COMMENTARY

The PROMIS® Smoking Assessment Toolkit—Background and Introduction to Supplement

Maria Orlando Edelen PhD

RAND Health, RAND Corporation, Boston, MA

Corresponding Author: Maria Orlando Edelen, PhD, RAND Health, RAND Corporation, 20 Park Plaza, Suite 920, Boston MA 02116, USA. Telephone: 617-338-2059, ext. 8634; Fax: 617-357-7470; E-mail: orlando@rand.org

ABSTRACT

Introduction: The PROMIS® Smoking Initiative has developed an assessment toolkit for measuring 6 domains of interest to cigarette smoking research: nicotine dependence, coping expectancies, emotional and sensory expectancies, health expectancies, psychosocial expectancies, and social motivations for smoking. The papers in this supplement describe the methods used to develop these item banks, their psychometric properties, and the preliminary evidence for their validity. This commentary is meant to provide background information for the material in this supplement.

Methods: After discussing the use of item response theory in behavioral measurement, I will briefly review the initial developmental steps for the smoking assessment toolkit. Finally, I will describe the contents of this supplement and provide some closing remarks.

Results: Psychometric evidence strongly supports the utility of the toolkit of item banks, short forms (SFs), and computer adaptive tests (CATs). The item banks for daily smokers produce scores with reliability estimates above 0.90 for a wide range of each cigarette smoking domain continuum, and SF and CAT administrations also achieve high reliability (generally greater than 0.85) using very few items (4–7 items for most banks). Performance of the banks for nondaily smokers is similar. Preliminary evidence supports the concurrent and the discriminant validity of the bank domains.

Conclusions: The new smoking assessment toolkit has attractive measurement features that are likely to benefit smoking research as researchers begin to utilize this resource. Information about the toolkit and access to the assessments is available at the project Web site (http://www.rand.org/health/projects/promis-smoking-initiative.html) and can also be accessed via the PROMIS Assessment Center (www.assessmentcenter.net).

INTRODUCTION

A major focus of recent NIH-funded research is the advancement of behavioral health measurement through development of new self-report tools based on the principles of item response theory (IRT). At the forefront of these efforts is the Patient Reported Outcomes Measurement Information System or PROMIS® (http://www.nihpromis.org/), an NIH Roadmap initiative that has set the standard for modern behavioral measurement development (Cella et al., 2007). The main goals of PROMIS are to standardize a set of assessment tools and to use modern measurement theory (i.e., IRT; Edelen & Reeve, 2007; Embretson & Reise, 2000; Jones & Thissen, 2007) and advances in computer technology to create and utilize item banks to measure patient reported outcomes (Ader, 2007; Cella et al., 2007; Fries, Bruce, & Cella, 2005). PROMIS was developed, in part, to increase the availability and use of a common set of standardized assessment tools that in the long term would enhance the comparability of findings across studies examining patient-reported constructs, reduce respondent burden, and increase measurement precision.

With funding from the National Institute on Drug Abuse (R01 DA026943; PI: Maria Edelen), and in partnership with lead researchers within and outside the PROMIS network, the PROMIS Smoking Initiative (Shadel, Edelen, & Tucker, 2011) has developed an assessment toolkit for current adult cigarette smokers that enables precise and efficient measurement of six constructs of central importance to smoking research: nicotine dependence, coping expectancies, emotional and sensory expectancies, health expectancies, psychosocial expectancies, and social motivations for smoking. The papers in this supplement describe the methods used to develop these item banks, present the derivation and psychometric properties of each of the banks, and provide preliminary evidence for their validity.

Self-report assessment instruments are used in a variety of smoking research contexts (e.g., to examine smoking trends over time, identify predictors of smoking behavior, examine the impact of smoking on disease development and outcomes, evaluate the efficacy of prevention and cessation programs), and the use of well-chosen, psychometrically sound smoking assessment instruments is critical in all of these situations (Panter & Reeve, 2002). However, assessment of

smoking-related constructs is complicated by the fact that several different items and measures exist to assess each construct and there is little in the way of a gold standard or consensus for how various aspects of smoking should be assessed. The PROMIS Smoking Initiative has sought to bolster smoking research by updating smoking assessment using IRT and item banking. Our use of IRT-based item banks has resulted in an efficient, flexible, and versatile assessment toolkit for sustained use in cigarette smoking research. It is hoped that the availability of a standardized toolkit for smoking assessment will contribute to a more integrated and coherent framework from which to improve understanding of smoking and cessation.

