POPULATION HEALTH MANAGEMENT
Volume 19, Number 4, 2016
Mary Ann Liebert, Inc.
DOI: 10.1089/pop.2015.0101

Adaptive Measurement of Well-Being: Maximizing Efficiency and Optimizing User Experience during Individual Assessment

Miriam Kraatz, PhD,[1] Lindsay E. Sears, PhD,[1] Carter R. Coberley, PhD,[1] and James E. Pope, MD[1]

Abstract

Well-being is linked to important societal factors such as health care costs and productivity and has experienced a surge in development activity of both theories and measurement. This study builds on validation of the Well-Being 5 survey and for the first time applies Item Response Theory, a modern and flexible measurement paradigm, to form the basis of adaptive population well-being measurement. Adaptive testing allows survey questions to be administered selectively, thereby reducing the number of questions required of the participant. After the graded response model was fit to a sample of size N = 12,035, theta scores were estimated based on both the full-item bank and a simulation of Computerized Adaptive Testing (CAT). Comparisons of these 2 sets of score estimates with each other and of their correlations with external outcomes of job performance, absenteeism, and hospital admissions demonstrate that the CAT well-being scores maintain accuracy and validity. The simulation indicates that the average survey taker can expect a reduction in number of items administered during the CAT process of almost 50%. An increase in efficiency of this extent is of considerable value because of the time savings during the administration of the survey and the potential improvement of user experience, which in turn can help secure the success of a total population-based well-being improvement program. (*Population Health Management* 2016;19:284–290)

Introduction

Although traditional views of health improvement focus on physiological dimensions, total health is not only the absence of physical and mental health problems, but also the presence of positive states of flourishing across areas of an individual's life.[1–3] Referred to as well-being by many researchers, this more holistic view of individual health has been linked to health care costs and utilization, absenteeism from work, and level of functioning while at work.[4–9] These links highlight the importance and potential value associated with well-being measurement and intervention. Although research supports the business case for well-being measurement as a means to manage and improve population well-being, little research has focused on improving the efficiency and accuracy of the actual methods by which population well-being is assessed.[10–12]

The literature on well-being distinguishes hedonic well-being, which represents people's feelings and thoughts about their lives, from eudemonic well-being, which captures individuals' sense of meaning and purpose in life.[3,13] Within the hedonic approach, there are further distinctions between evaluative well-being and experienced well-being, also characterized as cognitive and affective dimensions, and between global and domain satisfactions.[14–18]

A multitude of well-being measures have been developed over the years, including the satisfaction with life scale, the Short Form 36 (SF-36), a Spiritual Well-Being Questionnaire, and the Ryff Scales of Psychological Well-Being.[17,19–21] Recently, well-being measurement has expanded to a large-scale, continuous assessment by Gallup in its nightly poll in the United States, as well as the World Poll. The Gallup-Healthways Well-being Index (GHWBI) was built to measure both cognitively evaluated and affectively experienced hedonic well-being.[22] Principal component analysis and confirmatory factor analysis of an item pool developed on theoretical grounds identified 6 main factors: life evaluation, emotional health, physical health, healthy behaviors, work environment, and basic access.[23] Following this model, overall well-being is measured as a higher order construct that is a function of domain-specific and global, experienced and evaluative well-being dimensions.

[1]Healthways, Inc., Center for Health Research, Franklin, Tennessee.

Recently, the GHWBI model was evolved into the Well-Being 5 (WB5), based on a series of analyses that investigated items merged from prior validated well-being instruments.[10] Therefore, the WB5 model integrates both the evaluative and experienced dimensions from both hedonic and eudemonic perspectives. Factor analysis suggested 5 main factors: physical, purpose, social, financial, and community well-being in addition to maintaining global measures of experienced and evaluative well-being. Unidimensionality was overwhelmingly demonstrated for the purpose, community, financial, and social elements, which can be viewed as reflective constructs.[24,25] The physical element, however, was shown to have 3 subdomains whose introduction improved model fit significantly. Additional exploratory analyses identified multiple factors within the physical element of well-being. This supports the view of physical well-being as a formative construct in that it is made up of multiple independent yet theoretically connected parts (eg, exercise, smoking, health perception, body mass index).[24,25]

There is little research on measurement efficiency and user experience specific to population-based well-being instruments, even though these considerations may have implications for response rates and thereby the success of a well-being program. Making a survey assessment experience as smooth and pleasant as possible may always be a valuable goal, whether just a few dozen or hundreds of thousands of persons participate. However, for very large-scale assessments, time spent to complete a survey gains special significance, as every extra minute is multiplied manifold. The measurement paradigm of item response theory (IRT) combined with computerized adaptive testing (CAT) promises improvement in both survey duration and user appeal while not sacrificing accuracy or comprehensiveness.[26,27] IRT presents an alternative way to score assessment questions that also may be used to drive an adaptive test engine that only presents the relevant questions to the user needed to estimate their underlying well-being level.

