada*
[v] *London Regional Cancer Centre, London, Ont., Canada*

[*] Corresponding author. Tel.: +1 206 616 2626; fax: +1 206 543 2958.
E-mail address: Turkdc@u.washington.edu (D.C. Turk).
[1] Present address: Wyeth, Madison, NJ, USA.
[2] Present address: The National Institute of Nursing Research, Bethesda, MD, USA.
[3] Present address: Wyeth Consumer Healthcare, Madison, NJ, USA.
[4] Present address: Johnson and Johnson, Titusville, NJ, USA.
[5] Present address: Solstice Neurosciences, Malvern, PA, USA.
[6] Present address: Alpharma, Piscataway, NJ, USA.
[7] Present address: Tercica, Brisbane, CA, USA.

doi:10.1016/j.pain.2006.09.028

[w] University of Liverpool, Liverpool, UK
[x] GlaxoSmithKline, Research Triangle Park, NC, USA
[y] Johnson & Johnson, Raritan, NJ, USA
[z] Alpharma, Elizabeth, NJ, USA
[aa] Harvard Medical School, Boston, MA, USA
[ab] Alpharma, Piscataway, NY, USA
[ac] University of Munich, Munich, Germany
[ad] US Department of Veterans Affairs, Washington, DC, USA
[ae] NeurogesX, Inc, San Carlos, CA, USA
[af] Eli Lilly and Co., Indianapolis, IN, USA
[ag] Endo Pharmaceuticals Inc., Chadds Ford, PA, USA
[ah] St. Thomas Hospital, London, UK
[ai] Saint Louis University, St. Louis, MO, USA

Received 9 January 2006; received in revised form 22 May 2006; accepted 18 September 2006

1. Introduction

Two of the most difficult problems in evaluating treatment efficacy in pain research involve what constitutes a "successful" outcome and how best to measure it (Turk et al., 2003). Definitions of success reflect the agendas and values of the parties who evaluate the treatment. Because variability in outcome measures across clinical trials hinders evaluations of treatments, the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) recommended that six core outcome domains should be *considered* when designing chronic pain clinical trials: (1) pain; (2) physical functioning; (3) emotional functioning; (4) participant ratings of improvement and satisfaction with treatment; (5) symptoms and adverse events; and (6) participant disposition (Turk et al., 2003).

Study endpoints can be grouped into categories based on the source of the information: patient-reported outcomes (PROs), laboratory tests, device measurements, and behavioral observation, clinician-reported outcomes (CROs), and "third party" outcomes (e.g., disability, health care utilization). Each source provides unique information, for example:

- PROs document patients' perceptions of the impact of disease and treatment on health and functioning, and include patients' evaluations of their health status, symptoms, adherence to treatment, satisfaction, and the impact of disease on functioning and well-being (Acquadro et al., 2003; Willke et al., 2004).
- Laboratory, behavioral, and device measurements include objective and usually quantitative behavioral or physiological measures often performed by devices or by raters (e.g., sedimentation rate, quantitative sensory testing, structured observation protocols).
- CROs include outcomes either observed by a provider or requiring interpretation (e.g., radiologic results, blood chemistry). CROs also include scales completed by a health care provider using information about the patient. CROs that are completed by clinicians but that require patient input should be distinguished from PROs that clinicians administer because the former involve clinician judgment or interpretation whereas the latter involve unmodified patient responses (Willke et al., 2004).

The fact that the associations among laboratory endpoints, observation, CROs, and PROs are far from perfect highlights the complementary nature and importance of both subjective and objective assessments. PROs are particularly important for conditions that involve symptoms such as pain and fatigue where objective measures of patient perceptions are not available. In these instances, CROs, laboratory, device, and observation are at best surrogate markers.

Based on a set of evidence-based reviews, IMMPACT recommended consideration of the use of a set of PROs assessing pain, physical functioning, emotional functioning, participant ratings of improvement, and satisfaction with treatment (Dworkin et al., 2005). Although specific measures were recommended, caution was expressed as each measure had limitations. A general concern was that no attempts had been made to include relevant patient groups in decisions about what outcomes were meaningful, or whether the instructions, item content, or anchors of the scales were clearly described. Thus, the IMMPACT recommendations encouraged that such information be obtained not only in relation to currently available measures but also in the development of new outcome measures. A third IMMPACT meeting was convened to determine consensus recommendations regarding instrument evaluation and development.