The goal of this paper is to provide background information that may serve as an orientation to the material in the remainder of this supplement. To that end, I first provide a discussion of the use of IRT for smoking assessment. This is followed by a brief review of the initial developmental steps we undertook in generating the smoking assessment toolkit. Finally, I describe the contents of the supplement and provide some closing remarks including information on how to access the smoking assessment toolkit.

IRT AND IRT-BASED ITEM BANKING FOR SMOKING ASSESSMENT

The application of IRT and IRT-based item banking for instrument development and refinement has gained considerable momentum in behavioral research within the past decade (e.g., Edelen & Reeve, 2007; Hahn, Cella, Bode, Gershon, & Lai, 2006; Teresi, 2006), largely because the IRT approach has several attractive measurement features. To begin with, an essential characteristic of IRT models is that *reliability*, or measurement precision, is conditional on values of the measured construct. Precision is a function of the item parameters that can be calculated for individual items, sets of items, or entire scales. These parameters, estimated based on the IRT model, convey the strength of each item's relationship to the measured construct and indicate the range along the construct score continuum where an item provides the most reliable responses. Because the item parameters are estimated with respect to a clearly defined scale for the underlying latent variable that is being measured, IRT is said to have a "built-in" *linking* mechanism (Embretson, 1996; Linn, 1992; Mislevy, 1992). This important feature facilitates comparability of scores that represent a common construct but are derived from different sets of items.

These and other measurement features of IRT allow for the development of *item banks*, or sets of linked items with known properties (i.e., estimated IRT parameters). New items can be added to the bank or existing items can be modified and tested so that the building of an item bank can be ongoing as needed. In effect, the IRT-based item banking framework results in a sustainable measurement solution in the sense that the item bank scores representing a particular construct have a consistent interpretation even if some items are deleted or new ones are added to the bank over time.

The existence of banks of items makes possible the construction of tests *tailored* to specific purposes. For example, in a study where the outcome of interest is meeting criteria for a nicotine dependence diagnosis, a tailored assessment could select from the nicotine dependence bank only those items that

discriminate best at or near the diagnostic cut point (assuming the cut point has been determined previously). Computer adaptive testing (CAT) extends this idea further and essentially treats all tests as tailored, but to the individual rather than to a point on the underlying continuum. Because the computer-adaptive scoring approach selectively administers items based on known item parameters, the number of items required to calculate a precise score for a given individual, and thus the respondent burden, can be reduced substantially.

The IRT framework also offers a straightforward way to test for *differential item functioning* or DIF. An item exhibits DIF if two respondents who differ on the grouping variable being considered (e.g., gender) have equal levels of the construct being measured but do not have the same probability of endorsing each response category of that item (i.e., have different item parameters). For example, men and women with equal levels of depression have unequal probabilities of endorsing the *crying* symptom. Ignoring DIF can lead to misleading group differences and inaccurate bivariate associations (Holland & Wainer, 1993); thus the ease with which problematic DIF can be eradicated from IRT-based scales leads to more robust measures.

In summary, IRT-based item bank development yields a highly versatile set of tools. The fact that the item banks have known characteristics allows developers to evaluate and exclude items that show unacceptable levels of DIF, and enables linking of scores from different forms and tailoring of tests for specific purposes while maintaining a pre-specified degree of measurement precision. This measurement flexibility also extends to a wide array of administration options and platforms—such as computer-based assessment, use of handheld devices such as smartphones and notepads, CAT, and tailored paper and pencil short forms—all of which minimize respondent burden without sacrificing reliability and precision (Embretson, 1996; Hambleton & Swaminathan, 1985; Lord, 1980; Wainer, 2000; Wainer & Mislevy, 2000). This flexibility has the potential to directly impact smoking research, particularly in situations where there is a need to be economical regarding item count and respondent burden. Apart from traditional data collection approaches where respondent burden is often an issue, the field is turning increasingly to high frequency data collection methods (e.g., ecological momentary assessment) to more fully capture the temporal nuances of the smoking experience (Shiffman, 2009). This approach is inherently burdensome, and researchers are often obliged to compromise measurement precision by assessing constructs with only a very few items. The ability to obtain comparable scores from a variety of short assessments administered across multiple platforms will do much to improve the status quo in this regard.

