Qual Life Res (2013) 22:491–499
DOI 10.1007/s11136-012-0179-6

Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement

Yi Zheng · Chih-Hung Chang · Hua-Hua Chang

Accepted: 4 April 2012 / Published online: 27 April 2012

Abstract

Purpose Most multidimensional patient-reported outcomes (PRO) measures are lengthy to complete. Computerized adaptive testing (CAT) that selects the most informative items can potentially reduce respondent burden without sacrificing measurement accuracy. The commonly used maximum Fisher information item selection method has been reported to lead to highly unbalanced item bank usage and potentially imprecise trait estimation. This study employs the content-balancing strategy in a bifactor-modeled CAT item selection and examines its impact on measurement accuracy and item bank usage.

Methods Item responses from a population-based SF-36 survey were first calibrated using the bifactor graded response model. Four post hoc CATs using items and responses from the SF-36 data set were then created. The content-balancing strategy was adopted in the item selection procedure of the bifactor-modeled CAT. The measurement accuracy and usage of items of the CAT were compared between the tests with and without the content-balancing strategy.

Results The results indicate that the CAT implemented with the content-balancing strategy offers a better overall measurement accuracy of both the general health status and the two health domains (physical and mental) of the SF-36.

Conclusions The content-balancing strategy helps the CAT–PRO to balance the selection of items and achieve improved measurement accuracy. Its implementation in real-time CAT administration to measure multidimensional PRO traits merits further studies.

Keywords Patient-reported outcomes (PRO) · Computerized adaptive testing · Content-balancing · Bifactor model · Item selection · SF-36

Y. Zheng (✉)
University of Illinois at Urbana-Champaign,
1310 S. Sixth St Rm 210, Champaign, IL 61820, USA
e-mail: yizheng1@illinois.edu

C.-H. Chang
Buehler Center on Aging, Health & Society, Northwestern University Feinberg School of Medicine, 750 N. Lake Shore Drive, Suite 601, Chicago, IL 60611, USA
e-mail: chchang@northwestern.edu

C.-H. Chang
Graduate Institute of Biostatistics, China Medical University, Taichung, Taiwan, ROC

H.-H. Chang
University of Illinois at Urbana-Champaign,
1310 S. Sixth St Rm 236, Champaign, IL 61820, USA
e-mail: hhchang@illinois.edu

Introduction

Patient-reported outcomes (PRO) measurement has gained its popularity and acceptance in health outcome research and clinical practices in recent years. Patient self-reported health outcome data, besides clinical and laboratory data, are now being used more frequently to assess the medical endpoints of interest (e.g., symptom reduction and functioning improvement) and to facilitate clinical decision-making and disease management. Integrating PRO data into clinical practice may potentially improve the quality of patient care and reduce the health care costs by: (1) early detecting a patient's clinical issues and care needs; (2) tailoring therapeutic interventions or treatments; and (3) improving a patient's health-related quality of life and their satisfaction with care [1]. Because rapid, accurate, yet comprehensive assessment is an important first step to optimal patient care, numerous PRO assessment tools have been developed to measure various aspects of generic and disease-specific health outcomes, such as fatigue, pain, depression, and physical functions [2]. However, despite the demand and potential benefit, the collection and use of PRO data is still in its infancy and has not reached the desired level of acceptance and utilization in clinical settings. More research utilizing advanced and well-accepted psychometric methods to improve the PRO measurement tools is still needed.

As Chang [1] noted, "in order to put PRO into routine operation during clinical encounters in busy clinics, the measures must be brief and easy for the patient to complete, impose little or no burden on clinic staff to collect and analyze, and provide critical clinical information during the clinical encounter." To achieve this goal, it requires a more efficient and effective method. Item response theory (IRT) [3] and computerized adaptive testing (CAT) [4] together offer promises to address these measurement challenges.

Based on the mathematical foundation of IRT, the CAT version of the PRO assessments can tailor questionnaires and select the most suitable questions for each patient. This strategy shortens the questionnaire administration without sacrificing measurement accuracy. In fact, IRT and CAT have already been well developed and widely applied to measure latent traits (e.g., ability and competency) in the fields of educational measurement and psychological assessments, such as the Graduate Management Admission Test (GMAT), the National Council of State Boards of Nursing (NCLEX), and the Armed Services Vocational Aptitude Battery (ASVAB). When applied to PRO assessments, the correspondence of concepts being measured is straightforward: the health status or severity of illness of a patient can be regarded as the latent trait that can be easily modeled within the IRT framework. Chang and Reeve [2] described extensively how IRT can be utilized in PRO assessments and the potential benefits of developing a CAT-based PRO measurement system. Fayers [5] also thoroughly discussed the concerns and issues related to the application of IRT and CAT to PRO measurement and argued that IRT and CAT are two extremely powerful and exciting tools to add to our toolbox when developing PRO questionnaires.

