Journal of Computerized Adaptive Testing

Volume 1 Number 3

April 2013

Item Ordering in Stochastically Curtailed Health Questionnaires With an Observable Outcome

Matthew D. Finkelman, Wonsuk Kim, Yulei He, and Albert M. Lai

DOI 10.7333/1304-0103038

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing
Volume 1, Number 3, April 2013
DOI 10.7333/1304-0103038
ISSN 2165-6592

Item Ordering in Stochastically Curtailed Health Questionnaires With an Observable Outcome

Matthew D. Finkelman, ***Tufts University School of Dental Medicine***
Wonsuk Kim, ***Measured Progress***
Yulei He, ***Harvard Medical School***
Albert M. Lai, ***The Ohio State University***

Stochastic curtailment has been proposed as a method of shortening health questionnaires that predict an observable outcome. This study investigated whether the efficiency gains resulting from this approach can be enhanced by judiciously ordering the items within a questionnaire. Several new statistical procedures for ordering items are introduced and compared with an existing item ordering procedure, as well as with random orderings. In a post-hoc simulation using data from the Medicare Health Outcomes Survey, the orderings based on statistical criteria exhibited larger efficiency gains than the random item orderings. Comparisons between the different statistical methods depended on the simulation condition studied. Practical considerations are discussed.

Keywords: *curtailment, stochastic curtailment, computerized classification testing, variable-length testing, respondent burden.*

Over the past several decades, the use of questionnaires in psychiatric and medical research has become common practice. The percentage of biomedical publications citing the word *questionnaire* steadily increased from 1970 to 2000, paralleling the acceptance of self-report measures in clinical work (Walter, 2010). Instruments such as the CES-D (Radloff, 1977) and SF-36 (Ware & Sherbourne, 1992) have figured prominently in articles related to depression and health-related quality of life, respectively. Other assessments have been developed to detect diabetes mellitus type 2 (Ruige, de Neeling, Kostense, Bouter, & Heine, 1997), screen subjects for obstructive sleep apnea (Chung et al., 2008), and evaluate the physical functioning and mental well-being of Medicare beneficiaries (Haffer & Bowen, 2004), to name a few applications.

Because answering a large number of items can be burdensome to respondents, practitioners strive to make their health questionnaires as efficient as possible (Adams & Gale, 1982; Herzog & Bachman, 1981; Rogers, Wilson, Bungay, Cynn, & Adler, 2002; Scientific Advisory Committee of the Medical Outcomes Trust, 2002). One way to enhance efficiency is to employ *computerized classification testing* (CCT), an approach to assessment that originated in the literature of educational measurement. CCT involves the administration of items to respondents who are to be classified into one of multiple mutually exclusive categories (often two). Conducting the ques-

tionnaire by computer facilitates the use of variable-length testing, where the number of items presented is different from individual to individual (Thompson, 2007, 2011). In variable-length testing, interim analyses are conducted for each respondent while the questionnaire is underway. Once enough evidence has mounted in favor of one category, testing is terminated and the appropriate classification decision is made. By judiciously determining the stopping point for each respondent, variable-length tests can achieve low error rates, while reducing the average number of items presented (Thompson, 2007).