INITIAL PHASES OF SMOKING ASSESSMENT TOOLKIT DEVELOPMENT

Our formulation of the conceptual model and study design for the smoking assessment toolkit was influenced by the increasing rates of nondaily smokers and corresponding interest in understanding this subgroup of smokers. To ensure that the assessments would be relevant for both daily and nondaily smokers we explicitly included sufficient numbers of both types of smokers in each phase of development to allow adequate representation. Further, our analytic procedures were designed

such that any substantial differences between daily and non-daily smokers in the expression of the smoking domains could be identified and accommodated. Thus distinct results for daily and nondaily smokers are referenced throughout this supplement. A recent paper by Edelen, Tucker, Shadel, and Stuckey (2012) describes in detail the initial phases of toolkit development. We provide a brief overview here to orient the reader to the starting point for the analyses reported in this supplement.

Development of Item Pool for Field Testing

An initial pool of cigarette smoking items was selected for possible inclusion in the item banks using a qualitative item development phase (DeWalt, Rothrock, Yount, & Stone, 2007). This phase included a systematic literature review, which established a preliminary pool of 1,622 items from widely used measures, binning and winnowing of items to reduce redundancies and exclude items that were outside the intended scope of the banks (e.g., items about smoking during recovery from illicit substance use), item standardization with respect to response categories, time frame and person orientation, solicitation of feedback from daily and nondaily cigarette smokers via focus groups to identify any gaps in content coverage and cognitive interviews to ensure comprehensibility of items. This qualitative item development process resulted in 277 items representing eleven key conceptual domains relevant to current smokers (e.g., dependence, health concerns, positive smoking experiences, social influences, temptations to smoke).

Identification of Item Bank Domains

The item pool was administered to a large nationally representative sample of daily ($N = 4,201$) and nondaily ($N = 1,183$) smokers (total $N = 5,384$). All respondents completed thirteen of the 277 smoking items which assessed their smoking behavior and quitting history. The remaining 264 items were candidate items that were being considered for inclusion in one of the smoking item banks. These items were distributed across 26 overlapping forms containing an average of 147 items (range = 134–158); each respondent was randomly assigned one of the 26 forms.

We conducted a series of exploratory factor analyses (EFA) of the 264 items using IRTPRO's (Cai, du Toit, & Thissen, 2011) EFA module and data from the daily smokers. The goal of these analyses was to identify distinct groups of items representing key smoking behavior domains. At this stage of development, we assumed no differences between daily and nondaily smokers. Subsequent analyses to refine the content of the domains (described in other papers in this supplement) allowed for daily/nondaily differences to emerge. After close examination of the EFA results, we ultimately selected a 19-factor model to characterize the relationships among the items, as this solution was judged to have the most meaningful substantive content and useful factors. However, a number of the factors were highly correlated, which would limit their utility as distinct item banks. To arrive at a final set of factors that would later form the basis for the item banks, we merged a number of factors with relatively high factor intercorrelations (about $r = .7$, for most). This process resulted in the content of the six item banks which at this stage were labeled smoking dependence/craving (55 items), coping aspects of smoking (30 items), positive consequences of smoking (40 items), health consequences of smoking (26 items), psychosocial consequences of smoking (37 items), and social factors of smoking (23 items).