2. Consensus meeting procedure

The IMMPACT-III meeting was held on November 20–22, 2003 and included an international group of 44

participants from universities, United States governmental agencies, the international pharmaceutical industry, and an international self-help organization. Participants were invited on the basis of their research, clinical, or administrative expertise relevant to the design and evaluation of pain treatment outcomes. An attempt was made to include broad representation of various disciplines while keeping the size of the meeting relatively small to promote frank discussion. The primary purpose of the IMMPACT meeting was to develop consensus recommendations for methods that should be used for developing new patient reported outcome (PRO) measures and for evaluating the adequacy and appropriateness of existing measures for clinical trials of the efficacy and effectiveness of treatments for pain.

Since not all of the attendees at IMMPACT-III were familiar with recent innovations in psychometric theory and the possible contribution of item response theory (IRT) and computer adaptive testing (CAT), a background paper on IRT (Reeve, 2003) was circulated prior to the meeting. In addition, an overview of IRT and CAT was presented by one of the authors of this article (RG).

3. Recommended approaches for developing PRO measures for pain trials

The development of PRO measures involves a series of sequential steps beginning with consideration of what construct (latent variable) or constructs will be assessed. Attention must be given to the specific goals of the measure, its intended uses, and the characteristics of the individuals to whom it will be administered.

Selection of an outcome measure should be justified based on the domains of interest for the patient and the characteristics of the treatment and its putative effects. The lack of existing well-validated measures that address the hypotheses being tested may necessitate the developed of a new measure. In some instances, however, adding and testing a few items to a well-validated scale will obviate the need to develop a new instrument. The outcome measure must assess the domains of interest and the specific measures selected must be appropriate for the population for which the treatment is being considered, and must be reliable and valid with the minimum of patient burden possible. Table 1 contains a step-by-step sequence that should be used in developing a new assessment instrument. Although presented in a linear fashion, the results obtained at certain points may lead to return to previous steps to refine the instrument.

Because instrument development takes considerable time, effort, and resources, existing measures should be examined to determine whether a new measure is necessary. In reviewing available measures, attention should not only be given to the adequacy of the psychometric properties of the measure but also to the availability of appropriate information to confirm the measure's psychometric properties for the population of interest. Of course, some latitude should be allowed according to reasonable scientific judgment. There are likely to be instances where an investigator could use a PRO in a somewhat different patient group without having to go through an entirely new, time-consuming, and costly development and evaluation process. Once the need for a new measure is established, the formal process of instrument development can begin as outlined in

Table 1
Recommended process for developing outcome measures for pain clinical trials[a]

- I. Identify scientific approach
  - A. Overall question
  - B. Conceptual model or theoretical approach
  - C. Scope of assessment
- II. Establish
  - A. Target population
  - B. Factors or concepts to be included
    - 1. Specific goal of outcome measure
    - 2. Specific traits
    - 3. Need for independent or overlapping subscales
- III. Develop item pool
  - A. Methods
    - 1. Literature review
    - 2. Focus groups with patients and experts
    - 3. In-depth interviews with patients and experts
  - B. Determine format
    - 1. Individual items
    - 2. Scale properties
  - C. Consider methods of
    - 1. Data collection
    - 2. Scoring
    - 3. Analysis
- IV. Item evaluation
  - A. Components
    - 1. Minimize patient burden
    - 2. Evaluate language and cross-cultural equivalence
    - 3. Test in target population (cognitive interviewing or debriefing)
    - 4. Revise and repeat as necessary to finalize format and item wording
    - 5. Develop scoring algorithm
  - B. Measurement approach
    - 1. Classical test theory
    - 2. Item response theory
  - C. Field test items
    - 1. Collect response data for items from target population
    - 2. Assess dimensionality of items
    - 3. Locate "gaps" in the construct assessment
- V. Instrument evaluation – evaluate psychometric properties in target populations
  - A. Reliability
  - B. Validity
  - C. Responsiveness
- VI. Complete instrument development
  - A. Revise instrument if necessary
  - B. Finalize instrument
  - C. Develop user manual and instructions to respondents

[a] Although the recommended sequence is presented as if it were a linear process, the development of measures is frequently an iterative process.