Inspired by the potential benefits of CAT for clinical use, more than 30 research studies have been conducted on the application of CAT in PRO assessment and the results are promising. To list a few, Ware et al. [6] developed a CAT system to measure the impact of headaches and concluded that the CAT-based test achieved very large reductions in respondent burden without compromising validity for purposes of patient screening or monitoring changes in headache impact over time. Walter et al. [7] developed a CAT system to measure anxiety and found the correlation between the Anxiety-CAT and the State Trait Anxiety Inventory was 0.60. They concluded that the Anxiety-CAT did indeed show the advantages as expected theoretically, but suggested that further studies are still needed in order to evaluate its full potential for research and clinical practice. Choi and Swartz [8] compared six CAT item selection methods for polytomous items from a bank of 62 depression items and concluded that the simplest maximum Fisher information method performs almost as well as other complicated item selection methods.

Health outcomes are inherently multidimensional. For PRO instruments that are comprised of subscales or domains to capture multidimensional health outcomes, the unidimensionality assumption may not hold when tested using unidimensional IRT models. In this case, multidimensional IRT models may be more appropriate. There have been studies on multidimensional IRT-based CAT in PRO assessment. Petersen et al. [9] evaluated a multidimensional CAT incorporating three (physical functioning, emotional functioning, and fatigue) scales from the EORTC QLQ-C30. They concluded that multidimensional CAT may significantly improve measurement precision and efficiency. Haley et al. [10] studied the properties of a multidimensional CAT using data from two of the functional skills scales of the Pediatric Evaluation of Disability Inventory (PEDI) and concluded that the multidimensional CAT applications appeared to demonstrate both precision and efficiency advantages over the separate unidimensional CAT when the content subdomains are highly correlated.

Among various multidimensional IRT models, the *full-information bifactor model* [11, 12] has become increasingly popular among PRO assessment studies [e.g., [13–16]]. In the bifactor model, each item in a multiscale or multidomain instrument is modeled to have a primary, nonzero loading on the "general" factor and a secondary loading on no more than one of the "domain-specific" factors. The general factor and domain-specific factors are all orthogonal. This double-loading constraint greatly simplifies the computation, because the integration of likelihood for parameter estimation always includes only two dimensions regardless the number of scales/domains in the instrument. Besides its computational simplicity, another reason for the bifactor model's popularity is that it does not only address the multidimensionality issue but also generates an estimate to represent the general latent trait of a patient's health being measured. This overall estimate allows the PRO researchers or practitioners to know what his or her patient's current health status is. Another advantage the bifactor model offers in the development of computerized adaptive PRO assessment (CAT-PRO) is that the general factor can be the focus of the adaptive algorithm when the general factor loadings are greater than the domain-specific factor loadings [2, 13], which leads to a CAT that uses unidimensional adaptive algorithms. Specifically, as part of the strategy proposed by Weiss and Gibbons [16], a unidimensional model is used in the item selection and provisional trait estimation process, and the general factor discrimination parameters and difficulty parameters calibrated from the bifactor model are used as the discrimination parameters and difficulty parameters in the unidimensional model. If doing so, the bifactor CAT will enjoy both a better model fit by modeling the multidimensionality and a simple computation by adapting the test administration only based the unidimensional model.