IRT subsumes a family of models that estimate parameters such as item difficulty and item discrimination. Item difficulty places items on a continuum of the trait of interest (such as math ability, depression, or physical functioning) needed to endorse the item, while item discrimination is used to judge with how much confidence one can assign a trait estimate to a test taker. Typically, sets of items with a wide range of difficulties and fairly high discriminations are chosen for a survey. This contrasts with classical test theory that only emphasizes properties of the overall survey. CAT then utilizes the item parameters estimated within the IRT framework to provide a customized assessment to each respondent, thereby minimizing the number of items that do not optimally apply to the individual.

Applications of IRT and CAT have become widespread across fields ranging from education to medicine. CAT has become popular predominantly for aptitude assessment, as is evidenced by its use for large-scale assessments such as the GRE (Graduate Record Examination) and the GMAT (Graduate Management Administration Test).[28,29] Additionally, CAT has been tested and adopted in fields related to well-being such as headache impact, health-related quality of life, the SF-36, depression, and personality assessment.[19,30–34]

Although not the focus of this particular study, a primary motivation for introducing CAT into population well-being assessment is the improvement of user experience. Conversions of health and quality of life assessments have been motivated by the attempt to reduce burden on older or ill patients; examples are the Health-Related Quality of Life and the SF-36.[31,32] Some studies have found support for user preference for a CAT.[35] It seems plausible that such user experience improvements also would benefit a well-being survey for the general population.

The present study applies IRT to the measurement of well-being in order to improve efficiency and optimize user experience, building on a previously validated measurement model for overall well-being. First, analyses are conducted to validate the dimensional structure established in previous papers. Then, through the application of IRT analysis and scoring, accuracy and efficiency of the new approach are investigated.

Method

Sample

The data consisted of 2 independent samples: a sample collected by Gallup with 10,105 participants; the second sample stemmed from 1930 employees from 1 company. Both samples are described extensively in the original measurement development paper.[10] Participation was voluntary and not incentivized. A 2-step process handled missing data: First, cases with more than 50% of responses missing (as many of these cases had almost no responses at all) were deleted, resulting in a sample size of N = 11,640; then the SAS procedure PROC MI (SAS Institute, Inc., Cary, NC) was used to impute remaining missing data.

Measures

Well-being. The well-being measure utilized in this study is the WB5, which is described in detail in Sears et al.[10] Measures of the elements that have been found to be unidimensional and reflective were included in the present analysis: purpose (five 5-point Likert type items), community (7 items: six 5-point Likert type and 1 binary), financial (5 items: three 5-point Likert type and 2 binary), and social well-being scales (four 5-point Likert type items). For the financial element, the 2 dichotomous items were scored as 1 indicator for analyses of dimensionality.

Self-reported Outcomes. Job performance was assessed via a single item from the Health and Performance Questionnaire (HPQ) survey that asked participants to rate their overall job performance during the past 4 weeks on a scale from 0 to 10.[36] Another item also taken from the HPQ inquired about absenteeism for 1 or more entire days during the past 4 weeks. The resulting count variable was converted to a dichotomous variable with a value of zero representing no absences and a value of 1 representing 1 or more days absent. To measure hospital admissions, participants were asked to report the number of times they were admitted to the hospital in the past year. Again, the resulting count variable was converted to a dichotomous variable with no admissions being assigned a value of zero and 1 or more admissions receiving a value of 1.