Table 1. Finally, it is incumbent on authors of any new measure to demonstrate whether a newly developed measure has incremental advantages including decreased participant burden or increased reliability or validity – to recommend it over relevant existing instruments. For example, a new measure that has comparable psychometric properties to existing measures but that requires less time for patients to complete might be preferred.

3.1. Measurement theory

After the initial development of a measure is completed, its quality must be assessed. Specific steps are taken to demonstrate the adequacy of the psychometric properties of a measure. In particular, measures must be precise, reliable, and valid. Formal scale development is based on a set of important and well-established methods (e.g., Nunnally and Bernstein, 1994; Committee on Standards for Educational and Psychological Testing of the American Educational Research Association et al., 1999). We will not review these methods here but refer the reader to the classic texts in the field cited above. It is important to emphasize, however, that the responsiveness of a potential outcome measure – that is, its ability to assess changes in patients – must be established as early as possible in its development. Moreover, outcome measures must be able to assess changes that are clinically relevant. These properties are essential and without these characteristics a measure would not contribute in a meaningful way to the assessment of treatment outcomes.

3.2. Content and item development

The specific items to be included in a measure should represent important features of the construct being assessed. For example, a measure being developed to assess the impact of pain on emotional functioning must include questions that assess the range of positive and negative moods, such as depression, anxiety, anger, and vigor. A measure of the impact of pain on physical functioning could be designed to either capture the entire content domain or different components of the construct of physical functioning (e.g., role functioning vs. activities of daily living vs. movement or posture) depending on its intended use. Furthermore, a measure of physical functioning should include different content depending on its planned use (e.g., people with spinal cord injuries vs. those with carpal tunnel syndrome). Different pain-related impairments would be expected to have different impacts on specific types of physical functioning (e.g., walking versus ability to write or type).

Numerous studies have shown discrepant perceptions between patients and clinicians regarding health status, impact of disease, and treatment outcome priorities and preferences (e.g., Rothwell et al., 1997; Clinch et al., 2001; Hewlett et al., 2001). The value that patients place on different outcomes is largely unknown (Kvien and Heiberg, 2003). People with particular diseases or symptoms have unique perspectives on the impact of the disease and its treatment on their everyday functioning and well-being and thus are of critical importance in developing a new measure. Exploring subjective experiences of people with personal knowledge of the targeted syndrome or symptoms makes it possible to identify additional important experiences beyond those considered relevant by clinicians, and these should be incorporated within the core outcome measures used in clinical trials (Kirwan et al., 2003).

Focus groups and individual interviews should be used to identify content domains that are considered important by patients. Relevant information that has intuitive value to patients is a key element in determining the content to be included in a measure. It is important that focus groups include individuals with a range of pain or symptom severity as the severity of current symptoms can influence the importance that is ascribed to different outcomes (e.g., Casarett et al., 2001). The composition of the groups should also reflect the demographics of the patients to whom the measure will be applied because factors such as age, sex, and ethnicity might affect priorities and preferences for different outcomes (Ganz, 2002).

There are specific issues in the development of outcome measures for pain clinical trials that have not received adequate attention. First, it is not clear whether outcome measures of physical functioning should assess the interference of pain with physical functioning, the patient's overall level of physical functioning, or both. In deciding which of these approaches should be used, it is necessary to consider the characteristics of the patients being evaluated. For example, a sample of participants in a study with spinal cord injuries might distinguish between the inability to walk up a flight of stairs as a result of a functional impairment rather than pain, whereas participants with osteoarthritis might attribute their inability to walk up the stairs to pain rather than structural impairment. It is therefore likely that outcome measures that directly assess the interference of pain on physical functioning will be more responsive to analgesic treatment benefits than measures that more generally assess physical limitations. Similarly, measures of emotional functioning could assess, for example, general levels of anxiety and depression, but it is also possible to assess the impact of pain on these and other aspects of emotional functioning. Likewise, in an assessment of sleep in patients with pain, is the investigator interested in the impact of pain on sleep or patients' overall level of sleep disturbance?

