Nicotine & Tobacco Research, 2016, 361–365
doi:10.1093/ntr/ntv083
Brief report
Advance Access publication April 8, 2015

Brief report

The Psychometric Performance of the PROMIS Smoking Assessment Toolkit: Comparisons of Real-Data Computer Adaptive Tests, Short Forms, and Mode of Administration

Brian D. Stucky PhD, Wenjing Huang PhD, Maria Orlando Edelen PhD

Department of Health, RAND Corporation, Santa Monica, CA

Corresponding Author: Brian D. Stucky, PhD, Department of Health, RAND Corporation, 1776 Main Street, Santa Monica, CA 90407-2138, USA. Telephone: 310-393-0411, x6336; Fax: 310-393-4818; E-mail: bstucky@rand.org

Abstract

Introduction: The PROMIS Smoking Initiative has developed six item banks for assessment related to cigarette smoking among adult smokers (Nicotine Dependence, Coping Expectancies, Emotional and Sensory Expectancies, Health Expectancies, Psychosocial Expectancies, and Social Motivations). This article evaluates the psychometric performance of the banks when administered via short form (SF), computer adaptive test (CAT), and by mode of administration (computer vs. paper-and-pencil).

Methods: Data are from two sources: an internet sample ($N = 491$) of daily and nondaily smokers who completed both SFs and CATs via the web and a community sample ($N = 369$) that completed either paper-and-pencil or computer administration of the SFs at two time points. First a CAT version of the PROMIS Smoking Assessment Toolkit was evaluated by comparing item administration rates and scores to the SF administration. Next, we considered the effect of computer versus paper-and-pencil administration on scoring and test-retest reliability.

Results: Across the domains approximately 5.4 to 10.3 items were administered on average for the CAT. SF and CAT item response theory-scores were correlated from 0.82 to 0.92 across the domains. Cronbach's alpha for the four- to eight-item SFs among daily smokers ranged from .80 to .91 and .82 to .91 for paper-and-pencil and computer administrations, respectively. Test-retest reliability of the SFs ranged from 0.79 to 0.89 across mode of administration.

Conclusions: Results indicate that the SF and CAT and computer and paper-and-pencil administrations provide highly comparable scores for daily and nondaily smokers, but preference for SF or CAT administration may vary by smoking domain.

Introduction

The Patient Reported Outcomes Measurement Information System (PROMIS) Smoking Assessment Toolkit consists of six item banks that assess unique smoking behavior domains: Nicotine Dependence,[1] Coping Expectancies,[2] Positive Emotional and Sensory Expectancies,[3] Health Expectancies,[4] Psychosocial Expectancies,[5] and Social Motivations for Smoking.[6] The domains were identified following an extensive qualitative process[7] and were then further psychometrically refined using modern psychometric techniques (eg, item-factor analysis and item response theory [IRT]) to ensure that the item sets were unidimensional, provided a high level of scoring precision, and measured distinct content.[8,9] This process resulted in the development of fixed-length short forms (SFs) ranging in length from 4 to 8 items, and dynamic computer adaptive tests (CATs) for daily and nondaily smokers.

The CAT engine employed in the PROMIS Smoking Assessment Toolkit uses an algorithm that in real-time selects and administers the most informative item from the bank of items based on responses to prior questions. This process provides a uniquely tailored set of items for each participant, and commonly results in higher score reliability or shorter survey administrations.

A potential limitation of the Toolkit is that the CATs and SFs were developed based on an online sample of participants who responded to only random subsets of the items from each domain. Thus the SFs were not administered as a complete set, the performance of the CATs are known only through simulations, the impact of internet administration is unknown, and test-retest reliability has not yet been established. To address these limitations, this article presents results from new data collections investigating: (1) the utility of real-data CATs compared to SFs (Study 1); (2) the effect of mode of administration (ie, paper-and-pencil or internet) on participants' responses and SF scores (Study 2); and (3) test-retest reliability of the SFs (Study 2).