Analysis

Software. Although the sample was prepared in SAS, all IRT and CAT analyses were conducted using R (The R

Foundation, Vienna, Austria). Item parameter estimates, standard error estimates, and model fit statistics were obtained using the ltm package, while CAT administration was simulated using the catIrt package.[37,38]

Dimensionality. Previous authors found unidimensionality for each of the 4 reflective constructs based on Pearson correlations between the items.[10] The present study further confirmed dimensionality, but acknowledged the mix of 5-point Likert and binary items by conducting confirmatory factor analyses (CFAs) on the polychoric correlation matrix. The evaluation of model fit rested on the standardized root mean square residual (SRMSR), the comparative fit index (CFI), and the root mean square error of approximation (RMSEA) together with its 90% confidence interval.[39–41]

IRT analysis. The following section summarizes the process by which a graded response model was fit, model fit was evaluated, and item information was built into a CAT simulation. A graded response model (GRM) to each of the 4 constructs because of its capability of fitting mixed item format data, its interpretability, and the fact that it forces the category thresholds to be ordered, something alternative models do not call for.[26,42]

Researchers evaluated IRT model fit using item fit, person fit, and model comparisons.[26] They demonstrated the need for the complexity of an unconstrained GRM (U-GRM) that estimates a discrimination parameter for each item over a constrained GRM (C-GRM) that estimates 1 discrimination parameter for all items by conducting a likelihood ratio test, calculated as the difference between the −2 log likelihoods for the U-GRM and the C-GRM. This procedure is mentioned in Rizopoulos (see also Thissen et al) and has been applied by Andrae et al.[37,43,44] The researchers also reported the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), both of which are part of the ltm package output. For both the AIC and BIC, lower values indicate the preferred model.

After fitting the models, the R package catIrt was used to obtain theta score estimates for each element based on (*a*) the full item bank (FIB scores) and (*b*) simulating a CAT administration (CAT scores). The stopping rule for the CAT simulation was based on a precision goal; the CAT was halted when the standard error estimate for the theta score reached a value of 0.5 or lower.

Accuracy and Efficiency. To test the accuracy of CAT scores relative to FIB scores, the researchers calculated correlations between these scores. To examine efficiency, they presented frequency data on the number of items that individuals in a population would receive in an adaptive scenario as opposed to receiving the full instrument.

Validity. Correlations of FIB scores and CAT scores were included with the self-reported outcomes to investigate criterion validity, in the hope of demonstrating that utilizing scores produced with CAT does not reduce predictive power for these outcomes by any practical means. The interested reader may choose to compare effect sizes observed in this study with effect sizes reported in Sears et al.[10] To test whether the relationship between trait score estimate and outcomes differs depending on the method by which the trait score estimate is obtained (FIB or CAT), 95% confidence intervals were constructed for the difference between 2 dependent correlations.[45]

Results

Dimensionality

The results of CFAs on polychoric correlation matrices of the 4 elements are presented in Table 1. Compared to commonly accepted standards, both the SRMSR and CFI statistics still indicate good fit for all 4 reflective elements. For the SRMSR and the RMSEA, values less than .08 and .06, respectively, are generally considered indicating a good fit, while for the CFI, values of 0.95 or larger suggest good fit.[39]

IRT analysis

Evidence of overall model fit is presented in Table 2, which lists log likelihoods, AIC, and BIC values for and the chi-square difference test between the C-GRM and U-GRM on all elements. In each instance, the difference between the constrained and unconstrained model is highly significant and both AIC and BIC assume a smaller value for the U-GRM, indicating the necessity for the increased complexity of the U-GRM. Item parameter estimates together with their standard error estimates are presented in Table 3. Most items demonstrate good to very good discrimination with many estimated item discriminations varying between 1.5 and 3.5. Particularly striking is the low discrimination parameter for the seventh item in the community scale, with a value of 0.544. Also noteworthy is the distribution of threshold estimates. Although thresholds between the 2 lowest response categories (low theta range) lie consistently at least 2 standard deviations below the mean, the high theta range does not receive the same amount of coverage: aside from the purpose element, most thresholds between the 2 highest response categories lie around a value of 1. Consequently, test information drops and the standard error will be high for trait values of more than 1 standard deviation above the mean.

Accuracy, validity, and efficiency

Table 4 displays correlations between FIB scores and CAT scores as well as correlations of both sets of scores with the 3 outcome variables of self-reported job performance, absenteeism, and hospital admissions. The correlations between CAT and FIB scores were consistently very high, with 0.969 for the community element being the smallest, followed by

Table 1. SRMSR, CFI, and RMSEA for Unidimensional Models of WB5 Well-being Elements Purpose, Community, Financial, and Social

Well-being Element	*SRMSR*	*CFI*	*RMSEA*
Purpose	0.029	0.979	0.095 (0.088, 0.102)
Community	0.040	0.962	0.110 (0.106, 0.115)
Financial	0.026	0.985	0.118 (0.107, 0.129)
Social	0.010	0.997	0.044 (0.033, 0.055)

CFI, Comparative Fit Index; RMSEA, root mean square error of approximation; SRMSR, standardized root mean square residual; WB5, Well-Being 5.