CONTENTS OF THIS SUPPLEMENT

The papers in this supplement describe the final phase of item bank development, which involved extensive psychometric analyses including evaluations of local dependence (Chen & Thissen, 1997), estimation of exploratory multidimensional IRT models (Cai, 2010), and confirmatory item bifactor models (Cai, Yang, & Hansen, 2011; Gibbons & Hedeker, 1992), testing for item bias (i.e., DIF; Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Holland & Wainer, 1993), performing concurrent calibrations of the item banks for the daily and nondaily smoker groups using a nonequivalent anchor test design (Dorans, 2007), developing short form representations of the item banks, and simulating the properties of the CAT assessment. The papers describing the derivation of each bank provide some description of these analyses. However, due to their sophistication and in the interest of avoiding unnecessary redundancy, these methodologic steps are described in detail in the first report of this supplemental issue. This is followed by six reports describing development of each of the item banks in turn. Each individual bank paper includes a discussion of the theoretical basis for the bank's content, a brief explanation of the analyses conducted to finalize the item bank contents, a list of the actual items in the bank, a description of the bank's psychometric properties (including CAT performance), identification of a subset of items suggested for use as a short form, and a scoring translation table for the short form.

Overall, psychometric evidence strongly supports the utility of the PROMIS Smoking Initiative item bank development process. This toolkit of item banks, short forms, and CATs provides researchers and clinicians with an array of highly reliable approaches to assessment of key smoking domains. The item banks for daily smokers contain from 12 to 27 items each and produce scores with reliability estimates above 0.90 for a wide range of each cigarette smoking domain continuum. As can be seen in Table 1, short form and CAT administrations provide additional means of achieving high reliability (generally greater than 0.85) using very few items (4–7 items for most banks), with nondaily banks performing similarly. In addition, preliminary validity evidence, reported in the final full length paper in this supplement, indicates that the six item banks are differentially associated with smoking and quitting patterns and are related as expected to domains of health-related quality of life. For example, nicotine dependence is most strongly associated with smoking quantity and time to first cigarette of the day; health and psychosocial expectancies are most related to quitting recency and interest, and coping expectancies are strongly associated with anxiety.

HOW THE BANKS CAN BE ACCESSED AND USED

The item banks in the smoking assessment toolkit have recently been incorporated into the PROMIS item library,

Table 1. Psychometric Properties of 6 Item Banks for Daily Smokers

Bank	Number of items			Marginal reliability		
	Full	SF	CAT[a]	Full	SF	CAT[a]
Nicotine Dependence	27	8/4	4.7	.97	.91/.81	.91
Coping Expectancies of Smoking	15	4	4.3	.96	.85	.91
Emotional and Sensory Expectancies of Smoking	16	6	7.5	.95	.86	.90
Health Expectancies of Smoking	19	6	5.3	.96	.87	.91
Psychosocial Expectancies of Smoking	16	6	6.4	.95	.85	.90
Social Motivations for Smoking	12	4	9.7	.90	.77	.88

Note. CAT = computer adaptive test; SF = short form.
[a]CAT item count and reliability are averages based on simulation results with maximum number of administered items set to 10.

and assessments based on these banks (e.g., suggested or tailored short forms, CATs) can be accessed via Assessment Center (www.assessmentcenter.net). The banks are also available for download from the PROMIS Smoking Initiative project web site, hosted at http://www.rand.org/health/projects/promis-smoking-initiative.html. This site provides background and psychometric information on the smoking assessment toolkit, access to item banks, short forms, and CAT assessment (through RAND's MMIC), links to published papers on the topic, summaries of ongoing analytic results, and future research directions.

Although we did identify some differences in domain expression according to daily/nondaily smoker status, as is clear from the results reported in the remainder of this supplement the differences are not as substantial as we expected. For each assessment domain, we generated distinct item banks according to daily/nondaily smoker status. However, the majority of items and item properties in the banks for a given domain are identical, with only a handful of items that are specific to either daily or nondaily smokers. This high correspondence in item content and properties allowed us to develop short forms for each domain that can be used for all smokers, regardless of daily/nondaily status. Thus, although researchers who wish to use the CAT option for assessment will need to incorporate the daily/nondaily distinction into their design, the domain short forms can be used "off the shelf," without concern for the smoker status of the sample. Further, scoring tables for tailored short forms can be generated to be applicable across daily and nondaily smokers upon request.