Although both theoretical and empirical studies have proven the benefits and advantages of CAT-PROs over the traditional paper-and-pencil methods, some limitations remain. One significant limitation is related to the item selection method. The *maximum Fisher information* (MFI) item selection method, a commonly used method in previous studies, picks the item that is considered to be the most informative for the provisional estimated latent trait level. However, this algorithm has two well-known limitations: (1) there is usually an unbalanced or disproportionate use of the item bank, with a few "good" items overexposed and a large proportion of items under-utilized; (2) the assembled tests may not cover all the domains of health that the clinicians and researchers are most concerned about [17]. The strategy Weiss and Gibbons [16] used to solve this problem in their bifactor CAT was to continue administering the adaptive test for each individual domain after finishing the adaptive test on the general factor. For each domain, only the items belonging to that domain can be selected and administered, and the adaptive process continues using the unidimensional model. The difficulty parameters in the unidimensional model continue to be the same with those used on the general factor, but the discriminating parameters are now their corresponding domain-specific discrimination parameters calibrated from the bifactor model. One potential concern is that this strategy may lengthen the test administration. An alternative solution is the content-balancing strategies, which embed statements that explicitly control the domain of the selected item into the item selection algorithm. In this way, the test only needs to adapt based on the general factor, while the balance of domains is done by the content-balancing algorithm.

According to Leung et al. [18], there are three simple and easy content-balancing strategies: (a) the constrained CAT (CCAT), (b) the modified multinomial model (MMM), and (c) the modified constrained CAT (MCCAT). The CCAT, proposed by Kingsbury and Zara [19], selects the most optimal item from the content area with the current exposure rate farthest below its target administration percentage. However, this algorithm may yield undesirable order effects as the sequence of content areas is highly predictable [20]. Therefore, Chen et al. [20] developed the MMM that generates a multinomial variable based on the content category size and uses a uniform random number to decide which category the next item is drawn from. However, this method is subject to randomness and thus cannot guarantee the strict satisfaction of the content constraints. In 2000, Leung, Chang, and Hau proposed the MCCAT as a modification of the CCAT [21]. In the MCCAT, the next optimal item can be chosen from all the content areas that still have the quota not fully used up. This can eliminate the predictability of the sequence of content areas of the CCAT, and it also outperforms the MMM in the aspect of strictly satisfying the content constraints. Although the three content-balancing strategies yield almost identical measurement accuracy for the latent trait in Leung et al. [18] comparison study, our study adopts the MCCAT, because it is better than MMM in guaranteeing all the content constraints to be met and better than CCAT in avoiding undesired order effects.

The purpose of this study is to apply the MCCAT in a bifactor CAT designed for the 36-item Short Form Health Survey (SF-36) [22, 23] and to compare the performance on the measurement accuracy and item bank usage between the CATs with and without the content-balancing strategy.