As clinical trials are increasingly conducted in multiple countries, the ease of translating concepts and items

into other languages for use with cultures other than the one(s) for whom they were originally developed needs to be evaluated. Even if questions can be literally translated, it is important to consider whether the concepts are meaningful and are being interpreted similarly across cultures.

Once the overall content domain has been selected and specific items have been developed, attention must be given to the instructions, item wording, time-frame, response categories, scale anchors, and response format. Groups in whom the measure is going to be used must be able to clearly understand the instructions and item wording. The time-frame for the items (e.g., past week, month, current) also needs to be carefully considered as the responses will have different interpretations depending on the time interval used. Moreover, retrospective reports of long periods of time may be influenced by memory, current symptoms, or anchoring events (Stone and Shiffman, 1994).

Once a preliminary set of items has been selected and instructions have been developed, pilot testing with cognitive interviewing should be conducted on a sample drawn from the study population to establish that the targeted patient groups clearly understand the instructions, item wording, reference period, and response format. At this point, the interim measure should be pilot tested in the target population. Depending on the results of the pilot test, the items may need to be revised and retested. Pilot testing may reveal that there are insufficient items covering particular aspects of the construct. In this case, new items may need to be written in order to fill the "gaps" (e.g., insufficient items at the low and high ends of a scale creating floor or ceiling effects). This process might need to be repeated several times in order to finalize the instructions and the item content.

It is essential that detailed records be kept of the processes and steps used in the development and testing of the instructions, items, recall period, and response categories. These should explicitly specify the methods that were used to include relevant groups in determining the content covered by the measure. Test developers need to determine, in advance, the appropriate sample size and representation of the sample that will be used to evaluate the psychometric properties of an instrument (e.g., Nunnally and Bernstein, 1994; Committee on Standards for Educational and Psychological Testing of the American Educational Research Association et al., 1999).

3.3. Item Response Theory (IRT)

Classical test theory (CTT) has a long history and continues to provide valuable tools for assessing the adequacy of measures. In CTT, it is assumed that a respondent's score is a linear combination of responses to a set of questions that are sampled from a universe of questions measuring a common trait (construct) such as pain. Observed scores are partitioned into their true and error components. However, it has become increasingly recognized that CTT has important limitations (Table 2). In particular, CTT produces measures that are sample dependent. Demonstration that an instrument is reliable in one sample does not mean that it will be reliable when used in another. It is practically impossible to create measures that capture the full range of responses (e.g., activities essential for independent function such as bathing oneself vs. discretionary activities such as ability to engage in gainful employment) because any one sample would be forced to answer many irrelevant items, creating an unreasonable respondent burden.

Another important limitation to CTT is that even though there are a large number of measures that have been developed for certain outcomes (e.g., depression, physical functioning), it is impossible to compare scores on one measure with scores on another equally reliable measure. This is because different measures not only consist of different items that do not necessarily reflect highly similar constructs, but also have different item difficulty (e.g., toileting vs. driving), breadth and depth of measurement, formats, response scales, and time frames. For example, based on commissioned critiques of the literature, IMMPACT (Dworkin et al., 2005) recommended use of either the Brief Pain Inventory Interference Scale (Cleeland and Ryan, 1994) or the Multidimensional Pain Inventory Interference Scale (Kerns et al., 1985) to assess physical functioning. Although these two scales are believed to measure similar constructs, they comprise different items and use different time frames and response scales. Comparing studies using these two different scales is therefore problematic, even though they putatively are measuring the same construct, because they may not provide comparable assessments.

Item Response Theory (IRT) was developed in response to the limitations of CTT. IRT is a statistical theory consisting of various nonlinear mathematical models and related statistical techniques that attempt to link responses by individuals to locations on a continuum that reflects an unobservable construct or "latent trait" (Hambleton, 2000). IRT models express the probability of a particular response to a scale item as a function of the quantitative attribute (unobservable, latent trait) of the person and certain characteristics (parameters) of an item (Chang and Reeve, 2005). IRT is designed to describe explicitly the functional relationship between individuals' responses at the item level and the characteristics (parameters) of the items on the test. Because there is a non-linear relationship between response to an item and the latent variable (i.e., the concept the item is intended to measure), estimates based on IRT are not sample dependent, redundant items are unnecessary and possibly detrimental (e.g., in small

