Published in final edited form as:
J Pain Symptom Manage. 2009 October ; 38(4): 615–628. doi:10.1016/j.jpainsymman.2008.11.016.

Linking Pain Items from Two Studies onto a Common Scale using Item Response Theory

Wen-Hung Chen, PhD[1], **Dennis A. Revicki, PhD**[1], **Jin-Shei Lai, PhD**[2], **Karon F. Cook, PhD**[3], and **Dagmar Amtmann, PhD**[3]

[1]Center for Health Outcomes Research, United BioSource Corporation, Bethesda, MD

[2]Center for Outcomes Research and Education, Evanston Northwestern Healthcare and Northwestern University School of Medicine, Evanston, IL

[3]Department of Rehabilitation Medicine, University of Washington, Seattle, WA

Abstract

This study examined two approaches to linking items from two pain surveys to form a single item bank with a common measurement scale. Secondary analysis of two independent surveys: IMMPACT Survey with Main Survey (959 chronic pain patients; 42 pain items) and Pain Modules (N=148; 36 pain items); and CORE Survey (400 cancer patients; 43 pain items). There were common items included among the three data sets. The two approaches were examined, one in which all items were calibrated to an item response theory (IRT) model simultaneously and another in which items were calibrated separately and then the scales were transformed to a common metric. The two approaches produced similar linking result across the two sets of pain interference items because there was sufficient number of common items and large enough sample size. For pain intensity, simultaneous calibration yielded more stable results. Separated calibration yielded unsatisfactory linking result for pain intensity because of a single common item with small sample size. The results suggested that simultaneous IRT calibration method produced the more stable item parameters across independent samples, hence, is recommended for developing comprehensive item banks. Patient reported health outcome surveys are often limited in sample sizes and the number of items owing to the difficulty of recruitment and the burden to the patients. As a result, the surveys either lack statistical power or limited in scope. Using IRT methodology, surveys data can be pooled to lend strength to each other to expand the scope and to increase the sample sizes.

Keywords

item response theory; pain intensity; pain interference; linking studies

INTRODUCTION

Modern measurement theory methods are being used increasingly to develop and evaluate existing and new symptom assessment and health outcome instruments (1,2,3,4,5,6,7). For example, the NIH sponsored Patient-Reported Outcome Measurement Information System (PROMIS) initiative is focused on developing and evaluating item banks designed to measure pain, fatigue, physical function, social function, emotional distress and other domains. Item response theory (IRT) analyses have been instrumental in the psychometric evaluation of these

For more information contact Wen-Hung Chen, PhD, Center for Health Outcome Research, United BioSource Corporation, 7101 Wisconsin Ave., Suite 600, Bethesda, MD 20814; Telephone: 1-301-654-9729; Fax: 1-301-654-9864; wen-hung.chen@unitedbiosource.com.

item banks (8,9). While common in the educational testing field, scale linking has not been frequently applied to health outcome measures (4). However, there are situations where scale linking is needed such as when researchers are developing several large health outcome item banks each consisting of a large number of items. In order to reduce burden on the participants, item banks may be divided into several comparable subsets (8,9). Participants are administered only a subset of the items that comprise the item bank(s). Through scale linking, all subsets can be put on a common measurement metric. Another example where test linking may be needed is when a researcher wants to create one instrument by pooling health outcome items from two studies where there is overlap in the items administered across the two studies.

This study examines several methods for linking pain-related items across two patient surveys. A survey conducted by the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) committee (10) and a survey conducted by the Center on Outcomes, Research and Education (CORE), Evanston Northwestern Healthcare and Northwestern University (6). The IMMPACT Survey included 959 patients with chronic pain conditions and the CORE Survey included 400 patients with various cancers. Because the two surveys included common pain interference and intensity items, we were able to combine the two datasets to conduct the linking study. The obvious benefits of combining the datasets are that the subject sample size and number of items is larger and the domain coverage of items is broader.

The process of creating a common scale is usually referred to as scale linking in IRT analyses. The application of IRT starts with item calibration—the process of fitting item response data to an IRT model and obtaining item parameters for each item (11,12). There are different ways of linking sets of items using IRT depending on the data collection designs. Two set of items can be linked through common subjects, common items, or both. The item banks for the IMMPACT and CORE surveys included common items but the different samples constituted non-equivalent groups. There are two ways to achieve linking in such a situation, that is, a common item non-equivalent group design. One is to link by “simultaneous (concurrent) IRT calibration” (13). The other is by “separated IRT calibration” in which each sample is calibrated separately and then linked through “scale transformation” (13). For this study, we compared these two scale linking methods.

METHODS

Data Sources

The CORE and IMMPACT survey item sets included items measuring three pain domains: pain intensity, pain quality and pain interference.

CORE Survey—The CORE pain item bank data were collected in a cross-sectional survey of 400 cancer patients. Table 1 summarizes the demographic and clinical characteristics of this sample. The original CORE pain item bank contained 61 items, but this item bank was reduced to a final data set of 43 items (6). Items without good psychometric characteristics were excluded from the final pain item bank. The development of this bank is described elsewhere (6).

IMMPACT Survey—The IMMPACT Survey was a cross sectional internet-based survey of individuals with chronic pain recruited through the American Chronic Pain Association website (10). The purpose of the survey was to gather data on the importance and relevance of different domains impacted by chronic pain and to measure pain intensity, quality and interference, and functional and psychological well-being outcomes in people with chronic pain. The sample consisted of 959 respondents who had one or more chronic pain conditions (see Table 1). The survey included a Main Survey section with a total of 42 pain-related items, and the Pain

Module that included 36 items from the CORE pain item bank. All 959 patients took the Main Survey; only a subset of 148 also took the Pain Module.