Methods

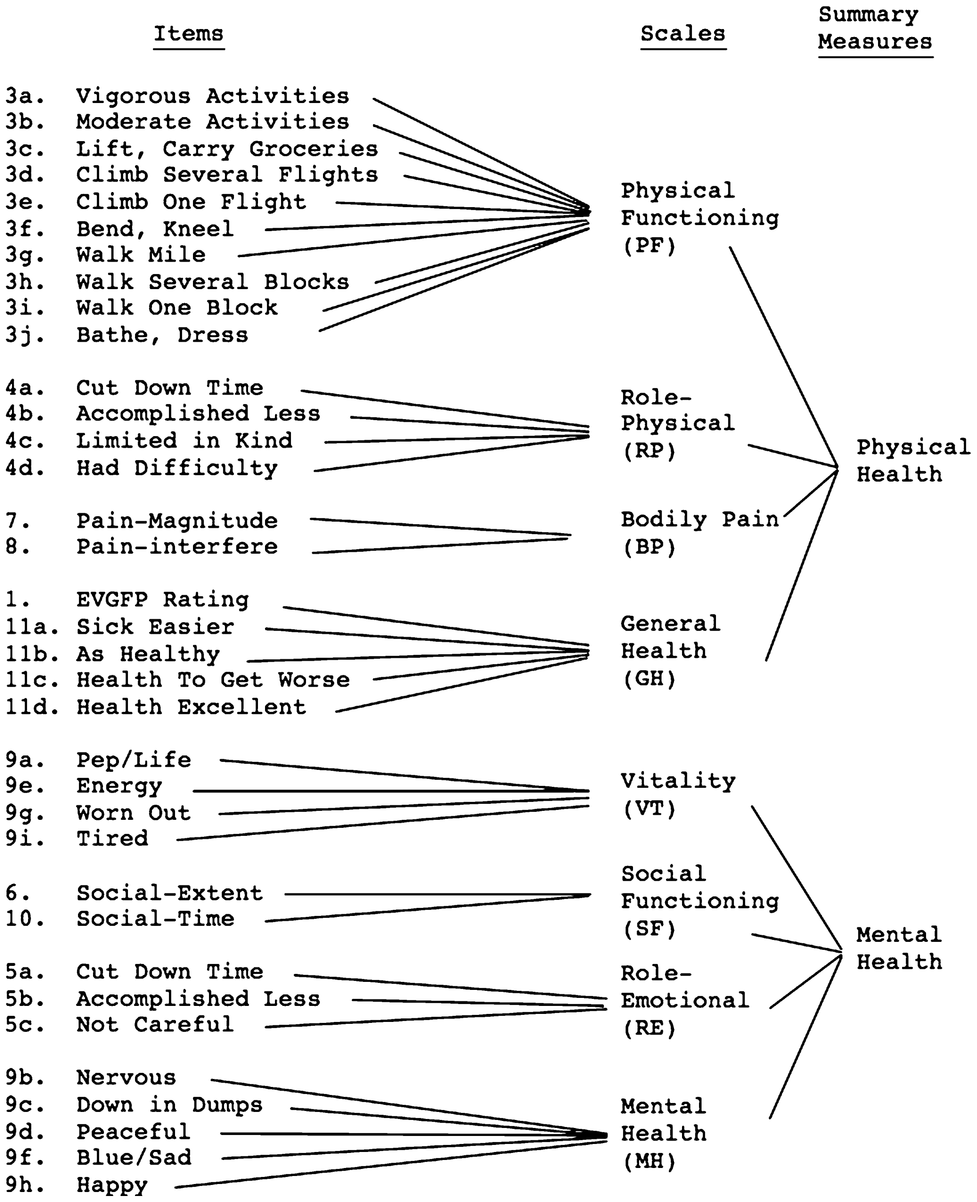
Source of data

The data used for item calibration and CAT simulation in this study are the baseline assessment data of the Medicare Health Outcomes Survey 2003 Cohort VI Baseline ($N = 6{,}801$). Categories of the demographic characteristics have been collapsed in this public use file to ensure confidentiality of the respondents. Only the item responses to the SF-36 questions were used.

Measurement model

The SF-36 is comprised of 35 scored items (and one unscored item) on eight health-related domains that generate two summary components: physical health (PH) and mental health (MH) (see Fig. 1) [24]. Due to its multidimensional and hierarchical nature, the bifactor model was deemed appropriate and used for this study. Considering model simplicity and computation feasibility, instead of having eight domain-specific factors in the model, this study specifies only two domain-specific factors: the physical health factor and the mental health factor. These

Fig. 1 Model structure of SF-36 (adapted from Ware et al. [24])



two dimensions have also been confirmed through factor analysis in previous research [25].

Because the instrument is composed of Likert-type scale items, the *graded response model* (GRM), which implies ordered categories and supports varied discrimination power and varied category step width [26], was chosen. In summary, the model used in the study is the bifactor graded response model [12]. To support its use, the model fit was compared between the unidimensional graded response model and the bifactor graded response model (both fitted to all the 35 items). The *Akaike information criterion* (AIC) for the unidimensional model is 456294.33, while the AIC for the bifactor model is 431732.75, indicating that the bifactor model improves the model fit by a great extent. Moreover, in the bifactor model fitting, three items designed for the physical health dimension (11a, 11b, 11c) go along with the mental health dimension better: they load negatively on the physical health factor under the original factor structure as illustrated in Fig. 1 (AIC = 431732.75), whereas they load positively on the mental health dimension when switched to the mental health factor (AIC = 431247.51). Based on result of the two rounds of model fit comparison, the final model used in the simulation study was chosen to be the bifactor model with item 11a, 11b, and 11c on the mental health factor.

CAT specifications and methods

First, the responses from all the subjects were used to calibrate item parameters using the IRTPRO program

(Scientific Software International, Inc.). Then, a random sample of 1000 subjects were drawn from the dataset, and their responses were utilized in the post hoc CAT simulation. A total of four post hoc CATs with and without the content-balancing strategy incorporated that utilize the real dataset of the SF-36 were simulated. The CAT system was built in MATLAB (R2011a, Mathworks, Inc), a powerful and flexible programming environment.