Table 2. Model Fit Indices for the Constrained and Unconstrained GRM

Well-Being Element	*Log Likelihood*	*AIC*	*BIC*
Purpose, unconstrained	−72211.44	144472.9	144656.9
Purpose, constrained	−72984.50	146011.0	146165.6
Log likelihood difference test	−2*ΔLL = 1546.12, df = 4		
Community, unconstrained	−85967.21	171998.4	172234.0
Community, constrained	−90340.68	180733.4	180924.8
Log likelihood difference test	−2*ΔLL = 8746.94, df = 6		
Financial, unconstrained	−52533.89	105105.8	105245.7
Financial, constrained	−53016.31	106062.6	106173.1
Log likelihood difference test	−2*ΔLL = 964.84, df = 4		
Social, unconstrained	−58965.63	117971.3	118118.5
Social, constrained	−59649.68	119333.4	119458.5
Log likelihood difference test	−2*ΔLL = 1368.1, df = 3		

AIC, Akaike information criterion; BIC, Bayesian information criterion; GRM, graded response model.
Constrained: models with 1 discrimination parameter for all items. Unconstrained: models where 1 discrimination parameter is estimated for each individual item.

0.975 for the financial element, 0.980 for the purpose element, and 0.990 for the social element.

Although several of the correlations of CAT scores with outcomes are significantly different from their counterparts based on FIB scores (indicated by a confidence interval that does not include zero), these differences are very small.

Figure 1 presents a frequency plot of the total number of items delivered during the CAT simulation. The average number of items administered for the purpose, community, financial, and social elements were 2.87, 2.60, 2.37, and 2.93, respectively. Adding the number of items across all 4 elements together, the average number of items administered was 10.77 from the total item pool of 21, which equals a 49% reduction in the number of items completed. This demonstrates a considerable reduction, especially when keeping in mind that hundreds of thousands of these surveys are completed every year.

Table 3. Graded Response Model Item Parameter Estimates for All Four Reflective Elements of the WB5 with Standard Error Estimates in Parentheses

Well-Being Element	*Item*	*Slope*	*Thresholds*			
		α	β_1	β_2	β_3	β_4
Purpose						
	Item 1	1.261 (0.025)	−2.062 (0.041)	−0.948 (0.034)	0.263 (0.024)	1.655 (0.107)
	Item 2	3.065 (0.064)	−2.441 (0.037)	−1.590 (0.058)	−0.589 (0.047)	0.836 (0.060)
	Item 3	1.753 (0.032)	−2.041 (0.034)	−1.021 (0.035)	0.185 (0.024)	1.827 (0.214)
	Item 4	3.105 (0.064)	−2.114 (0.030)	−1.278 (0.048)	−0.335 (0.038)	0.944 (0.091)
	Item 5	1.616 (0.030)	−2.680 (0.048)	−1.351 (0.047)	−0.038 (0.035)	1.472 (0.099)
Community						
	Item 1	3.528 *(0.057)*	−1.583 *(0.018)*	−0.787 *(0.027)*	0.060 *(0.020)*	1.219 *(0.328)*
	Item 2	3.040 *(0.078)*	−1.255 *(0.019)*			
	Item 3	4.150 *(0.073)*	−1.716 *(0.020)*	−1.002 *(0.034)*	−0.104 *(0.027)*	1.003 *(0.332)*
	Item 4	3.210 *(0.050)*	−2.029 *(0.025)*	−1.286 *(0.038)*	−0.242 *(0.030)*	0.943 *(0.089)*
	Item 5	1.251 *(0.024)*	−3.973 *(0.086)*	−2.512 *(0.081)*	−0.927 *(0.065)*	1.237 *(0.072)*
	Item 6	1.514 *(0.027)*	−2.824 *(0.051)*	−1.772 *(0.052)*	−0.760 *(0.041)*	0.736 *(0.040)*
	Item 7	0.544 *(0.019)*	−1.077 *(0.050)*	0.875 *(0.054)*	2.556 *(0.171)*	4.782 *(0.637)*
Financial						
	Item 1	1.444 *(0.050)*	−2.131 *(0.052)*			
	Item 2	4.051 *(0.127)*	−1.418 *(0.019)*	−0.673 *(0.047)*	0.125 *(0.030)*	1.218 *(1.103)*
	Item 3	1.867 *(0.035)*	−1.074 *(0.022)*	−0.240 *(0.018)*	0.319 *(0.021)*	1.028 *(0.076)*
	Item 4	1.519 *(0.044)*	−1.475 *(0.033)*			
	Item 5	2.326 *(0.043)*	−2.483 *(0.039)*	−1.577 *(0.052)*	−0.582 *(0.041)*	0.838 *(0.045)*
Social						
	Item 1	2.014 (0.041)	−2.228 (0.037)	−1.543 (0.049)	−0.680 (0.037)	0.445 (0.029)
	Item 2	3.420 (0.096)	−2.332 (0.035)	−1.466 (0.073)	−0.439 (0.058)	0.827 (0.085)
	Item 3	1.131 (0.025)	−2.164 (0.046)	−0.954 (0.037)	0.115 (0.025)	1.532 (0.085)
	Item 4	1.976 (0.040)	−2.644 (0.046)	−1.665 (0.056)	−0.613 (0.044)	0.680 (0.040)