Methods

Data Collection, Procedures, and Demographics

Eligibility criteria were the same for Studies 1 and 2: participants were current smokers, 18 years or older, who did not have plans to quit smoking in the next 6 months. Smokers were classified as "daily smokers" if they had smoked on 28–30 of the past 30 days, and as "nondaily smokers" if they smoked less than 28 of the past 30 days.

Study 1 participants (N = 491) were a subset of smokers from the original calibration data collection[10] who were recontacted through Harris Interactive's online panel via the internet and received points for their participation that are redeemable for gifts. Participants were routed from the Harris Interactive website to the PROMIS Assessment Center website where they completed a CAT administration of each smoking domain followed by any remaining SF items that were not administered via the CAT. The Study 1 sample had a mean age of 48, was 53% female, 30% had completed a 4-year college, 76% were non-Hispanic white, and 64% of the sample had smoked more than 10 cigarettes per day.

To assess the potential for selection bias among Study 1 participants, we also compared their demographic characteristics to the subset of participants in the original calibration data collection who did not participate in Study 1. Briefly, there were no significant differences between the samples for gender, daily smoker status, employment status, or education. There were significant differences in age at the time of the original data collection (calibration sample not participating in Study 1 mean age = 46 compared to 48 in Study 1; t = 3.51, P < .05) and minority status (68% of calibration sample not participating in Study 1 were non-Hispanic white compared to 76% for Study 1; χ^2 (1) = 15.7, P < .05).

Study 2 participants (N = 369) were recruited via flyers and advertising (eg, craigslist, campus newspapers) at various community venues in several large US cities. Participants were randomly assigned to complete either a paper-and-pencil or internet-based version of the survey containing SFs for all domains (N_{paper} = 192, $N_{internet}$ = 177). They were paid $5 for completing a screening instrument and $25 for completing a survey containing the smoking domain SFs. To assess test-retest reliability of the SFs, a subset of participants were recontacted approximately 1 week after baseline to complete a follow-up survey using the same mode of administration that was used in the baseline survey (N_{paper} = 113, $N_{internet}$ = 106). The Study 2 sample had a mean age of 44, was 50% female, 22% had completed a 4-year college, 31% were non-Hispanic white, and 45% of the sample had smoked more than 10 cigarettes per day.

Measures

Study 1 participants received CAT and SF representations of the six item banks; Study 2 participants completed SFs. The Nicotine Dependence domain (four- and eight-item SFs) measures craving, withdrawal, and smoking temptations (Marginal reliability (MR) = 0.81 and 0.91 for the four- and eight-item SFs, respectively.[1] MR is an IRT approach to summarizing the precision of scores across the latent continuum. Values range from 0 to 1 and can be interpreted somewhat similarly to Cronbach's coefficient alpha.). The Coping Expectancies domain (four-item SF) assesses the degree which smoking is used to cope with negative affect and stress (MR = 0.85[2]). The Emotional and Sensory Expectancies domain (six-item SF) measures positive affective states and pleasurable sensations due to smoking (MR = 0.86[3]). The Health Expectancies domain (six-item SF) measures perceptions of the health consequences of smoking (MR = 0.87[4]). The Psychosocial Expectancies domain (six-item SF) measures disapproval and normative values associated with smoking, and the negative beliefs about one's appearance when smoking (MR = 0.85[5]). The Social Motivations domain (four-item SF) measures anticipated social benefits of smoking (MR = 0.80[6]). All smoking domain IRT-scores for Study 1 and 2 were transformed onto a T-score metric with a mean of 50 and SD of 10.

Analytic Approach

Study 1: Evaluating Differences in Performance of CAT and SF Administration

First, items were administered via CATs until either a level of score precision equal to a reliability of 0.90 was reached, or until the maximum number of items (12) was reached. When either criterion was satisfied all remaining SF items were administered that were not presented during the CAT. This yielded two sets of IRT scores: response pattern expected a posteriori for the CATs[11] and summed score-to-IRT score expected a posteriori for the SFs.[12] Within-subjects differences between the CAT and SF IRT-scores were evaluated across domains using correlations, t tests, and effect size estimates (ie, Cohen's d). In addition, we report average length of the CAT and the proportion of the SF items that were administered on average via the CAT (eg, 1.0 indicates all the SF items were administered via the CAT; 0.5 indicates that on average one-half of the SF items were administered via the CAT). Finally, the MR of the CATs and SFs is reported.