Being a post hoc CAT, the program simulates the item selection procedure, but the responses of an individual person are directly drawn from the real data set. Based on the responses to the administered items, provisional trait estimates are computed using the *maximum likelihood estimation* (MLE) method, based on which the next item is selected. In the process of item selection and provisional trait estimation, Weiss and Gibbons' [16] strategy of "unidimensional" CAT, as explained in the introduction, was followed in this study. Because a reasonable prerequisite of adopting this strategy is that the general factor loadings should be greater than the domain-specific factor loadings [13], the factor loadings of the final model were checked. As can be seen in Table 1, most of the items (30 out of 35) have greater loadings on the general factor than the domain-specific factor, supporting our adoption of Weiss and Gibbons' strategy.

Table 1 Factor loadings of the SF-36 scored items in the final model

Item	Factors		
	General	Physical	Mental
3a	0.59	0.36	0
3b	0.73	0.47	0
3c	0.71	0.49	0
3d	0.66	0.58	0
3e	0.69	0.59	0
3f	0.62	0.49	0
3g	0.65	0.64	0
3h	0.68	0.67	0
3i	0.69	0.63	0
3j	0.67	0.36	0
4a	0.86	0	0
4b	0.87	−0.04†	0
4c	0.90	0.08	0
4d	0.91	0.02	0
7	0.73	0.05	0
8	0.81	0.07	0
1	0.78	0.07	0
11a	0.59	0	0.27
11b	0.66	0	0.09
11c	0.46	0	0.11
11d	0.77	0.02	0
9a	0.74	0	0.12
9e	0.77	0	0.15
9g	0.67	0	0.24
9i	0.70	0	0.19
6	0.80	0	0.23
10	0.77	0	0.31
5a	0.64	0	0.53
5b	0.59	0	0.51
5c	0.63	0	0.48
9b	0.37	0	0.63
9c	0.50	0	0.76
9d	0.50	0	0.53
9f	0.44	0	0.74
9h	0.43	0	0.54

† Item 12 also loads negatively when switched to the mental health factor. Thus we decided to leave it in the original factor in this study

In terms of the item selection method, we adopted the simplest *maximum Fisher information* (MFI) method in both the content-balanced condition and non-content-balanced condition. The reason is that Choi and Swartz [8] showed that it performs almost as well as other more complicated item selection methods. We also included the random item selection method to serve as a frame of reference for comparison purposes. A 2 × 2 design was therefore employed: 2 item selection methods (random selection method vs. maximum Fisher information method) × 2 content-balancing factors (content-balanced vs. non-content-balanced).

The item bank of the post hoc CAT is composed of the 35 scored items of the SF-36. In the content-balanced CAT, the domain constraint of the CAT was determined according to the structure of the SF-12, a short form of the SF-36 [27, 28]. Table 2 shows the number of items in each of the eight domains of the original SF-36, SF-12, and the content-balanced CAT. As can be seen in Table 2, the assembled CAT is required to draw the exact same number of items from each of the eight domains. As a result, the CAT test length was set at 12. The MCCAT content-balancing strategy was used in the CAT simulation.

After the total of 12 items were administered, the health status of a person was estimated by the Bayesian *expected* a priori *estimation* (EAP) method [12] based upon the responses to those administered items. The general health status is reflected by the trait on the general factor in the bifactor model. However, the traits on the two subscales (physical health and mental health) were not estimated directly through the bifactor model. As Gibbons et al. [12] noted, the trait estimates of the domain-specific factors in the bifactor model "describe associations among the residuals between the items within each subdomain, once the primary dimension has been accounted for." Therefore, the bifactor model domain-specific factor loadings "may underestimate the unconditional subdomain estimates." Gibbons et al. suggested that a reasonable way to obtain estimates for subscales is to apply traditional unidimensional IRT models separately to each subscale and obtain a corresponding subscale trait estimate. Following their suggestion, two separate unidimensional graded response models were fitted to the two subscales, and two sets of unidimensional item parameters were calibrated. In the simulated CAT, among all the items administered to each subject, the responses to the items that belong to the physical subscale along with the corresponding unidimensional item parameters were used to estimate the physical health status. Same procedures were applied to estimate the mental health status.

Evaluation criteria

The performance of the tests was evaluated in two aspects: (a) measurement accuracy and (b) item bank usage. To evaluate the measurement accuracy, the *root mean square error* (RMSE) of the three trait estimates (general, physical, and mental) was computed by Eq. (1):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}, \tag{1}$$

where θ_i is the health status estimated using the responses to all the items in the item bank, and $\hat{\theta}_\mathrm{i}$ is the health status estimated using only those items administered in the CAT.