One of the most important components of a CCT is the method by which items are selected for presentation. In the context of latent variable measurement, Huebner (2012) noted that item selection for *classifying* a respondent into one of two categories is generally different from item selection for *estimating* the respondent's latent trait. In particular, the distinction between the two can be described as *sequential* versus *adaptive* item selection (Huebner, 2012; Thompson, 2007). Unlike the adaptive item selection methods that are commonly used when estimating the latent construct, item selection methods for classification often focus on obtaining information at or near the cut point between the two possible classifications (Eggen, 1999; Huebner, 2012; Huebner & Li, 2012; Spray & Reckase, 1994; Thompson, 2007, 2011). A typical CCT strategy is then to arrange the items in descending order of their ability to discriminate between values near the cut point, and therefore to discriminate between the possible categories (Huebner, 2012). Maximum Fisher information at the cut point (Spray & Reckase, 1994) and maximum Kullback-Leibler information around the cut point (Eggen, 1999) are two such methods for assessments that classify a latent variable into one of two categories. It is noteworthy that if one of these methods is adopted, the item selection process will yield identical selections for each respondent (Huebner, 2012; Thompson, 2011); of course, the *number* of items actually administered might differ between two respondents, due to the use of variable-length testing. The CCT item selection methods are referred to as *sequential* rather than *adaptive* because they administer a pre-determined list of items in sequence (until the termination criterion is satisfied), rather than adapting to the respondent's answers (Huebner, 2012; Thompson, 2007). Both the Fisher information and Kullback-Leibler information approaches have been studied for use alongside the sequential probability ratio test (SPRT), a variable-length testing termination criterion that originated in the sequential analysis literature (Wald, 1947) and was later investigated in the CCT setting (Reckase, 1983). Eggen and Straetmans (2000) and Weissman (2007) also examined item selection methods to be coupled with the SPRT. Other termination criteria that have been used to classify a latent variable include ability confidence intervals (Thompson, 2007, 2011; Weiss & Kingsbury, 1984), the generalized likelihood ratio test (Bartroff, Finkelman, & Lai, 2008; Thompson, 2011), and Bayesian decision theory (Lewis & Sheehan, 1990; Rudner, 2009; Vos, 2000); the termination criterion that is used might affect which item selection method exhibits the best efficiency (Thompson, 2011).

Although most prior research on CCT has focused on the classification of latent variables, computerized questionnaires can also be used to predict an *observable* outcome, such as the respondent's vital status (alive or deceased) after a follow-up period of two years. For the latter objective, Finkelman, He, Kim, and Lai (2011) proposed the use of stochastic curtailment, a sequential analysis method that was developed in the field of clinical trials (Betensky, 1997; Davis & Hardy, 1994; Lan, Simon, & Halperin, 1982; Leung, Wang, & Amar, 2003) and has also been applied in educational contexts (Finkelman, 2008, 2010). Stochastic curtailment dictates that a respondent's test be terminated if the future items scheduled for that respondent are unlikely to affect his or her final classification. A study using data from the Medicare Health Outcomes Survey (MHOS; Finkelman et al., 2011) showed that this method can reduce the respondent burden

of a questionnaire that predicts an observable outcome, while maintaining comparable sensitivity and specificity.

Finkelman et al. (2011) conjectured that stochastic curtailment's impact on predicting an observable outcome might be influenced by the arrangement of items—which they referred to as the *item ordering*—within a questionnaire. That is, they hypothesized that the efficiency of a stochastically curtailed test to predict an observable outcome depends on which item is presented first, which item is presented second, and so forth. However, they only examined one item ordering method, leaving the comparison of orderings as a topic of future work. As a result, it is currently unknown how to order a questionnaire's items so that the efficiency gains achieved by stochastic curtailment are maximized, in the case where the outcome is observable rather than latent. The extent to which statistical item ordering methods can outperform random orderings is also unknown when the outcome is observable. The aim of this study was to address these open questions by comparing different item ordering methods, both statistical and random, in terms of their efficiencies applied to a stochastically curtailed health questionnaire to predict an observable outcome.

One notable distinction between the measurement of a latent variable and the prediction of an observable outcome lies in the choice of statistical model to be used. When CCT is used to classify a latent variable, it is typically coupled with item response theory (IRT; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). When combining stochastic curtailment with the prediction of a dichotomous observable outcome, Finkelman et al. (2011) used logistic regression to model the data; logistic regression modeling was used in the current research as well. The item ordering methods compared herein are thus suitable for use alongside logistic regression, rather than IRT, and will also be generalizable to other statistical procedures that classify an observable outcome.

Stochastic Curtailment of Health Questionnaires

Suppose that an assessment is being used to predict an observable health outcome for each respondent who completes it. This prediction is based on a classification model that has been fitted to a *training dataset*, i.e., a dataset that is collected in order to conduct statistical modeling (as opposed to a *test dataset* that is used independently to evaluate the performance of the fitted model). The goal is to classify future respondents as efficiently as possible; some respondents might be administered the full set of items, while others receive a subset of it.