Common and Unique Items—Pooling of the IMMPACT Main Survey, the Pain Module, and the CORE pain item bank yields 29 pain interference, 10 pain intensity, and 38 pain quality items. The three item sets shared eight common items: one pain intensity item and 7 pain inference items. That left the IMMPACT Main Survey with 10 unique pain inference items and 24 unique pain quality items. The IMMPACT Pain Module did not have a unique item to itself. All its items were common with the CORE pain item bank. These included 12 pain interference items, 10 pain intensity items, and 14 pain quality items. Figure 1 shows the compositions of the unique and shared common items of the three item sets for the three pain domains.

Since there was no common item shared between the IMMPACT Main survey with Pain Module and CORE item bank for the pain quality domain, they could not be linked using the separated calibration approach. Thus, pain quality domain was excluded from this study. The two pain domains, pain interference and pain intensity, were analyzed separately as they each was assumed to be unidimensional.

Statistical Analyses

Samejima's Graded Response Model (14) as implemented in MULTILOG (15) was used to calibrate the items. The pain interference and pain intensity domains were analyzed separately. Higher response categories indicate worse pain or severe pain-related problems. The extreme response categories for some of the items were excluded from the IRT calibration because no patients endorsed these categories. The maximum number of response categories allowed by MULTILOG is ten. Therefore, for the 11-category items, the two highest response categories were combined into one category.

Simultaneous IRT Calibration (with multiple groups)—Simultaneous IRT calibration represents the easiest and most direct way to link two sets of items when there are shared common items. In this method, the item parameters of the common items are constrained to be equal across the samples. In essence, each common item is treated as one single item administered to different groups of samples. When the simultaneous calibration is completed, all items are automatically on the same measurement scale (16).

The mean and standard deviation of the population distribution of latent trait (θ) are also estimated during the calibration. In the linking, when the multiple samples are considered equivalent, the single group IRT calibration is conducted. In this study, the groups are non-equivalent as they are from different populations (e.g., cancer and chronic pain); therefore, multiple group IRT calibration is used, in which, one of the groups will have mean fixed to zero, and other groups' means will be estimated (13).

For multiple groups simultaneous calibration, we included all items and all subjects in a single calibration run. The sample from the CORE survey was treated as the reference group with its population mean fixed to zero. The IMMPACT sample mean was estimated. Linking using simultaneous IRT calibration with multiple groups was conducted separately for the pain interference and pain intensity domains.

Separated IRT Calibration—The second linking approach that we used was to conduct separate item calibrations followed by scale transformation. One IRT calibration was conducted for the combined IMMPACT Main Survey and IMMPACT Pain Module. The second IRT calibration was conducted for the CORE pain item bank.

This method results in separate sets of item parameters for each data set that are not on the same measurement scale because the samples of the two studies are considered non-equivalent. To link the separate sets of item parameters, "scale transformation" is performed on the common items. (13,16). Scale "transformation constants" are calculated and used to place item parameters on a common mathematical metric. This method takes advantage of the parameter invariance of IRT models. When the assumptions of the model are met, item parameters calibrated in different samples are the same *within a linear transformation*. Because the measurement scale in IRT is unobservable, there is no natural origin for the scale and it is arbitrarily given a mean of 0 and standard deviation of 1. The linear transformation of one metric to the other can be expressed as,

$$\theta^{*}=A\theta+B, \tag{1}$$

where, *A* is the slope, and *B* is the intercept. The new item parameters then can be transformed to the target metric using the same coefficients as follows,

$$a_j^{*}=\frac{a_j}{A}, \tag{2}$$

$$b_j^{*}=Ab_j+B, \tag{3}$$

Where a_j and a^*_j are the slope parameters, and b_j and b^*_j are the location or threshold parameters.

In this study, we transformed the item parameters from the IMMPACT item calibration to be the same as the CORE item calibration, using the CORE calibration as the common metric. There are different scale transformation methods (13,16). The most commonly used are the mean/mean, mean/sigma, and test characteristic curve methods. We used four transformation methods to obtain the transformation constants by using the computer program STUIRT (17). STUIRT is one of many equating and scale transformations computer programs (available at www.uiowa.edu/~casma). These programs can be used in conjunction (13). In this study, linking using separated IRT calibration was conducted separately for pain interference and pain intensity.

Evaluation of the Two Approaches—To evaluate how well the two calibration approaches work, we first examine whether the calibration converged. There is no closed solution for the item parameters in IRT. MULTILOG uses marginal maximum likelihood (18) method to estimate the item parameters. The calibration is said to converge when item parameter estimates are obtained. Usually, non-convergence indicates problem in fitting of the data to the IRT models.

Non-convergence is evident when there is extreme item parameter. The underlying scale of the IRT model is conventionally set a mean equal 0 and standard deviation of 1. Therefore, any threshold parameter that exceeds ± 6.0 is considered extreme since it is six standard deviations away from the mean. Other evidence of non-convergence related to the order of the threshold parameters. The graded response model assumes that the response options are monotonically ordered, subsequently, the threshold parameters are ordered. Non-ordered threshold parameters, in graded response model, are indication of non-convergence or problematic model fitting. In this study, we examine the value of the item parameters and the order of the threshold parameters to evaluate how well the two calibration approaches work.

RESULTS

Simultaneous IRT Calibration

The results of the simultaneous IRT calibration are summarized in Table 2 for the pain interference domain. There were 29 pooled items and 1,364 subjects for the pain interference domain. The slope parameters were all reasonable large from 1.84 to 3.74, and all the threshold parameters were monotonically increasing. The mean θ for the CORE sample was fixed at zero, and the mean θ for the IMMPACT sample was 2.22. The item characteristic curves suggest that 10 response categories may be too many. Some of the middle response categories overlap with each other. Subjects with similar degrees of interference (based on estimated θ) were equally likely to endorse more than one overlapping response categories. On average, the IMMPACT sample reported higher levels of pain interference. This was expected since all CORE subjects were cancer patients and not all of them experienced significant pain (6). Whereas, the IMMPACT sample was comprised of patients who had chronic pain from rheumatoid arthritis, osteoarthritis, lower back pain, fibromyalgia, diabetic neuropathy, and other neuropathic pain.