To evaluate the item bank usage, the item exposure rate was computed for every item by Eq. (2):

$$\text{exposure rate of the } i\text{th item} = \frac{\text{number of examinees who see the item}}{\text{total number of examinees}}. \tag{2}$$

Results

Measurement accuracy

The RMSE of the three trait estimates (general, physical, and mental) were computed for all the experiment conditions by Eq. (1). Table 3 shows the values of RMSE. A low RMSE value indicates a more accurate measurement. The absolute magnitude of RMSE is closely related to test length and the quality of items. Therefore, this study used the CATs with the content-balanced and non-content-balanced random item selection methods as a comparison baseline.

For the estimation of the general health status, the two MFI item selection methods (content-balanced and non-content-balanced) provide better measurement accuracy than the two random item selection methods. Although this is never a surprising result for a traditional unidimensional CAT, the effectiveness of the MFI method is not always guaranteed in the context of this study, where the item parameter calibration and general health estimation used a bifactor model but item selection used a unidimensional model. This result indicates that the unidimensional model works well for the purpose of item selection in this test. This also confirms that higher loading on the general factor than on their respective domain-specific factor is a reasonable and effective prerequisite for adopting the unidimensional CAT strategy under the bifactor model.

Among the MFI methods, the non-content-balanced and content-balanced conditions have similar level of measurement accuracy of the general health status, with the non-content-balanced condition being slightly better. In terms of physical health status, the non-content-balanced condition is more accurate than the content-balanced condition. The mental health status shows an opposite direction that the content-balanced condition is better. This result can be explained by the inclination of the MFI method to selecting physical health items when there is no content-balancing control. The physical health items are generally more discriminating than the mental health items, and therefore, they tend to be more informative. As a result, the non-content-balanced MFI method tends to select the physical health items more. The content-balancing strategy controls the number of items selected from each domain, which indirectly controls the number of items selected from the two domains. Overall, the content-balanced tests have fewer physical health items and more mental health items than the non-content-balanced tests. Consequently, the content-balanced condition measures the mental health better, whereas the non-content-balanced condition measures the physical health better. Although there is a trade-off between the two domains, the result still favors the content-balanced method, because its improvement on the mental side is much greater than the loss on the physical side.

Item bank usage

The exposure rates of the 35 items in the simulated CATs are shown in Fig. 2. The plot supports the explanation in the

Table 2 Number of items from each domain in the SF-36, SF-12, and content-balanced CATs

Domain	PF	RF	BP	GH	VT	SF	RE	MH
SF-36 (CAT item bank)	10	4	2	5	4	2	3	5
SF-12	2	2	1	1	1	1	2	2
Content-balanced CATs	2	2	1	1	1	1	2	2

Table 3 RMSE of health status estimation in the simulated CATs

		Latent traits		
		General	Physical	Mental
MFI	Non-content-balanced	0.2215	0.2629	1.0280
	Content-balanced	0.2262	0.3682	0.3516
RND	Non-content-balanced	0.3122	0.7036	0.5160
	Content-balanced	0.2643	0.5732	0.3897