For example, consider a questionnaire that predicts what a respondent's vital status will be after a specified follow-up period. It is assumed that a logistic regression model has been fitted to training data; the dependent variable is vital status at follow-up (0 = alive, 1 = deceased), and the independent variables are the item responses at baseline. It is further assumed that items are either dichotomous (e.g., coded either 0 or 1), or ordinal (coded 1–3 for items with three options, 1–4 for items with four options, and so forth). Extensions to nominal or continuous items are straightforward, but are suppressed here for simplicity. Now let $\hat{\alpha}$ denote the estimated intercept of the logistic regression model, and $\{\hat{\beta}_i\}$ the estimated coefficients of the independent variables. Here, $i = 1,...,N$ indexes the order that items are presented in the questionnaire, e.g., $\hat{\beta}_1$ is the coefficient of the first item. Finally, let x_i denote the respondent's answer to item i. For respondents who receive all N items, the estimated probability of death before follow-up is given by the standard logistic regression equation

$$\hat{p} = \frac{\exp\left(\hat{\alpha} + \sum_{i=1}^{N} \hat{\beta}_i x_i\right)}{1 + \exp\left(\hat{\alpha} + \sum_{i=1}^{N} \hat{\beta}_i x_i\right)}. \quad (1)$$

An *at-risk* classification is made if $\hat{p} \geq p^*$, and a *not-at-risk* classification is made if $\hat{p} < p^*$. Here, p^* is a cut point set by the practitioner; it represents the level of probability that must be reached in order for an at-risk classification to be made. For example, if p^* is specified to be 0.10, then a respondent is classified as at-risk if and only if the logistic regression model estimates his or her chance of death before follow-up as 10% or higher. As noted in Finkelman et al. (2011), an equivalent classification rule is to label respondents as at-risk if and only if

$$\hat{\alpha} + \sum_{i=1}^{N} \hat{\beta}_i x_i \geq C, \quad (2)$$

where

$$C = \log\left[\frac{p^*}{\left(1 - p^*\right)}\right]. \quad (3)$$

Once the classification rule (Expression 2) has been specified for questionnaires that go the full length, stochastic curtailment can be used to determine early stopping. Suppose $k < N$ items have been administered to a given respondent, eliciting an answer vector $\mathbf{x}_k = (x_1, ..., x_k)$. Let $(X_{k+1}, ..., X_N)$ denote the respondent's future answers to items $k+1$ through N, assuming that all items will be presented (capital letters are used because future responses are treated as random variables). Using this notation, Expression 2 can be rewritten as

$$\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_i + \sum_{i=k+1}^{N} \hat{\beta}_i X_i \geq C. \quad (4)$$

As mentioned above, stochastic curtailment stipulates that the test be terminated if the respondent's future answers are unlikely to influence his or her classification. Therefore, the stopping rule is invoked if the conditional probability of Expression 4, given the data observed thus far, is very high or low. Rearranging the terms of Expression 4 to place only the random part on the left-hand side, the probability of interest is

$$P\left(\sum_{i=k+1}^{N} \hat{\beta}_i X_i \geq C - \hat{\alpha} - \sum_{i=1}^{k} \hat{\beta}_i x_i \mid \mathbf{x}_k\right), \quad (5)$$

where $P(Y \mid Z)$ represents the probability of event Y given Z.

Finkelman et al. (2011) proposed the following steps to estimate Expression 5 and decide whether to end the respondent's questionnaire after k items:

1. Among subjects in the training dataset who experienced the outcome of interest (here, death before follow-up), find the proportion for whom

$$\sum_{i=k+1}^{N} \hat{\beta}_i X_i \geq C - \hat{\alpha} - \sum_{i=1}^{k} \hat{\beta}_i x_i. \tag{6}$$

   This proportion, which is denoted $\hat{P}_{\mathrm{O}}$, can be obtained simply by examining the training subjects' responses to items $k + 1$ through N, taking the appropriate linear combination, and comparing to

$$C - \hat{\alpha} - \sum_{i=1}^{k} \hat{\beta}_i x_i. \tag{7}$$

   The latter expression takes into account the current respondent's vector of answers up to item k.
2. Similarly, among subjects in the training dataset who did not experience the outcome of interest, find the proportion for whom

$$\sum_{i=k+1}^{N} \hat{\beta}_i X_i \geq C - \hat{\alpha} - \sum_{i=1}^{k} \hat{\beta}_i x_i. \tag{8}$$