Table 2
Advantages and limitations of classical test theory (CTT) and item response theory (IRT)

CTT	IRT
Major advantages	
1. Minimal assumptions about data (no distributional assumptions)	1. Improves efficiency and reduced respondent burden by reducing the number of items required to establish measures with comparable precision and reliability
2. Basic model about test scores is simple (test score = true score + error)	2. Yield scores that do not vary with the characteristics of the population with respect to the underlying trait
3. Formulas for estimating amount of error in test sores, influence length of test on reliability and validity, the impact of score range restrictions on correlations, and the correction of the correlations between variables are readily available	3. Facilitates evaluation of whether items are equivalent in meaning to different populations
4. Long history	4. Permits inclusion of items with different response formats in the same scale
5. Easily understood by consumers	5. Permits identification of item and scale performance associated with group membership
	6. Allows for linking of scores from different questionnaires to measure the same construct
	7. Assesses group differences in both item and scale functioning
	8. Item statistics are independent of the ability of the sample used in item calibration
	9. Permits the development of tests (e.g., CAT) that can be individualized to each participant
Major limitations	
1. Statistics that describe items, performance tasks, item difficulty, and discrimination are sample dependent	1. Assumes only one construct (pain, emotional or physical functioning) is measured by the items in a unidimensional scale. Assumption cannot strictly be met because cognitive, personality, and test-taking factors always affect test performance
2. Test score is scale dependent, thus difficult to compare scores of different measures of comparable constructs	2. Assumes that if the trait level is held constant, item responses should be uncorrelated for for individuals at the same level
3. Single estimate of measurement error for all individuals	3. Requires large sample sizes
4. Focus on test characteristics rather than on item characteristics	4. Software currently available is not ‘‘user- friendly’’
5. Requires strictly parallel test forms	5. Parameter estimates from different samples are only linearly related, and adjustments for ability differences must be made for the item statistics from two samples to be comparable
6. Floor and ceiling effects are common	6. Choosing a model and determining consequences of model misfit require considerable experience and data analytic skills
	7. IRT models are complex and difficult for consumers (investigators, clinicians) to understand

multi-item scales redundant items can hurt the validity of a scale), and scales and items can be evaluated for having comparable characteristics for the outcome of interest (Reeve and Fayers, 2005).

The aim of IRT models is to make predictions about constructs such as health status and functional abilities of individuals from as few a number of items as possible. In contrast, CTT does not focus on the person but is applied at the level of the scale or test (the psychometric properties of the scale or test). IRT-based measures have important properties that provide advantages over CTT for health outcomes measurement: (1) each question in a scale is characterized with a set of properties that describes its relationship with a measured construct and how the item functions within a study population; (2) IRT item properties are relatively invariant with respect to the sample of respondents and respondent scores are relatively invariant with respect to the set of items used; and (3) scores can be compared or combined despite individuals receiving different sets of IRT-calibrated questions. The advantage of IRT-based measures is that not every item needs to be administered to an entire sample and hence IRT provides a potentially more efficient method of data collection. This makes IRT an important analytical tool to evaluate items, to scale the performance of a questionnaire, to evaluate the equivalence of content for an instrument that is used for different populations or in different settings, and to link two or more instruments on a common metric (Gershon et al., 2003; Chang and Reeve, 2005). Several recent efforts to apply IRT methodologies to pain and rehabilitation have been reported (e.g., Roorda et al., 2005). Recently, in the United States, the National Institutes of Health have funded a collaborative project, Patient Reported Outcomes Measurement Information System (PROMIS), to examine the application of IRT in areas related to pain and physical and emotional functioning. This project is in its early stages and the

potential of IRT in pain-related research remains to be demonstrated.