MFI maximum Fisher information item selection, *RND* random item selection

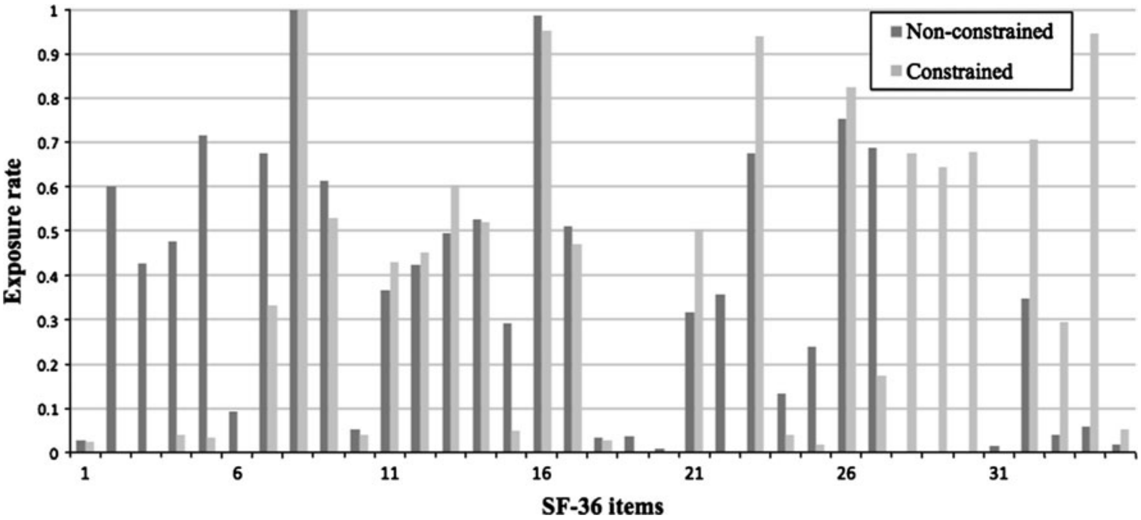


Fig. 2 The exposure rates of SF-36 scored items under the two maximum Fisher information conditions. Item 1 through 17 and 21 are physical health items; item 18 through 20 and 22 through 35 are mental health items

previous subsection that the non-content-balanced tests generally lack mental health items, while the content-balanced tests are more balanced between the two subscales. Within a domain, however, the content-balancing strategy does not prevent the maximum information method to select relatively more discriminating items. For example, the items in the first domain (item 1 through 10) are generally more discriminating than the others, so most of the ten items have high exposure rates under the non-content-balanced condition. But when the content constraint requires only two items to be selected from domain 1, only those relatively more discriminating items (item 7 through 9) are frequently administered, while the remaining are left unused. Unlike educational testing, the high exposure rate may not be a serious issue in CAT-PRO, as it does not seem harmful if a respondent has prior knowledge of the content of the PRO items. However, the under-utilization of some items as indicated by the extremely low exposure rates are still of concern.

Discussion

The study results support that the implementation of the content-balancing strategy in a bifactor CAT can adequately address both the measurement accuracy and item bank usage concerns in measuring multidimensional PROs. The content-balanced CAT has the same level of measurement accuracy in general health status with the non-content-balanced CAT. Although the content-balancing strategy sacrifices the measurement accuracy of physical health by a small amount, it prevents the estimation of mental health from being far away from the true values or even unavailable. Therefore, the content-balanced CAT would be a better choice than the traditional non-content-balanced CAT when obtaining precise estimates for all the three aspects of health is of primary consideration.

Besides the SF-36 used in this study, many other PRO instruments are composed of multiple domains/subscales. It is also common in the PRO instruments that some items in certain domains or subscales outperform the others within the same measurement tool as they relate highly to the underlying trait being measured. This often leads to an item selection procedure that favors the “more informative” domains or subscales (more items with higher Fisher information) and neglects the “less informative” ones. As a result, the scales or domains on which the item selection algorithm neglects suffer from a lack of administered items, which leads to unstable estimation greatly due to randomness. When the discriminating power of items from each domain or subscale varies, the content-balancing strategies are needed to assure sufficient measurement accuracy of each domain or subscale of interest, resulting in a better overall measurement accuracy.

In addition to benefiting the computerized adaptive administration of those already-established multiscale PRO instruments, the content-balancing strategy may also be a powerful tool for assembling instruments adaptively and dynamically from a large item bank. Given an item bank containing items from many different health content domains, the content-balancing strategy can easily restrict the CAT to only select items from those clinically relevant domains deemed important both to the clinicians and their patients. The numbers of items to be selected from each relevant domain can also be specified. The assembled CAT can focus on a single domain. For instance, a depression estimate can be derived from responses to those depression items. It can also be a combination of several health domains as in this study when multidimensional health is of interest. Both overall health and domain-specific health can be estimated using item responses to the assembled CAT.

One limitation of this study is that the content-balancing strategy did not solve the problem related to some items never being selected in the adaptive tests. Further studies may look into methods to improve the usage of those under-selected items. Strategies such as the *a-stratification method* [29, 30] may be applied. In addition, the item bank used in this study has a relatively small number of items (i.e., 35 items). It would be important and necessary to utilize a larger item bank, such as the original SF-36 200-item bank [22, 23, 31] to further examine the impact of the content-balancing strategy on measurement accuracy and item usage in simulated or real CATs.