WB5, Well-Being 5.
Italicized standard error estimates are ''approximate estimates'' based on several modifications of options in the numerical optimization process.

Table 4. Correlations of FIB and simulated CAT Scores with Each Other and Outcomes Self-reported Job Performance, Absenteeism, and Hospital Admissions

Well-Being Element	*CAT Score*	*Job Performance (n = 7332)*	*Absenteeism (n = 10,765)*	*Admissions (n = 11,347)*
Purpose				
FIB	0.980	0.399	–0.092	–0.045
CAT		0.392	–0.091	–0.046
Difference [CI]		0.007* [0.003, 0.011]	–0.001 [–0.004, 0.003]	0.001 [–0.003, 0.005]
Community				
FIB	0.969	0.255	–0.072	–0.012
CAT		0.226	–0.065	–0.015
Difference [CI]		0.029* [0.023, 0.034]	–0.006* [–0.011, –0.002]	0.002 [–0.002, 0.007]
Financial				
FIB	0.975	0.220	–0.138	–0.036
CAT		0.219	–0.130	–0.038
Difference [CI]		0.001 [–0.004, 0.006]	–0.009* [–0.013, –0.005]	0.002 [–0.002, 0.006]
Social				
FIB	0.990	0.289	–0.079	–0.006
CAT		0.284	–0.076	–0.004
Difference [CI]		0.005* [0.002, 0.008]	–0.003* [–0.006, –0.001]	–0.003* [–0.005, 0.000]

*Difference between correlations is significantly different from zero with a Type I Error rate of 0.05.
FIB, full item bank; CAT, computerized adaptive testing.

Discussion

IRT and CAT present viable methods for a more efficient and user-friendly well-being assessment. Key findings of the present study are: (1) The U-GRM fits the 4 reflective elements of the WB5 survey well, (2) applying CAT by all practical means retains the precision of measurement achieved when employing the entire item bank while (3) leading to considerable savings in the number of items presented to participants.[42]

The first finding listed above is supported by a demonstration of adequate unidimensionality of the items within each reflective element and a comparison of the restricted vs. the less restricted GRM. The GRM was chosen over other models such as the generalized partial credit model for theoretical considerations, as it allows for ordered categories and additive response probabilities. Unanimously better fit of its unconstrained version when compared to a GRM with just 1 slope for all items highlighted the need for individual slopes. The researchers investigated the second finding by estimating trait scores based on the entire item bank and contrasting them with simulated CAT scores. Results show an overwhelming support of the viability of CAT in this scenario with very high correlations between FIB and CAT scores. Construct validity was not impaired by the transition from FIB to CAT scores. Although maintaining precision, the CAT simulation resulted in a significant reduction in number of items administered, reducing the test load by 49%, despite the shortness of the original scales (third finding). Given the large scale at which the WB5 is administered, this constitutes meaningful time savings. A shorter survey can be expected to result in decreased costs for well-being programs and health management, and overall increased efficiency in any care or lifestyle management guided or initiated by well-being measurement.