Study 2: Evaluate the Effect of Mode of Administration on Domain-level Scores and Score Reliability

First, reliability estimates were obtained separately for the paper-and-pencil and internet-based administrations using test-retest correlations and Cronbach's alpha. Next, separate analysis of variances were evaluated for each domain that included as predictors mode (between-subjects), daily/nondaily smoker classification, and the interaction between both predictors. Standard effect size calculations are presented for the effect of mode. Finally, we used a Structural Equation Modeling approach to simultaneously test the overall effect of mode of administration among the combined sample of daily and nondaily smokers.[13] This joint test is operationally equivalent to a two-group (ie, paper-and-pencil vs. internet-based administration) Structural Equation Modeling analysis with invariance constraints imposed on the group means and covariances.

Results

Study 1

The top of Table 1 displays the differences in means between the CAT and SF IRT-based T-scores. Using dependent samples *t* tests, score differences were statistically significant for five of the seven comparisons. However, with the exception of Coping Expectancies, effect size estimates indicate that the standardized mean differences between the domains were small at 0.10 *SD*s or less, suggesting that the CATs and SFs result in very similar scores. Consistent with this finding, the correlations between the scoring systems were high, ranging from 0.86–0.92.

As anticipated (based on the CAT stopping rule), the MRs of the CATs were generally greater than 0.90, while the SF reliabilities ranged from 0.80 to 0.88 (the reliability for the four-item version of Nicotine Dependence is 0.76). For illustrative purposes Figure 1 shows the difference between the reliability of CAT and SF T-scores across the range of the Nicotine Dependence score. The eight-item SF is similar in reliability to the CAT at around ±1 *SD* of the mean (50); however, in the tails of the distribution the CAT provides greater reliability as a result of selecting more targeted items that measure higher (or lower) levels of nicotine dependence.

Table 1 also provides the CAT item administration rates in comparison to the SFs. The CATs administered slightly more items than that are present on the SFs (mean CAT length across domains = 5.2 to 10.3 items). Further, the CATs on average administered from 10% to 88% of the SF items (Table 1). Notably, Nicotine Dependence and Coping Expectancies had the lowest proportions of SF items administered (from 10% to 13%, respectively) and the lowest number of average items administered (5.4 and 5.2, respectively), while Social Motivations for Smoking had the highest proportion of SF items administered (0.88) and the highest average number of items administered (10.3).

Study 2

Reliability

The SF test-retest reliabilities range from 0.79 to 0.87 for the paper-and-pencil administration and 0.82–0.89 for the internet-based administration (see bottom of Table 1). Estimates of internal consistency were also similar across the domains with Cronbach's alphas ranging from 0.83 to 0.92 and 0.83–0.91 for the paper-and-pencil and internet-based administrations, respectively. Reliability estimates for the daily and nondaily smoker subgroups were largely comparable across domain (available from the first author).

Effect of Mode

The effect of mode of administration on SF domain scores was first evaluated by smoker classification using a two-way (mode × smoker-type) analysis of variance for each domain. Across four of the six analysis of variances, domain scores were significantly higher at $P < .05$ for daily smokers (the exceptions were Health and Psychosocial Expectancies). The effect of mode was only significant for the eight-item Nicotine Dependence (means = 50.6 and 48.0, for paper-and-pencil and internet administration, respectively; effect size = 0.26). All other domain comparisons resulted in effect sizes from 0.00 to 0.13. There were no significant interactions between mode and smoker classification.