There were 10 pooled items and 1,364 subjects for the pain intensity domain. There was only one item that was answered by all subjects (Pain on average in past week). Only the patients administered the IMMPACT Pain Module survey and the CORE patients answered the other 9 items. The slope parameters were very high, from 2.78 to 4.73, except for the item “I have minor aches and pains” which had a slope of 0.90. This item was one of the unfolding model items (19), where the patients with no pain or with severe pain would both answer “None of the time.” This item does not fit the graded response model. Mean θ of the IMMPACT sample was 2.41, higher than the CORE sample. The item characteristic curves show that, again, the items with 10 response categories can be reduced to fewer response categories with minimal loss of information.

These results indicated that linking was applied successfully using the simultaneous calibration approach. The calibrations converged for both domains, and there were no significant problems observed for either the item parameters or item characteristic curves.

Separated IRT Calibration

The pain interference domain provided a good test of the separated calibration approach. There were 7 common items between the IMMPACT Main survey and CORE items with 959 and 400 subjects in each study, respectively. In addition, the IMMPACT Pain Module and CORE surveys shared 12 common items and had 148 and 400 subjects, respectively. We first estimated item parameters for the pain interference items for the IMMPACT sample. These included 959 subjects who answered the Main survey and 148 subjects who also answered the Pain Module survey. There were 29 interference items in this data set, 17 items from the Main survey and 12 from the Pain Module. Item calibration also was completed for the CORE pain interference items. There were 19 pain interference items completed by 400 CORE patients. There were 19 common items between these two data sets (7 from the Main survey and 12 from the Pain Module survey). Different from the simultaneous calibration where the mean θ was fixed for one of the samples and estimated for the other, for the each separated calibration the mean θ was always fixed at zero. That is, the mean θ was fixed at zero for both the IMMPACT and CORE samples when their parameters were being calibrated in separate runs. The slope parameters ranged from 1.20 to 2.99 for the items in the IMMPACT sample, and from 2.49 to 5.96 for the CORE sample. The slopes were higher for the IMMPACT items were due to higher correlation between the items, and the more homogenous responses among the CORE survey patients. The threshold parameters for the IMMPACT items ranged from −5.56 to 0.66, and

ranged from −0.11 to 1.92 for the CORE items. The ranges of the threshold parameters cannot be directly compared between the IMMPACT and CORE items at this time.

We also conducted separate calibration for the items within the pain intensity domain. For the IMMPACT sample, only one item was answered by all IMMPACT patients, and the other 9 items were answered by the 148 Pain Module survey patients. Because of the small sample size and since many of the response categories had zero counts the calibration was not stable. This was evidenced in the item parameters estimates. One item had a slope of 0.13. Thresholds for some items were as high as 16.6 and even 37.0, and were not monotonically increasing. Based on this result, we chose not to attempt to link items of the intensity domain based on the separated calibration approach.

After the item parameters were estimated, the item parameters from the IMMPACT calibration were transformed to be on the same metric as the CORE calibration. Using the computer program, STUIRT, we obtained transformation constants, A and B (see equation 1), using four different scaling methods. The results are shown in Table 3 for the pain interference domain. We found that the transformation constants were very close among the four transformation methods. However, the transformation constants obtained using the Haebara and Stocking-Lord approaches were more similar. We used these constants to transform the item parameters of the non-common items, and the items from the two separated calibration onto a common metric. The transformed item parameters using the Stocking-Lord transformation method are shown in Table 4.

Simultaneous vs. Separate Calibration

Next, we compared the simultaneous approach and the separate approach for the pain interference domain. Note that although we were comparing the item parameters of the same items between the two approaches, the comparison was not direct. The IMMPACT and CORE items were on the same scale as a result of the simultaneous calibration approach. The IMMPACT and CORE items were also on the same scale as a result of the separated calibration approach. However, the measurement scales for the two approaches were not on the same metric scale. The two sets of “item bank” calibrations differed by a linear transformation. Therefore, another scale transformation was necessary to make the two banks on the same mathematical metric in order to compare the simultaneous calibration and separated calibration approaches.

We correlated the item parameters of the two approaches. The correlation between the slope parameters was 0.923 (Figure 2). The correlations between the threshold parameters ranged between 0.911 and 0.992, except the first threshold parameter (b1, the threshold parameter indicates the point where the probability passes 50% for endorsing the second or higher responses). Figure 3, the scatter plots of the item threshold parameters, shows the straight line relationship between linking methods, except an outlier for b1. The b1 parameter for the item “Pain interference with your daily activities,” a 4-point scale item, was very small because there were zero observations for the lowest response option. Without this outlier, the two sets of item parameters were very similar. We also examined the descriptive statistics and correlations of the IRT scores of the two approaches. Table 5 shows these statistics including the correlation of between IRT scores for the pain interference domain obtained with the two approaches. The two scales differed by a factor of 0.784, the ratio of the standard deviations for the IRT scores of the CORE sample (1.047/0.821). The correlations between the IRT scores of the two approaches (see Table 5), were as high as 0.999 for the IMMPACT and CORE samples, and for overall. Thus the two calibration approaches produced very similar item characteristics.

DISCUSSION

In this study, we demonstrated how pain items from separate surveys can be linked to the same measurement scale to form a single item bank when there were some shared common items. For the datasets that were used in this study, we obtained better linking results using a simultaneous calibration when the sample size was relatively small. This is demonstrated by comparing the results for the pain intensity domain, where the separated calibration did not converge. The common items for the pain intensity domain were answered by only 148 subjects. When the items were calibrated separately, we obtained extreme item parameter estimates (threshold parameters estimates as high as 16.6 and 37.0) and some of the threshold parameters were not monotonically ordered correctly..

These findings correspond to simulation studies that have found the simultaneous calibration method produced more accurate results than separated calibration when the IRT model fits the data (13,16). Cook et al. (20) also found that sample sizes of 300 or more subjects were necessary for linking health outcome measures. When sample size is small, the separated calibration approach is not advisable based on the finding of non-convergence results.