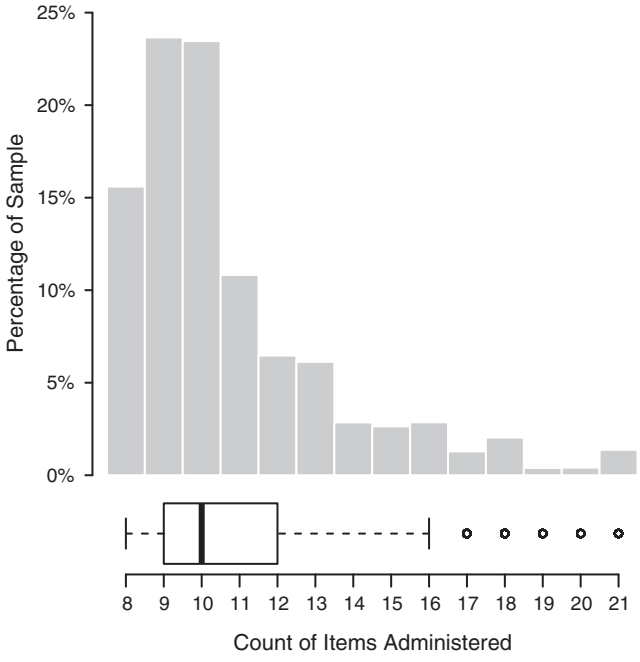

FIG. 1. Number of items delivered to a sample of adaptive well-being measurement administration.

It is vital to note that a more concise and relevant survey experience for the end user may assure higher survey completion rates. This is essential to the success of well-being improvement programs because the survey responses guide prioritization and outreach to individuals for coaching. Improvements in efficiency go hand in hand with improvements in user experience. One study found that participants perceived CAT assessment as clear and easy to use as compared to other measurement instruments and described the CAT test length as ''better than expected.''[35] In Turner-Bowker et al, the majority of participants rated the DYNHA SF-36, a computerized adaptive version of the SF-36 Health Survey, as relevant and easy to complete.[19,32] One study asked 26 participants 5 questions regarding the preference of the CAT or static versions of the survey.[46] For all 5 questions, the CAT

was preferred, and despite the small sample size, 2 of those comparisons were significant. They conclude that ''Pilot study patients preferred self-administered CAT surveys.''

In addition to the reduction in the number of items needed to reach a certain precision goal, IRT and CAT offer even more: successively choosing and presenting items whose difficulty applies to the survey taker's trait level very well may create a much more pleasant user experience. Despite a lack of research focused solely on the user experience of CAT vs. traditional assessment, one theory that may support the idea that CAT will improve user experience is Csikszentmihalyi's theory of the experience of flow.[47,48] The core of Csikszentmihalyi's theory is the balance between skill(s) and challenges. When challenge and confidence are well balanced, the person involved in an activity has the best chance to experience flow, which Csikszentmihalyi describes as a ''sense of discovery'' and a special ''state of consciousness.'' It is not too hard to imagine that the ability of CAT to adjust to a test taker's skill level can create a flow experience in aptitude assessments. With these results and theoretical considerations in mind, one goal for future research could be to assess the reception of an adaptive WB5 survey.

There are many opportunities to expand on the results from this study. Although the sample on which the item parameters were estimated covers large parts of the American public's age range, education levels, and employment status, among others, it is not a stratified sample of, and therefore does not fully represent, the American public. Replication of these analyses in other samples and investigation of differential item functioning will deepen the scope of the conclusions drawn.[26] IRT can be a useful tool for evaluating and selecting new items for an expanded item pool. Although the discrimination parameter indicates the strength of the relationship between the item and the underlying latent construct of interest, similar to the correlation in a factor analysis, threshold parameters can advise how well the construct of interest is covered across the entire range of the trait. The 4 item banks utilized in this study are quite small, containing between 4 and 7 items. They each cover the range of 2 standard deviations below and 1 standard deviation above the mean quite well. However, for some of the elements coverage has potential for expansion, especially of the high range of well-being.

Future research may involve fitting more complex IRT models to well-being data. The U-GRM is a unidimensional model that was fit individually to each of the 4 reflective elements of the WB5. Yet there are multidimensional IRT (MIRT) models available, and research centering on these models has gained momentum.[49,50] Because of their complexity, the researchers refrained from fitting such models in this first application of IRT and CAT to population well-being measurement, although MIRT might provide superior model fit as well as small improvements to test length during CAT, as scores on elements assessed first could be used for initial trait estimate prediction for the other elements. Last but not least, job performance, absenteeism, and hospital admissions are self-reported. As both well-being variables and outcomes are self-reported, the results are susceptible to method bias. Future research could attempt to employ more independent measurement of these constructs. Well-being research will benefit greatly from establishing and confirming relationships of well-being constructs with other concepts, such as self-efficacy or personality traits (eg, the Big Five) and examining long-term developments. The latter goal, in particular, is becoming more feasible as continuous and frequent assessment via employers or population assessments such as the nightly poll by Gallup-Healthways gain ground.