Mode of administration was next assessed in a two-group (paper-and-pencil vs. internet administration) Structural Equation Modeling model with equality constraints imposed on the means and covariances of the SF domain scores for both groups. The close model fit indicates that mode of administration does not affect the domain means or intercorrelations ($\chi^2 = 51.8$, $df = 27$, $P = .01$; RMSEA = 0.067; TLI = 0.96; CFI = 0.98),

Discussion

This article presents psychometric evidence supporting the equivalence of SF and CAT administration options as well as paper-and-pencil

Table 1. Comparisons of CAT and Short Form Psychometric Properties Across a Community and Follow-up Sample

	Nicotine Dependence (eight-item)	Nicotine Dependence (four-item)	Coping Expectancies	Positive Emotional and Sensory Expectancies	Health Expectancies	Psychosocial Expectancies	Social Motivations for Smoking
Study 1: CAT vs. SF validity							
Bank length/SF length	32/8	32/4	21/4	18/6	24/6	21/6	15/4
CAT mean (*SD*)	47.2 (10.9)	47.2 (10.9)	49.9 (10.1)	50.1 (9.7)	48.9 (9.8)	46.8 (10.1)	50.0 (9.7)
SF mean (*SD*)	47.3 (10.9)	48.2 (10.2)	47.5 (9.9)	50.1 (9.2)	48.1 (8.8)	46.1 (9.1)	50.9 (9.1)
t-value (*df*)	0.5 (485)	3.9 (485)*	−10.5 (486)*	0.3 (480)	−4.2 (480)*	−3.9 (480)*	5.2 (488)*
Effect size	−0.01	−0.09	0.24	0.00	0.08	0.07	−0.10
CAT and SF correlation	0.90	0.86	0.87	0.92	0.92	0.92	0.91
Study 1: CAT vs. SF reliability							
CAT marginal reliability	0.92[a]	0.92[a]	0.93	0.91	0.92	0.91	0.89
SF marginal reliability	0.88	0.76	0.85	0.86	0.88	0.84	0.80
Mean CAT length (SF length)	5.4 (8)	5.4 (4)	5.2 (4)	7.5 (6)	5.8 (6)	7.6 (6)	10.3 (4)
Proportion of SF items administered as CAT	0.10	0.13	0.13	0.48	0.43	0.46	0.88
Study 2: SF mode and reliability							
Test-retest: PP	0.87	0.82	0.85	0.79	0.80	0.81	0.81
Test-retest: Web	0.88	0.85	0.83	0.84	0.89	0.85	0.82
Cronbach's alpha: PP	0.92	0.84	0.89	0.87	0.84	0.83	0.90
Cronbach's alpha: Web	0.91	0.85	0.83	0.88	0.89	0.86	0.90

CAT = computer adaptive test; *df* = degrees of freedom; SF = short form; PP = paper and pencil administration; Web = internet-based administration.

[a]CAT marginal reliability for Nicotine Dependence is identical in both comparisons.

*Indicates values statistically significant at $P < .01$.

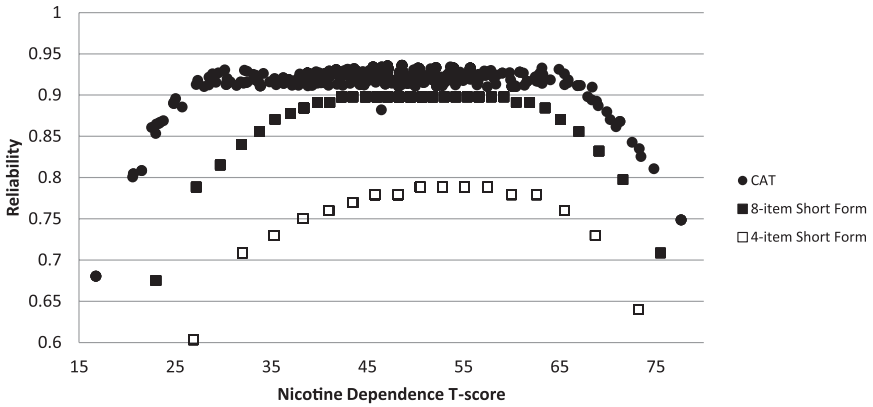

Figure 1. Comparisons of the Nicotine Dependence CAT, four-item short form, and eight-item short form administrations.

and internet-based administration modes. Regarding mode of administration, the relatively minor differences between paper-and-pencil and internet are consistent with findings from prior meta-analytic reviews that also identified very small effect size differences between paper-and-pencil and computerized tests.[14–16] However, we also note that the internet mode for Nicotine Dependence resulted in a slightly lower than expected mean score. Echoing the recommendations from Green, Bock, Humphreys, Linn, and Reckase[17] that computer-based and paper-and-pencil administered tests are only equivalent following empirical evaluation, future administrations of this domain are needed to determine if this finding reflects a true difference in administration modes. Regarding reliability differences across mode, there was a slight tendency for higher reliability among the internet-based administrations of the Psychosocial and Health Expectancies domains, suggesting that internet-based responses may be more consistent.