Features of IRT provide the potential for the use of Computer Adaptive Testing (CAT). CAT is a technology for interactive administration of questions tailored to the individual's responses. Testing is adaptive in the sense that questions are selected from a large item pool on the basis of the individual's ongoing responses. This item bank is calibrated using the appropriate IRT model based on responses to the item set. Once an item bank has been created and calibrated, a computer program then proceeds through an algorithm of choosing the item most appropriate for the trait level of a respondent (e.g., severity of pain), estimating the respondent's trait level based on his or her response to the item, and then choosing the next most appropriate item from the bank until a pre-specified level of measurement precision is reached. This approach reduces the need for redundant items and lowers respondent burden (Reeve, 2003). The value of CAT for measures that focus on multidimensional constructs, for which items are not assumed to be interchangeable, however, remains to be established (De Vet et al., 2003).

Despite the advantages noted, there are several limitations inherent in IRT (see Table 2). A key assumption in most currently employed IRT models is that items are unidimensional, that is, that the set of items are measuring a single continuous latent variable. The fact that it is possible to select a pool of items intended to measure very different concepts and still find underlying unidimensionality with IRT models creates problems when it is necessary to describe what the questionnaire is measuring. Thus, it is important to distinguish statistical from conceptual unidimensionality. Many of the constructs of importance in pain clinical trails are multidimensional and even though IRT may force unidimensionality by eliminating items, this process may eliminate items that are important in pain clinical trials. Specifically, the initial IMMPACT meeting suggested a number of important domains in pain clinical trials, namely, pain, physical functioning, emotional functioning, patient global ratings of improvement (change), and side effects. Combining these in a statistical manner, although potentially feasible, would likely involve the loss of important information about the specific effects of treatment. One strategy to address this potential limitation is the use of factor analysis to examine patterns of covariation among responses, and if multidimensionality is found, then each factor can be used as a unique scale if doing so would be consistent with the overall theoretical approach.

Other disadvantages of IRT include the need for large sample sizes, the absence of user-friendly software, and a lack of a standard set of recognized fit indices. Finally and importantly, the models used in IRT are complex and difficult to describe to many consumers, including many clinical researchers and health care providers.

IRT is a promising approach that will require further refinements before it can be recommended for use in clinical trials. It is most likely that the approach will be of use in measurements of physical functioning but probably less so for symptoms such as pain and fatigue as IRT assumes an underlying dimension (e.g., the pain experience) that may not be valid. IRT models are not panaceas that resolve all problems identified with CTT. There will continue to be need for the systematic work in item and instrument development discussed above and in Table 1.

4. Conclusions

There is a lack of standardized and comprehensive outcome measures for pain trials that have adequate comparative information for relevant samples, that can be used across a variety of research applications, and that allow investigators to combine or compare groups with different demographic or disease characteristics. The need to develop such measures provided the impetus for our recommendations, which are intended to be considered in the development of new measures for use in clinical trials of treatments for pain. As we have discussed, measures based on CTT have a number of limitations, including respondent burden, inability to compare measures putatively assessing the same construct, and assumptions about linearity that may be unwarranted. IRT methods offer potential advantages, and there is a growing awareness of the potential of IRT to complement and even replace some traditional psychometric approaches. CAT will likely be used increasingly as the technology improves and familiarity with this approach grows. However there are also important limitations of these newer approaches as discussed and enumerated in Table 2.

These recommendations have important implications for those who are conducting or planning clinical trials, as well as for regulators and the developers of analgesic interventions. As recommended in a previous IMMPACT consensus paper (Dworkin et al., 2005) a small set of existing instruments are already available that should be considered for use in chronic pain clinical trials. It would be useful to explore existing datasets in which such instruments are included and to refine outcome measures based on observations and the protocol outline in Table 1. The pooling and analysis of data from previously published studies using common outcome measures would permit the development of a validated, dynamic system to establish item pools from which individually tailored PROs could be selected, and would facilitate comparisons among outcome studies and enhance measurement precision of treatment outcomes. One such interdisciplinary

initiative is already being supported by the United States National Institutes of Health. This PROMIS collaboration is an attempt not only to examine the utility of IRT methodology but also to determine the feasibility of pooling data from different study sites and across different measures of a set of common outcome domains, including pain, sleep, emotional functioning, and physical functioning It would be useful for clinical investigators, regulators, and the pharmaceutical industry to join forces with patient advocacy groups in participating in the development of measures to be used in pain clinical trials.

5. Disclaimer

The views expressed in this article are those of the authors. No official endorsement by the US Department of Veterans Affairs, US Food and Drug Administration, US National Institutes of Health, or the pharmaceutical companies that provided unrestricted grants to the University of Rochester Office of Professional Education should be inferred.