Author Disclosure Statement

Drs. Kraatz, Sears, Coberley, and Pope declared the following conflicts of interest with respect to the research, authorship, and/or publication of this article: The authors were employed by and/or stakeholders of Healthways, Inc. when the work was conducted.

The authors received the following financial support for this article: This study was funded by Healthways, Inc.

References

1. Keyes CLM. Mental illness and/or mental health? Investigating axioms of the complete state model of health. J Consult Clin Psychol. 2005;73:539–548.
2. World Health Organization. Constitution of the World Health Organization. 1995. http://www.who.int/governance/eb/who_constitution_en.pdf. Accessed July 1, 2015.
3. Diener E. Assessing subjective well-being: progress and opportunities. SOCI. 1994;31:103–157.
4. Gandy WM, Coberley C, Pope JE, Wells A, Rula EY. Comparing the contributions of well-being and disease status to employee productivity. J Occup Environ Med. 2014;56:252–257.
5. Shi Y, Sears LE, Coberley CR, Pope JE. The association between modifiable well-being risks and productivity: a longitudinal study in pooled employer sample. J Occup Environ Med. 2013;55:353–364.
6. Sears LE, Shi Y, Coberley CR, Pope JE. Overall well-being as a predictor of health care, productivity, and retention outcomes in a large employer. Popul Health Manag. 2013;16:397–405.
7. Harrison PL, Pope JE, Coberley CR, Rula EY. Evaluation of the relationship between individual well-being and future health care utilization and cost. Popul Health Manag. 2012;15:325–330.
8. Shi Y, Sears LE, Coberley CR, Pope JE. Classification of individual well-being scores for the determination of adverse health and productivity outcomes in employee populations. Popul Health Manag. 2013;16:90–98.
9. Gandy WM, Coberley C, Pope JE, Rula EY. Well-being and employee health-how employees' well-being scores interact with demographic factors to influence risk of hospitalization or an emergency room visit. Popul Health Manag. 2014;17:13–20.
10. Sears LE, Agrawal S, Sidney JA, et al. The well-being 5: development and validation of a diagnostic instrument to improve population well-being. Popul Health Manag. 2014;17:357–365.
11. Diener E. Subjective well-being: the science of happiness and a proposal for a national index. Am Psychol. 2000;55(1):34–43.
12. Hagerty MR, Cummins RA, Ferriss AL, et al. Quality of life indexes for national policy: Review and agenda for research. SOCI. 2001;55(1):1–96.
13. Steptoe A, Deaton A, Stone AA. Psychological wellbeing, health and ageing. Lancet. 2015;385(9968):640–648.
14. Stone AA, Mackie C. Subjective Well-being: Measuring Happiness, Suffering, and Other Dimensions of Experience. Washington, DC: National Academies Press; 2014.