Study 1 results generally support the utility of both the CAT and SF administration options. There were some differences in CAT performance across the domains. While the Social Motivations for Smoking CAT scores were more reliable on average than the SF option, the CAT also required nearly the entire bank of items, calling into question the utility of the CAT option for this domain. In contrast, the Nicotine Dependence and Coping Expectancies domains appear to be well-suited for CAT administration. Both domains required only about five items to achieve an average reliability of 0.90. The preference for the CAT option is apparent when evaluating the tails of the domain score distribution. For example, Figure 1 indicates that the Nicotine Dependence CAT provides score reliabilities greater than 0.90 for T-score values from 25 to about 70. In contrast the eight-item SF has a range limited to about 40 to 60. These results confirm that the CAT is tailoring the administration of items to account for participants' true level of dependence, and suggests that researchers interested in measuring highly dependent smokers or those who have recently started smoking, may benefit from the CAT administration option. Further, the difference between CAT and SF reliability is even more apparent when considering the four-item Nicotine Dependence SF (MR = 0.76). The relatively low MR and lack of reliability coverage (Figure 1) suggests that researchers administering the Nicotine Dependence domain may benefit from using either the eight-item SF (MR = 0.88) or CAT administration options.

The correlations between the SFs and CATs across domains, though large (range = 0.86–0.92) were lower than expected. For example, the pediatric PROMIS project reported CAT and SF correlations from 0.93 to 0.98 across eight quality of life domains.[18] However, the modest correlations reported here are likely an artifact of the smoking domain CATs administering items from the banks that do not appear on the SFs. As indicated in Table 1, with the exception of the Social Motivations domain, on average only 10% to 48% of the SF items were administered via the CAT. By contrast 66% to 88% of the pediatric PROMIS domain SF items were administered via the CAT, indicating that the CATs and SFs from the pediatric PROMIS project produce overlapping scores. The discrepancy in the proportions of the SF items administered via the CAT in this application is an artifact of how the SF items were selected and the lengths of the SFs. The pediatric PROMIS project utilize slightly longer SFs (8–10 items) and selected SF items in order to maximize reliability at various locations along the latent continuum[19]; the PROMIS Smoking Assessment project utilizes shorter-length SFs (4–6 items with the exception of the longer version of Nicotine Dependence) and, in addition to ensuring adequate SF reliability, emphasized content coverage and expert opinion when selecting itmes,[9] which leads to SFs that are somewhat more diverse in content than may be present via CAT administration. It speaks to the versatility of the PROMIS Smoking Assessment item banks that correlations between the SF and CAT are relatively high given that they include largely different items in their administration.

In light of these findings, future studies of smoking behaviors should consider the particular utility of CAT administration as a means of obtaining highly precise scores with relatively few items. At present the PROMIS Smoking Assessment Toolkit is the only smoking behavior measurement system that utilizes CAT technology. The results presented here suggest that both CAT and SF options have strong psychometric characteristics; evidence for the utility of the domains will emerge as they are adopted by the smoking behavior research community.

Funding

This work was supported by the National Institute on Drug Abuse (R01DA026943; PI: MOE).

Declaration of Interests

None declared.

References

1. Shadel WG, Edelen MO, Tucker JS, Stucky BD, Hansen M, Cai L. Development of the PROMIS® Nicotine Dependence Item Banks. *Nicotine Tob Res*. 2014;16(suppl 3):S190–S201. doi:10.1093/ntr/ntu032.