For the pain interference domain, we had sufficient sample size for linking scales. In this case, we obtained similar results when linking by simultaneous calibration and using separated calibrations (as shown in Table 5). Therefore, results of this study suggest that simultaneous calibration is preferable when linking sets of item from two surveys, particularly with smaller sample sizes and fewer common items. However, with sufficient sample size and enough common items, separate calibration approaches can yield similar results to the simultaneous calibration methods.

There is no fixed rule regarding the number of common items across two samples. Angoff (21) suggested that common items should contain the larger of 20 items or 20% of the total number of items. Other investigators (22) recommend from 5 to 10 items to form the link between samples. In general, a larger number of common items will result in more precise and stable item calibrations in the bank. In our comparison, even when there were few (< 5) common items, we found that the simultaneous IRT calibration method produced converged item parameter estimates across independent samples of pain intensity and pain interference items. The results of this study suggest that simultaneous calibration methods should be the recommended approach for linking sets of health domain items across two or more studies.

Linking scales are common in education owing to the need to maintain a consistent and comparable test scores across many test forms and test administrations. Reports of empirical studies of linking scales in education are numerous (13,18,23,24,25,26,27,28). Linking scales across samples are less common in health outcome research. While the methods and experiences maybe borrowed from education, there is uniqueness in health outcome content and items which require further exploration on the appropriateness of different linking methods. Testing in education is routine and longitudinal, and the population is relatively homogenous. The sample sizes are often large, and the number of test items is almost unlimited. In health outcome studies, the population is more heterogeneous, the sample sizes are frequently small, and the number of items may be limited. However, the small sample sizes and limited item banks in health outcome studies make linking scales and items more important. Through linking, small studies can be pooled to create larger sample sizes and to create larger and more diverse item banks. More empirical research is needed to understand and further explicate how best to link health outcome instruments and item banks. Further guidance is needed as to the best methods and to provide empirically based recommendations .for accomplishing linking between different health outcome studies.

In conclusion, patient reported health outcome surveys are often limited in sample sizes and the number of items owing to the difficulty of recruitment and the burden to the patients. Thus the surveys' content coverage may be restricted and statistical power limited. Using IRT methodology, surveys data can be pooled to lend strength to each other to expand the content coverage and to increase the sample sizes. This in turn increases the statistical power of the pooled study and provides for a more comprehensive item bank. In addition, linking the two studies allows scores to be placed on the same metric which enables improved comparison of findings between studies even if the same set of items were not included in both studies.

REFERENCES

1. McHorney CA. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med 1997;127:743–750. [PubMed: 9382391]
2. Revicki DA, Cella DF. Health status assessment for the 21st century: item response theory, item banking, and computer adaptive testing. Qual Life Res 1997;6:595–600. [PubMed: 9330558]
3. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38(9 Suppl):II28–II42. [PubMed: 10982088]
4. McHorney CA, Cohen AS. Equating health status measures with item response theory: illustration with functional status items. Med Care 2000;38(9 Suppl):II43–II59. [PubMed: 10982089]
5. Lai JS, Cella D, Chang CH, et al. Item banking to improve, shorten, and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-fatigue scale. Qual Life Res 2003;12:485–501. [PubMed: 13677494]
6. Lai JS, Dineen K, Reeve BB, et al. An item response theory-based pain item bank can enhance measurement precision. J Pain Symptom Manage 2005;30:278–288. [PubMed: 16183012]
7. Hays RD, Liu H, Spritzer K, Cella D. Item response theory analyses of physical functioning items in the Medical Outocmes Study. Med Care 2007;45:S12–S21. [PubMed: 17443114]
8. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B, Rose M. The patient-reported outcome measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Med Care 2007;45(5 Suppl 1):S3–S11. [PubMed: 17443116]
9. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Honghu L, Gershon R, Reise SP, Lai J, Cella D. Psychometric evaluation and calibration of health-related quality of life items banks: plans for the patient-reported outcome measurement information system (PROMIS). Med Care 2007;45(5 Suppl 1):S22–S31. [PubMed: 17443115]
10. Turk DC, Dworkin RH, Revicki DA, Harding G, Burke LB, Cella D, Cleeland CS, Cowan P, Farrar JT, Hertz S, Max MB, Rappaport BA. Identifying important outcome domains for chronic pain clinical trials: an IMMPACT survey of people with pain. Pain. 2007Epub ahead of print.
11. Reeve, BR.; Fayers, P. Applying item response theory modeling for evaluating questionnaire item and scale properties. In: Fayers, P.; Hays, R., editors. Assessing Quality of Life in Clinical Trials. Vol. 2nd ed.. New York: Oxford University Press; 2005. p. 55-73.
12. Van der Linden, WJ.; Hambleton, RK., editors. Handbook of Modern Item Response Theory. New York, NY: Springer-Velag; 1997.
13. Kolen, MJ.; Brennan, RL. Test Equating, Scaling, and Linking: Methods and Practices. New York: Springer-Verlag; 2004.
14. Samejima F. Estimation of latent trait ability using a response pattern of graded scores. Psychometrika Monogr 1969;34(4 Pt 2)whole No. 17.
15. Thissen, D. MULTILOG user's giude. Lincolnwood, IL: Scientific Software International; 2002.
16. Cohen AS, Kim SH. An investigation of linking methods under the graded response model. Appl Psych Measure 1998;22:116–130.
17. Kim, S.; Kolen, MJ. STUIRT: A computer program for scale transformation under unidimensional item response theory models, version 1.0. Iowa City: Iowa Testing Programs, University of Iowa; 2004.

18. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. Psychometrika 1981;46:443–459.

19. Roberts JS, Donoghue JR, Laughlin JE. A general item response theory model for unfolding unidimensional polytomous responses. Applied Psychological Measurement 2000;24(1):3–32.

20. Cook KF, Taylor PW, Dodd BG, Teal CR, McHorney CA. Evidence-based practice for equating health status items: sample size and IRT model. J Appl Measurement 2007;8:175–189.