15. Kahneman D, Riis J. Living, and thinking about it: two perspectives on life. In: Baylis N, Huppert FA, Keverne B, eds. The Science of Well-being. Oxford, UK: Oxford University Press; 2005:285–304.
16. Andrews FM, Withey SB. Social Indicators of Well-being: Americans' Perceptions of Life Quality. New York: Springer Publishing; 1976:63–106.
17. Diener E, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. J Pers Assess. 1985;49(1):71–75.
18. Diener E, Lucas RE, Scollon CN. Beyond the hedonic treadmill: revising the adaptation theory of well-being. Am Psychol. 2006;61:305–314.
19. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36). Med Care. 1992;30:473–483.
20. Gomez R, Fisher JW. Domains of spiritual well-being and development and validation of the spiritual well-being questionnaire. Person Indiv Differ. 2003;35:1975–1991.
21. Ryff CD. Happiness is everything, or is it? Explorations on the meaning of psychological well-being. J Pers Soc Psychol. 1989;57:1069–1081.
22. Gallup-Healthways. Gallup-Healthways Well-Being Index: Methodology Report for Indexes. 2009. http://wbi.meyouhealth.com/files/GallupHealthwaysWBI-Methodology.pdf. Accessed July 1, 2015.
23. Evers KE, Prochaska JO, Castle PH, et al. Development of an individual well-being scores assessment. Psychol of Well-Being. 2012;2(1):1–9.
24. Bollen K, Lennox R. Conventional wisdom on measurement: a structural equation perspective. Psychol Bull. 1991;110:305–314.
25. Diamantopoulos A. Export performance measurement: reflective versus formative indicators. Int Market Rev. 1999; 16:444–457.
26. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
27. Wainer H, Dorans NJ, Flaugher R, Green BF, Mislevy RJ. Computerized Adaptive Testing: A Primer. 2nd ed. London: Routledge; 2000.
28. Mills CN. Development and introduction of a computer adaptive Graduate Record Examinations General Test. In: Drasgow F, Olson-Buchanan JB, eds. Innovations in Computerized Assessment. Mahwah, NJ: Lawrence Erlbaum Associates; 1999:117–135.
29. Rudner LM. Implementing the graduate management admission test computerized adaptive test. In: van der Linden WJ, Glas CAW, eds. Elements of Adaptive Testing. New York: Springer; 2010:151–165.
30. Ware JE, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing. Med Care. 2000;38(9 suppl):II73–II82.
31. Rebollo P, Castejon I, Cuervo J, et al. Validation of a computer-adaptive test to evaluate generic health-related quality of life. Health Qual Life Outcomes. 2010;8:147.
32. Turner-Bowker DM, Saris-Baglama RN, DeRosa MA, Giovannetti ER, Jensen RE, Wu AW. A computerized adaptive version of the SF-36 is feasible for clinic and Internet administration in adults with HIV. AIDS Care. 2012;24:886–896.
33. Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a computer-adaptive test for depression (D-CAT). Qual Life Res. 2005;14:2277–2291.
34. Simms LJ, Clark LA. Validation of a computerized adaptive version of the schedule for nonadaptive and adaptive personality (SNAP). Psychol Assess. 2005;17(1):28–43.
35. Nikolaus S, Bode C, Taal E, Vonkeman HE, Glas CAW, van de Laar MAFJ. Acceptance of new technology: a usability test of a computerized adaptive test for fatigue in rheumatoid arthritis. JMIR Human Factors. 2014;1(1):e4.
36. Kessler RC, Barber C, Beck A, et al. The World Health Organization health and work performance questionnaire (HPQ). J Occup Environ Med. 2003;45:156–174.
37. Rizopoulos D. ltm: an R package for latent variable modeling and item response theory analyses. J Stat Softw. 2006;17(5):1–25.
38. Nydick SR. catIrt: An R Package for Simulating IRT-Based Computerized Adaptive Tests. 2014; R package version 0.5-0.
39. Hu Lt, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equat Model. 1999;6(1):1–55.
40. Bentler PM. Comparative fit indexes in structural models. Psych Bulletin. 1990;107:238–246.
41. Steiger JH, Lind JC. Statistically based tests for the number of factors. Paper presented at: Annual Meeting of the Psychometric Society;1980; Iowa City, IA.
42. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement #17. Richmond, VA: Psychometric Society; 1969.
43. Thissen D, Steinberg L, Gerrard M. Beyond group-mean differences: the concept of item bias. Psych Bull. 1986; 99(1):118–128.
44. Andrae DA, Covington PS, Patrick DL. Item-level assessment of the irritable bowel syndrome quality of life questionnaire in patients with diarrheal irritable bowel syndrome. Clin Ther. 2014;36:663–679.
45. Diedenhofen B, Musch J. cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One. 2015;10(3):e0121945-e0121945.
46. Ware JE, Gandek B, Sinclair SJ, Bjorner JB. Item response theory and computerized adaptive testing: implications for outcomes measurement in rehabilitation. Rehabil Psychol. 2005;50(1):71–78.
47. Csikszentmihalyi M. Beyond Boredom and Anxiety. San Francisco, CA: Jossey-Bass Inc Pub; 1975.
48. Linacre JM. Computer-adaptive testing: a methodology whose time has come. In: Chae S, Kang U, Jeon E, Linacre J, eds. Development of Computerised Middle School Achievement Tests. Vol 69. Seoul, South Korea: Komesa Press; 2000.
49. Reckase M. Multidimensional Item Response Theory. New York: Springer; 2009.
50. Ackerman TA, Gierl MJ, Walker CM. Using multidimensional item response theory to evaluate educational and psychological tests. Educ Meas 2003;22(3):37–51.

Adddress correspondence to:
Dr. Lindsay E. Sears
Healthways, Inc.
Center for Health Research
701 Cool Springs Blvd
Franklin TN 37067

E-mail: Lindsay.Sears@healthways.com