Acknowledgements

The authors thank Bryce Reeve for his careful reading, suggestions regarding limitations of IRT, and useful comments on an early draft of this manuscript, and to Paul J. Lambiase and Mary Gleichauf for their invaluable assistance in the organization of the IMMPACT meeting. Abbott Laboratories, Allergan, Alpharma, AstraZeneca, Celgene, Eli Lilly and Co., Endo Pharmaceuticals Inc., GlaxoSmithKline, Johnson and Johnson, Merck and Co., NeurogesX, Novartis Pharmaceuticals, Pfizer, and Schwarz Biosciences provided unrestricted grants to the University of Rochester Office of Professional Education to support the consensus meeting.

References

Acquadro C, Benson R, Dubois D, Leidy NK, Marquis P, Revicki D, et al. PRO Harmonization Group. Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the patient-reported outcomes (PRO) harmonization group meeting at the Food and Drug Administration, February 16, 2001. Value Health 2003;6:522–31.

Casarett D, Karlawish J, Sankar P, Hirschman K, Asch D. Designing pain research from the patient's perspective: what trial endpoints are important to patients with chronic pain? Pain Med 2001;2:309–16.

Chang C-H, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. Eval Health Prof 2005;28:264–82.

Clinch J, Tugwell P, Wells G, Shea B. Individualized functional priority approach to the assessment of health related quality of life in Rheumatology. J Rheumatol 2001;28:445–51.

Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med 1994;23:129–38.

De Vet JCW, Terwee B, Bouter LM. Current challenges in clinimetrics. J Clin Epidemiol 2003;56:1137–41.

Dworkin R, Turk D, Farrar J, Haythornthwaite J, Jensen M, Katz N, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113:9–19.

Ganz PA. What outcomes matter to patients. A physician–researcher point of view. Med Care 2002;40(Suppl III):III-11–III-19.

Gershon R, Cella D, Dineen K, et al. Item response theory and health-related quality of life in cancer. Expert Rev Pharmacoeconomics Outcomes Res 2003;3:783–91.

Hambleton RK. Emergence of item response modeling in instrument development and data analysis. Med Care 2000;38(Suppl. II):II-6–II-65.

Hewlett S, Smith AP, Kirwan JR. Values for function in rheumatoid arthritis: patients, professional, and public. Ann Rheum Dis 2001;60:928–33.

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association, 1999.

Kerns RD, Turk DC, Rudy TE. The West Haven–Yale Multidimensional Pain Inventory (WHYMPI). Pain 1985;23:345–56.

Kirwan J, Heiberg T, Hewlett S, Hughes R, Kvien T, Ahlmen M, et al. Outcomes from the patient perspective workshop at OMERACT 6. J Rheumatol 2003;30:868–72.

Kvien TK, Heiberg T. Patient perspective in outcome assessments – perceptions or something more? J Rheumatol 2003;30:873–6.

Nunnally J, Bernstein I. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.

Reeve BB. Item response theory modeling in health outcomes measurement. Expert Rev Pharmacoeconomics Outcomes Res. 2003;3:131–45.

Reeve BB, Fayers P. Applying item response theory modeling for evaluating questionnaire item and scale properties. In: Fayers P, Hays RD, editors. Assessing quality of life in clinical trials: methods of practice. 2nd ed. Oxford University Press, 2005. p. 55–73.

Roorda K, Molenaar IW, Lankhorst GJ, Bouter LM. Improvements of a questionnaire measuring activity limitation in rising and sitting down in patients with lower-extremity disorders living in home. Arch Phys Med Rehabil 2005;86:2204–10.

Rothwell PM, McDowell Z, Wong CK, Dorman PJ. Doctors and patients don't agree: cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. BMJ 1997;124:1580.

Stone AA, Shiffman S. Ecological momentary assessment (EMA) in behavioral medicine. Ann Behav Med 1994;16:199–202.

Turk DC, Dworkin RH, Allen RR, Bellamy N, Brandenburg N, Carr DB, et al. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. Pain 2003;106:337–45.

Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. Controlled Clin Trials 2004;25:535–52.