21. Angoff, WH. Scales, norming, and equivalent scores. In: Thorndike, RL., editor. Educational Measurement. Vol. 2nd ed.. Washington, DC: American Council on Education; 1971. p. 508-600.

22. Wright BD, Bell SR. Item banks: What, why, how. Journal of Educational Measurement 1984;21(4): 331–345.

23. Baker FB. Equating tests under the graded response model. Applied Psychological Measurement 1993;16:87–96.

24. Baker FB, Al-karni A. A comparison of two procedures for computing IRT equating coefficients. Journal of Educational Measurement 1991;28:147–162.

25. Cook LL, Peterson NS. Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement 1987;11:225–244.

26. Hung, P.; Wu, Y.; Chen, Y. IRT item parameter linking: Relevant issues for the purpose of item banking. Paper presented as the International Academic Symposium on Psychological Measurement; Tainan, Taiwan. 1991.

27. Kim SH, Cohen AS. A comparison of linking and concurrent calibration under item response theory. Applied Psychological Measurement 1998;22(2):131–143.

28. Tate RL. A cautionary note on IRT-based linking of tests with polytomous items. Journal of Educational Measurement 1999;36(4):336–346.

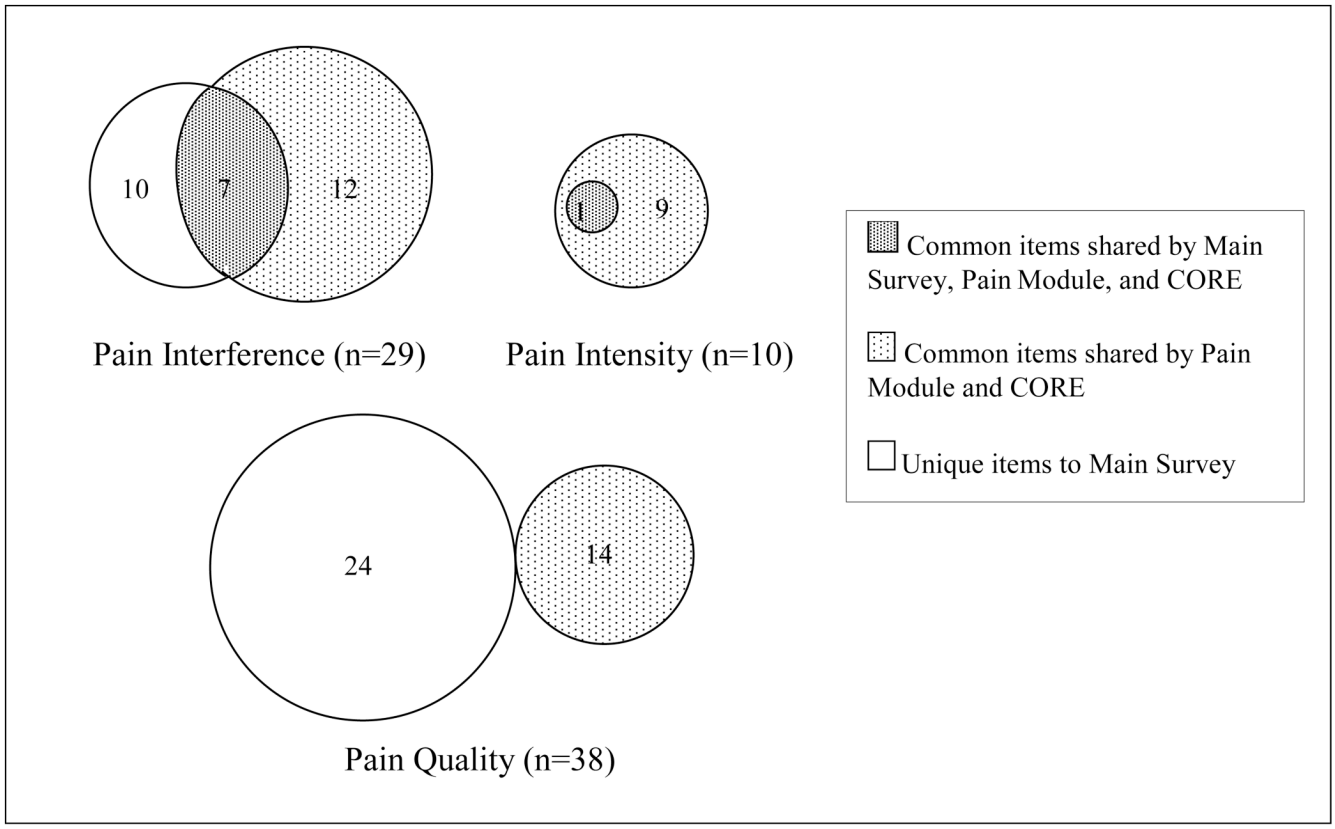

Figure 1.
Number of Unique and Common Items in IMMPACT Main Survey, IMMPACT Pain Module, and NU-CORE Item Bank by Pain Domains