2. Shadel WG, Edelen MO, Tucker JS, Stucky BD, Hansen M, Cai L. Development of the PROMIS® Coping Expectancies of Smoking Item Banks. *Nicotine Tob Res*. 2014;16(suppl 3):S202–S211. doi:10.1093/ntr/ntu040.
3. Tucker JS, Shadel WG, Edelen MO, et al. Development of the PROMIS® Positive Emotional and Sensory Expectancies of Smoking Item Banks. *Nicotine Tob Res*. 2014;16(suppl 3):S212–S222. doi:10.1093/ntr/ntt281.
4. Edelen MO, Tucker JS, Shadel WG, et al. Development of the PROMIS® Health Expectancies of Smoking Item Banks. *Nicotine Tob Res*. 2014;16(suppl 3):S223–S231. doi:10.1093/ntr/ntu053.
5. Stucky BD, Edelen MO, Tucker JS, et al. Development of the PROMIS® Negative Psychosocial Expectancies of Smoking Item Banks. *Nicotine Tob Res*. 2014;16(suppl 3):S232–S240. doi:10.1093/ntr/ntt282.
6. Tucker JS, Shadel WG, Edelen MO, et al. Development of the PROMIS® Social Motivations for Smoking Item Banks. *Nicotine Tob Res*. 2014;16(suppl 3):S241–S249. doi:10.1093/ntr/ntt283.
7. Edelen MO, Tucker JS, Shadel WG, Stucky BD, Cai L. Toward a more systematic assessment of smoking: development of a smoking module for PROMIS®. *Addict Behav*. 2012;37(11):1278–1284. doi:10.1016/j.addbeh.2012.06.016.
8. Edelen MO, Stucky BD, Hansen M, Tucker JS, Shadel WG, Cai L. The PROMIS® smoking initiative: initial validity evidence for six new smoking item banks. *Nicotine Tob Res*. 2014;16(suppl 3):S250–S260. doi:10.1093/ntr/ntu065.
9. Hansen M, Cai L, Stucky BD, Tucker JS, Shadel WG, Edelen MO. Methodology for Developing and Evaluating the PROMIS® Smoking Item Banks. *Nicotine Tob Res*. 2014;16(suppl 3):S175–S189. doi:10.1093/ntr/ntt123.
10. Edelen MO. The PROMIS® Smoking Assessment Toolkit—Background and Introduction to Supplement. *Nicotine Tob Res*. 2014;16(suppl 3):S170–S174. doi:10.1093/ntr/ntu086.
11. Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. *Appl Psych Meas*. 1982;6(4):431–444. doi:10.1177/014662168200600405.
12. Thissen D, Pommerich M, Billeaud K, Williams VS. Item response theory for scores on tests including polytomous items with ordered responses. *Appl Psych Meas*. 1995;19(1):39–49. doi:10.1177/014662169501900105.
13. Preacher KJ. Quantifying parsimony in structural equation modeling. *Multivar Behav Res*. 2006;41(3):227–259. doi:10.1207/s15327906mbr4103_1.
14. Mead AD, Dasgow F. Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. *Psychol Bull*. 1993;114(3):449. doi:10.1037/0033-2909.114.3.449.
15. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review. *Value Health*. 2008;11(2):322–333. doi:10.1111/j.1524-4733.2007.00231.x.
16. Wang S, Jiao H, Young MJ, Brooks T, Olson J. Comparability of computer-based and paper-and-pencil testing in K–12 Reading Assessments. A meta-analysis of testing mode effects. *Educ Psychol Meas*. 2008;68(1):5–24. doi:10.1177/0013164407305592.
17. Green BF, Bock RD, Humphreys LG, Linn RL, Reckase MD. Technical guidelines for assessing computerized adaptive tests. *J Educ Meas*. 1984;21(4):347–360. doi:10.1111/j.1745-3984.1984.tb01039.x.
18. Varni JW, Magnus B, Stucky BD, et al. Psychometric properties of the PROMIS® pediatric scales: precision, stability, and comparison of different scoring and administration options. *Qual Life Res*. 2014;23(4):1233–1243. doi:10.1007/s11136-013-0544-0.
19. Yeatts KB, Stucky BD, Thissen D, et al. Construction of the pediatric asthma impact scale (PAIS) for the patient-reported outcomes measurement information system (PROMIS). *J Asthma*. 2010;47(3):295–302. doi:10.3109/02770900903426997.