An important feature of the IRT-based item banking approach is that it represents a sustainable measurement solution for any given domain of behavioral research. Not only are the various scores generated based on this assessment toolkit comparable to one another, it is also straightforward to relate these scores back to existing measures of similar constructs (i.e., legacy measures) and to incorporate new items and subdomains into the system without creating the problem of "version control." We hope that these attractive measurement features will encourage smoking researchers to utilize the new smoking assessment toolkit, either in part or in full, and either instead of or in addition to use of other existing smoking measures.

FUNDING

DECLARATION OF INTERESTS

ACKNOWLEDGMENTS

REFERENCES

Ader, D. N. (2007). Developing the patient-reported outcomes measurement information system (PROMIS). *Medical Care*, *45*, S1–S2. Retrieved from http://uwcorr.washington.edu/publications/developing_ader.pdf.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612. doi:10.1007/s11336-010-9178-0

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling*. Chicago, IL: Scientific Software International.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*, 221–248. doi:10.1037/a0023350

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., … PROMIS Cooperative Group. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care*, *45*, S3–S11. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17443116.

Chen, W. H. & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289. Retrieved from www.jstor.org/stable/1165285.

DeWalt, D., Rothrock, N., Yount, S., & Stone, A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care*, *45*, S12–S21. Retrieved from http://dx.doi.org/10.1097/01.mlr.0000254567.79743.e2.

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, *16*, 85–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17286198.

Edelen, M., Tucker, J., Shadel, W. G., & Stuckey, B. (2012). Toward a more systematic assessment of smoking: Development of a smoking module for PROMIS®. *Addictive Behaviors*, *37*, 1278–1284. doi:10.1097/01.mlr.0000245251.83359.8c

Edelen, M. O., & Reeve, B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*, 5–18. doi:10.1007/s11136-007-9198-0

Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Application to the Mini-Mental Status Examination. *Medical Care*, *44*, S134–S142. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17060820.

Embretson, S. E. (1996). The new rules of measurement. *Psychol Assessment*, *8*, 341–349. doi:10.1037/1040-3590.8.4.341

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fries, J. F., Bruce, B., & Cella, D. (2005). The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. *Clinical and Experimental Rheumatology*, *23*, S53–S57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16273785.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436. doi:10.1007/BF02295430

Hahn, E. A., Cella, D., Bode, R. K., Gershon, R., & Lai, J.-S. (2006). Item banks and their potential applications to health status assessment in diverse populations. *Medical Care*, *44*(11 Suppl. 3), S189–S197. Retrieved from http://www.scholars.northwestern.edu/pubDetail.asp?t=&id=33750351039&.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: LEA.

Jones, L. V., and Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26: Psychometrics* (pp. 1–27). New York, NY: Elsevier.

Linn, R. L. (1992). *Linking results of distinct assessments*. Center for Research on Evaluation, Standards, and Student Testing, University of Colorado at Boulder, Boulder, CO. 6/11/92, 2nd Draft.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Earlbaum.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Educational Testing Services, Princeton, NJ. December 1992, Policy Issue/Perspective.

Panter, A. T., & Reeve, B. B. (2002). Assessing tobacco beliefs among youth using item response theory models. *Drug and Alcohol Dependence*, *68* (Suppl. 1) S21–S39. Retrieved from http://www.drugandalcoholdependence.com/article/S0376-8716(02)00213-2/fulltext.

Shadel, W. G., Edelen, M., & Tucker, J. S. (2011). A unified framework for smoking assessment: The PROMIS Smoking Initiative. *Nicotine & Tobacco Research*, *13*, 399–400. doi:10.1093/ntr/ntq253

Shiffman, S. (2009). Ecological momentary assessment (EMA) in studies of substance use. *Psychological Assessment*, *21*, 486. doi:10.1037/a0017074

Teresi, J. A. (2006). Overview of quantitative measurement methods. Equivalence, invariance, and differential item functioning in health applications. *Medical Care*, *44*(Suppl. 3), S39–S49. doi:10.1097/01.mlr.0000245452.48613.45

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: LEA.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 61–100). Mahwah, NJ: LEA.