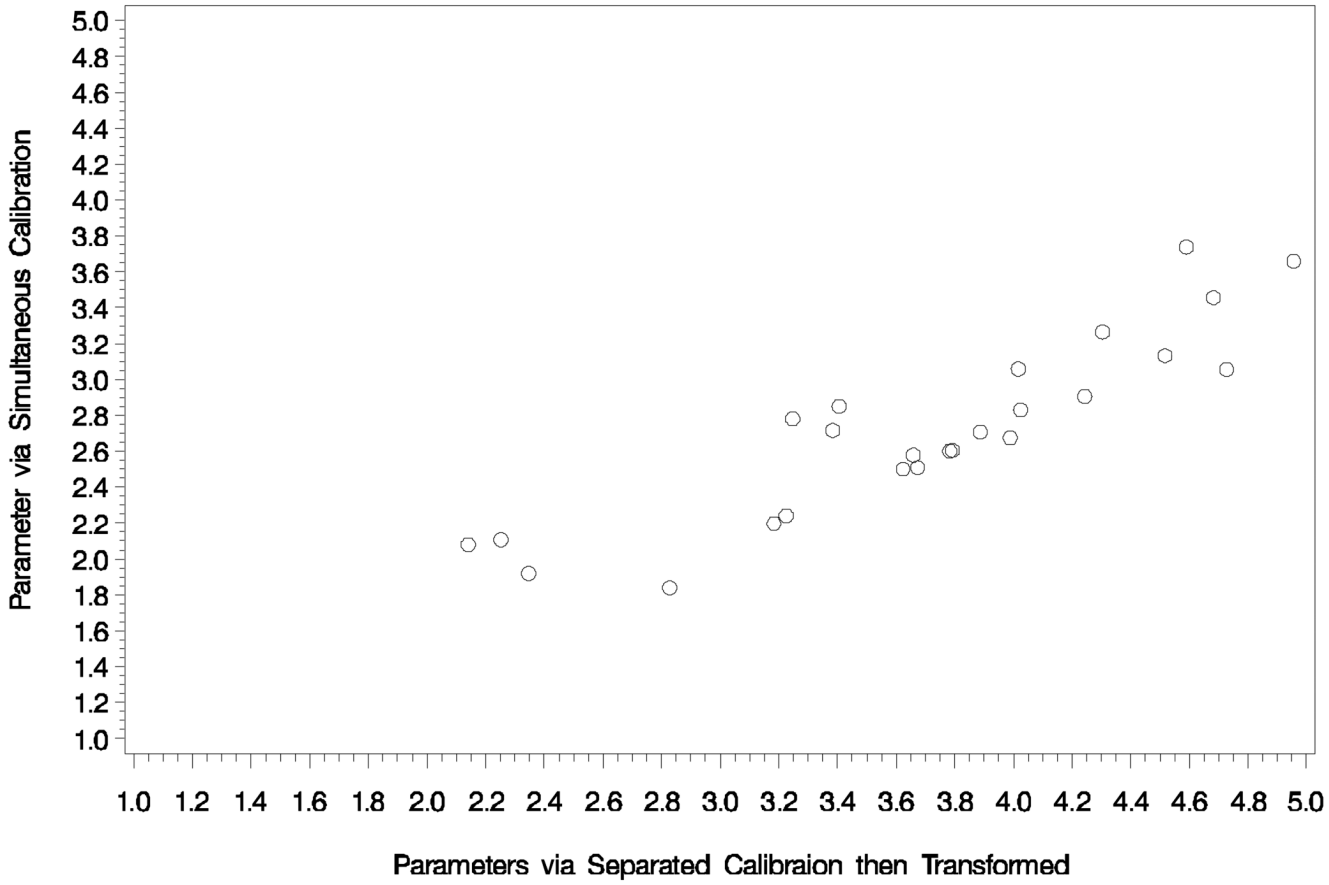

Figure 2.
Comparisons of Slope Parameters of Pain Interference Item between Simultaneous Calibration vs. Separated Calibration with Stocking-Lord Transformation

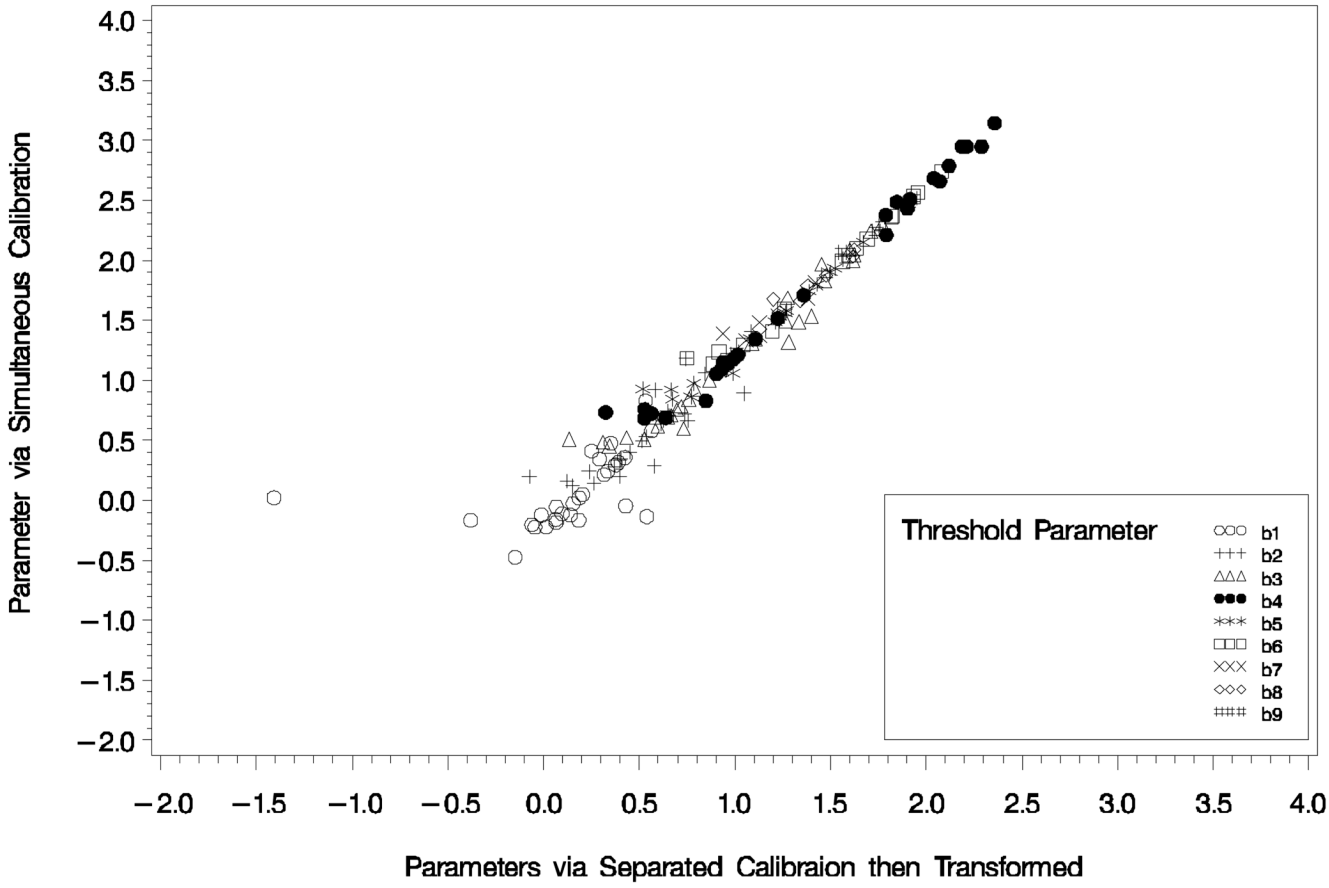

Figure 3.
Comparisons of Threshold Parameters of Pain Interference Item between Simultaneous Calibration vs. Separated Calibration with Stocking-Lord Transformation