References

1. Chang, C-H. (2007). Patient-reported outcomes measurement and management with innovative methodologies and technologies. *Quality of Life Research, 16*(Supplement I), 157–166.
2. Chang, C-H., & Reeve, B. B. (2005). Item response theory and its applications to Patient-Reported Outcomes Measurement. *Evaluation & the Health Professions, 28*(3), 264–282.
3. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. CA: Sage Publications.
4. Chang, H.-H. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 117–133). Thousand Oaks, CA: Sage.
5. Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: The challenges for health outcomes assessment. *Quality of Life Research, 16*(Supplement 1), 187–194.
6. Ware, J. E., Jr, & Kosinski, M. (2003). Applications of CAT to the assessment of headache impact. *Quality of Life Research, 12*, 935–952.
7. Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for 'Anxiety' (Axiety-CAT). *Quality of Life Research, 16*(Supplement I), 143–155.
8. Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*(6), 419–440.
9. Petersen, M. Aa., Groenvold, M., Aaronson, N., Fayers, P. M., Sprangers, M. A., & Bjorner, J. B. (2006). Multidimensional computerized adaptive testing of the EORTC QLQ-C30: Basic developments and evaluation. *Quality of Life Research, 15*, 315–329.
10. Haley, S. M., Ni, P. S., Ludlow, L. H., & Fragala-Pinkham, M. A. (2006). Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the pediatric evaluation of disability inventory. *Archives of Physical Medicine and Rehabilitation, 87*, 1223–1229.
11. Gibbons, R., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423–436.
12. Gibbons, R. D., Bock, R., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*(1), 4–19.
13. Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19–31.
14. Haley, S. M., Ni, P., Dumas, H. M., Fragala-Pinkham, M. A., Hambleton, R. K., Montpetit, K., Bilodeau, N., Gorton, G. E., Watson, K., & Tucker, C. A. (2009). Measuring global physical health in children with cerebral palsy: Illustration of a multidimensional bi-factor model and computerized adaptive testing. *Quality of Life Research, 18*, 359–370.
15. Immekus, J. C., Gibbons, R. D., & Rush, A. J. (2007). Patient-reported outcomes measurement and computerized adaptive testing: An application of post-hoc simulation to a diagnostic screening instrument. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC conference on computerized adaptive testing*.
16. Weiss, D. J., & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC conference on computerized adaptive testing*.
17. Cheng, Y., Chang, H-H., Douglas, J., & Guo, F. (2009). Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints. *Educational and Psychological Measurement, 69*(1), 35–49.
18. Leung, C-K., Chang, H.-H., & Hau, K.-T. (2003). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment*, *2*(*5*). Available from http://www.jtla.org.
19. Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*, 359–375.
20. Chen, S., Ankenmann, R. D., & Spray, J. A. (1999, April). *Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
21. Leung, C-K., Chang, H.-H., & Hau, K.-T. (2000, April). *Content-balancing in stratified computerized adaptive testing designs*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
22. Hays, R. D., Sherbourne, C. D., & Mazel, R. M. (1993). The RAND 36-item health survey 1.0. *Health Economics, 2*(3), 217–227.
23. Ware, J. E. Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30*(6), 473–483.

24. Ware, J. E., Kosinski, M., & Keller, S. D. (1994). SF-36 Physical and mental summary scale: A user's manual. Boston, MA: The Health Institute.
25. Chang, C.-H., Wright, B. D., Cella, D., & Hays, R. D. (2007). The SF-36 physical and mental health factors were confirmed in cancer and HIV_AIDS patients. *Journal of Clinical Epidemiology, 60*(1), 68–72.
26. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologist*. NJ: Lawrence Erlbaum Associates.
27. Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*(3), 220–233.
28. Ware, J. E. (2002). *User's manual for the SF-12v2 health survey (with a supplement documenting SF-12 health survey)*. Lincoln, RI: QualityMetric Inc.
29. Chang, H., & Ying, Z. (1999). A-stratified multistage computer adaptive testing. *Applied Psychological Measurement, 23*(3), 211–222.
30. Chang, H., Qian, J., & Ying, Z. (2001). A-stratified multistage computer adaptive testing with b blocking. *Applied Psychological Measurement, 25*(4), 333–341.
31. Stewart, A. L., Sherbourne, C. D., Hays, R. D., Wells, K. B., Nelson, E. C., Kamberg, C., Rogers, W. H., Berry, S. H., Ware, J. E. (1992). Summary and discussion of MOS measures. In A. L. Stewart & J. E. Ware (Eds.), *Measuring functioning and well-being: The medical outcomes study approach* (pp. 345–371). Durham, NC: Duke University Press.