Table 1

Demographic Characteristics of IMMPACT Survey and CORE Survey Participants

		CORE Survey (n=400)	
	IMMPACT Survey (n = 959)	**General Cancer (n=202)**	**Prostate Cancer (n=198)**
Age(years), mean (SD)	45.6 (11.6)	57.7 (13.7)	71.8 (8.1)
Gender			
Male, %	28.0	29.6	100.0
Female, %	72.0	70.4	
Race			
White, %	92.4	83.2	56.6
African-American. %	2.1	7.4	37.4
Other, %	5.5	9.5	6.0
College Graduate, %	39.7	57.5	33.3

Table 2

Item Parameter Estimates of Pooled Pain Interference Item by Simultaneous Calibration

	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5	Threshold 6	Threshold 7	Threshold 8	Threshold 9
Unique items: IMMPACT Main Survey										
Pain interferes day-to-day activity 0–6	3.69	−0.22	0.14	0.62	1.22	1.80	2.37			
Pain interferes ability to work 0–6	2.60	0.30	0.50	0.76	1.18	1.58	2.00			
Pain interferes sat/enjoy in soc/rec acts 0–6	3.53	0.05	0.34	0.69	1.10	1.57	2.18			
Pain interferes ability to take part in rec. 0–6	3.49	0.02	0.33	0.71	1.05	1.49	2.10			
Pain interferes sat/enjoy in family acts 0–6	3.13	0.24	0.70	1.01	1.35	1.94	2.53			
Pain interferes relationship with spouse/ fam 0–6	2.24	0.21	0.64	1.09	1.52	1.91	2.57			
Pain interferes sat/enjoy from work 0–6	2.50	0.32	0.53	0.85	1.14	1.53	2.04			
Pain interferes ability to do house/ chores 0–6	2.91	−0.11	0.40	0.78	1.21	1.74	2.36			
Pain interferes	2.71	0.69	1.12	1.35	1.71	2.15	2.75			

	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5	Threshold 6	Threshold 7	Threshold 8	Threshold 9
friendship with others 0–6										
Past 4 weeks pain interfered normal work 0–4	2.51	−0.47	0.28	1.31	2.51					
Common items: IMMPACT Main and Pain Module Surveys, and CORE Pain Item Bank										
Pain interferes ability to concentrate 0–4[1]	2.83	0.35	1.06	2.00	2.79					
Pain interferes ability to pay attention 0–4[1]	2.61	0.58	1.24	2.31	2.95					
Pain interferes ability to think clearly 0–4[1]	2.58	0.83	1.40	2.25	3.15					
Pain makes me depressed 0–4[1]	1.92	0.41	1.23	2.09	2.95					
Pain interferes my family life 0–4[1]	2.72	0.48	1.18	1.97	2.69					
Past week pain interfered: General	3.46	−0.16	0.12	0.45	0.73	0.98	1.30	1.53	1.88	2.29

J Pain Symptom Manage. Author manuscript; available in PMC 2009 October 14.

	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5	Threshold 6	Threshold 7	Threshold 8	Threshold 9
activity 0–10[2]										
Past week pain interfered: Mood 0–10[2]	2.85	−0.20	0.16	0.49	0.76	0.92	1.24	1.49	1.79	2.25
Past week pain interfered: Walking ability 0–10[2]	1.84	−0.05	0.29	0.60	0.83	1.06	1.41	1.68	2.04	2.48
Past week pain interfered: Normal work 0–10[2]	3.60	−0.16	0.20	0.51	0.69	0.87	1.16	1.37	1.66	2.01
Past week pain interfered: People relationship 0–10[2]	3.06	0.36	0.66	0.93	1.15	1.35	1.60	1.81	2.09	2.52
Past week pain interfered: Sleep 0–10[2]	2.08	−0.17	0.20	0.51	0.73	0.93	1.19	1.39	1.68	2.07
Past week pain interfered: Enjoyment of life 0–10[2]	3.74	−0.02	0.24	0.53	0.68	0.85	1.14	1.33	1.57	1.89
During past week pain interferes daily acts 1–4[1]	3.66	0.02	1.27	2.05						
Past 4 weeks pain interferes: mood 1–5[1]	2.78	−0.05	1.09	1.83	2.95					
Past 4 weeks pain interferes:	2.20	−0.13	0.89	1.54	2.66					

	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5	Threshold 6	Threshold 7	Threshold 8	Threshold 9
walk or move about 1–5[1]										
Past 4 weeks pain interferes: sleep 1–5[1]	2.11	−0.12	0.92	1.69	2.49					
Past 4 weeks pain interferes: normal work 1–5[1]	3.06	−0.18	0.72	1.49	2.43					
Past 4 weeks pain interferes: recreation 1–5[1]	2.68	−0.22	0.66	1.32	2.21					
Past 4 weeks pain interferes: enjoyment life 1–5[1]	3.27	−0.12	0.74	1.50	2.38					

[1] Common items appear both in IMMPACT Pain Module Survey and CORE Pain Item Bank.

[2] Common items appear in IMMPACT Main Survey, IMMPACT Pain Module Survey, and CORE Pain Item Bank.

Table 3

Transformation Constants for Pain Interference Domain

	Pain Interference Domain	
	A[1]	B[1]
Mean/Mean	0.561	1.757
Mean/Sigma	0.575	1.771
Haebara	0.605	1.776
Stocking-Lord	0.607	1.762

[1] A and B are the transformation constants in Equation 1

Table 4

Item Parameter Estimates of IMMPACT Pain Interference Item by Separated Calibration Then Transformed to CORE Pain Item Bank Scale Using Stocking-Lord Method via Common Items

	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5	Threshold 6	Threshold 7	Threshold 8	Threshold 9
Unique items: IMMPACT Main Survey										
Pain interferes day-to-day activity 0–6	5.34	0.02	0.27	0.60	1.02	1.43	1.82			
Pain interferes ability to work 0–6	3.78	0.38	0.52	0.70	0.99	1.27	1.56			
Pain interferes sat/enjoy in soc/rec acts 0–6	5.12	0.20	0.41	0.65	0.93	1.26	1.69			
Pain interferes ability to take part in rec. 0–6	5.09	0.19	0.40	0.67	0.90	1.21	1.64			
Pain interferes sat/enjoy in family acts 0–6	4.52	0.34	0.66	0.87	1.11	1.52	1.94			
Pain interferes relationship with spouse/ fam 0–6	3.23	0.32	0.61	0.93	1.23	1.50	1.96			
Pain interferes sat/enjoy from work 0–6	3.62	0.39	0.54	0.76	0.96	1.24	1.60			
Pain interferes ability to do house/ chores 0–6	4.24	0.10	0.45	0.72	1.02	1.39	1.82			

	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5	Threshold 6	Threshold 7	Threshold 8	Threshold 9
Pain interferes friendship with others 0–6	3.89	0.65	0.94	1.11	1.36	1.67	2.08			
Past 4 weeks pain interfered normal work 0–4	3.67	−0.15	0.37	1.09	1.92					
Common items: IMMPACT Main and Pain Module Surveys, and CORE Pain Item Bank										
Pain interferes ability to concentrate 0–4[1]	4.02	0.29	0.85	1.62	2.12					
Pain interferes ability to pay attention 0–4[1]	3.79	0.57	1.02	1.77	2.19					
Pain interferes ability to think clearly 0–4[1]	3.66	0.54	1.09	1.71	2.36					
Pain makes me depressed 0–4[1]	2.35	0.25	0.98	1.60	2.21					
Pain interferes my family life 0–4[1]	3.38	0.35	0.75	1.46	2.04					
Past week pain interfered:	4.68	0.07	0.15	0.34	0.57	0.79	1.04	1.23	1.48	1.77

J Pain Symptom Manage. Author manuscript; available in PMC 2009 October 14.

	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5	Threshold 6	Threshold 7	Threshold 8	Threshold 9
General activity 0–10[2]										
Past week pain interfered: Mood 0–10[2]	3.41	−0.06	0.12	0.31	0.53	0.67	0.92	1.13	1.38	1.74
Past week pain interfered: Walking ability 0–10[2]	2.83	0.43	0.58	0.73	0.85	0.99	1.20	1.39	1.62	1.91
Past week pain interfered: Normal work 0–10[2]	5.13	0.19	0.40	0.53	0.64	0.77	0.96	1.13	1.34	1.58
Past week pain interfered: People relationship 0–10[2]	4.02	0.43	0.62	0.79	0.94	1.08	1.26	1.42	1.62	1.94
Past week pain interfered: Sleep 0–10[2]	2.14	−0.38	−0.07	0.14	0.33	0.52	0.75	0.94	1.20	1.56
Past week pain interfered: Enjoyment of life 0–10[2]	4.59	0.16	0.24	0.44	0.53	0.68	0.89	1.06	1.23	1.48
During past week pain interferes daily acts 1–4[1]	4.96	−1.41	1.01	1.62						
Past 4 weeks pain interferes: mood 1–5[1]	3.25	0.07	0.95	1.47	2.29					
Past 4 weeks pain	3.18	0.54	1.05	1.40	2.07					

	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5	Threshold 6	Threshold 7	Threshold 8	Threshold 9
interferes: walk or move about 1–5[1]										
Past 4 weeks pain interferes: sleep 1–5[1]	2.25	0.14	0.58	1.27	1.85					
Past 4 weeks pain interferes: normal work 1–5[1]	4.73	0.07	0.74	1.33	1.90					
Past 4 weeks pain interferes: recreation 1–5[1]	3.99	−0.04	0.75	1.28	1.79					
Past 4 weeks pain interferes: enjoyment life 1–5[1]	4.30	−0.01	0.65	1.27	1.79					

[1] Common items appear both in IMMPACT Pain Module Survey and CORE Pain Item Bank.

[2] Common items appear in IMMPACT Main Survey, IMMPACT Pain Module Survey, and CORE Pain Item Bank.

Table 5

Descriptive Statistics for IRT Theta Scores and Their Correlation from Different Calibration Approaches and After Linking

					Separated Calibration		
					IMMPACT	CORE	
	N	Mean (SD)	Range	Median	Sample Pearson r	Sample Pearson r	Overall
Theta score from simultaneous calibration							
IMMPACT Sample	962	2.20 (0.738)	−0.05–4.00	2.22	0.999***		
CORE Sample	400	−0.09 (1.047)	−1.35–3.32	−0.18		0.999***	
Overall	1364	1.52 (1.341)	−1.35–4.00	1.86			0.999***
Theta score from separated calibration then scaletransformed by Stocking-Lord method							
IMMPACT Sample	962	1.70 (0.505)	0.11–2.87	1.71			
CORE Sample	400	−0.02 (0.821)	−1.04–2.70	−0.05			
Overall	1364	1.19 (0.994)	−1.04–2.87	1.47			
Overall (adjusted with factor 0.784=1.047/0.821)	1364	1.52 (1.268)	−1.33–3.66	1.88			

*** p<0.001

** p<0.01

* p<0.05