   This proportion is denoted $\hat{P}_{\mathrm{O}-}$
3. If both $\hat{P}_{\mathrm{O}}$ and $\hat{P}_{\mathrm{O}-}$ are greater than or equal to a specified constant γ_1, assessment is ceased and the current respondent is given an at-risk classification.
4. If both $\hat{P}_{\mathrm{O}}$ and $\hat{P}_{\mathrm{O}-}$ are less than or equal to a specified constant $1-\gamma_0$, assessment is ceased and the current respondent is given a not-at-risk classification. γ_1 and γ_0 are assumed to be greater than 0.5.
5. If neither of the conditional statements in Steps 3 or 4 is observed, another item is administered.

Step 1 can be thought of as a process whereby the current respondent's previous answers (items 1 to k) are concatenated with the future answers (items $k+1$ to N) of each subject in the training dataset who experienced the outcome. This process results in a set of "combined response patterns," the total number of which is equal to the number of training-set subjects experiencing the outcome. Each combined response pattern is then classified according to the logistic regression model; $\hat{P}_{\mathrm{O}}$ represents the proportion of such response patterns that receive an at-risk classification. For example, a value of $\hat{P}_{\mathrm{O}} = 0.97$ indicates that when concatenating the current respondent's previous answers with future answers of subjects experiencing the outcome, 97% of such concatenated response patterns result in an at-risk classification. Step 1 thus produces a nonparametric estimate of the probability that the current respondent will receive an at-risk classification, given his or her previous answers and assuming that his or her future answers will be similar to those of training-set subjects who experienced the outcome. Step 2 is similar to Step 1,

except that the current respondent's previous answers are now concatenated with the future answers of subjects in the training set who *did not* experience the outcome. $\hat{P}_{O-}$ represents the proportion of the combined Step 2 response patterns that receive an at-risk classification. For example, a value of $\hat{P}_{O-} = 0.84$ indicates that when concatenating the current respondent's previous answers with future answers of subjects *not* experiencing the outcome, 84% of such concatenated response patterns result in an at-risk classification. Step 2 thus estimates the same conditional probability as Step 1, but assuming that the respondent's future answers will be similar to training-set subjects who did not experience the outcome of interest.

Intuitively, if both estimated probabilities from Steps 1 and 2 ($\hat{P}_{O}$ and $\hat{P}_{O-}$) are either very high or very low, then the practitioner can be relatively confident of the respondent's classification, and testing can be halted. Therefore, in Steps 3 and 4, the estimated probabilities are compared to thresholds γ_1 and $1-\gamma_0$. The constant γ_1 can be considered as a "positive hurdle" value: for early stopping to occur in favor of an at-risk classification, this hurdle must be met or surpassed by both $\hat{P}_{O}$ and $\hat{P}_{O-}$. The constant $1-\gamma_0$ can be considered as a "negative hurdle" value: for early stopping to occur in favor of a not-at-risk classification, both $\hat{P}_{O}$ and $\hat{P}_{O-}$ must be at or below this hurdle. Setting γ_1 and γ_0 to high values (e.g., $\gamma_1 = \gamma_0 = 0.99$) indicates that the practitioner wishes to use a conservative stopping rule. It is also possible to specify different values for γ_1 and γ_0; for example, choosing $\gamma_1 = 0.75$ and $\gamma_0 = 0.95$ indicates that the practitioner wishes to be more conservative in making an early not-at-risk classification than an early at-risk classification. The special case $\gamma_1 = \gamma_0 = 1$ dictates that when the current respondent's previous answers are concatenated with future answers of training-set subjects, all such combined response patterns must have the same classification (at-risk or not-at-risk) in order for early stopping to occur. This special case is roughly equivalent to a rule whereby the test stops early only if the remaining items cannot possibly change the respondent's classification, regardless of his or her future answers. The latter stopping rule is called *curtailment*; see Eisenberg and Ghosh (1980) and Eisenberg and Simons (1978) for properties of curtailed tests.

Use of Steps 1–5 has been termed "stochastic curtailment via empirical proportions" (Finkelman et al., 2011). These steps not only delineate whether to stop early, but also which classification decision to make if early stopping does occur. For respondents whose tests do not stop early (i.e., respondents who receive the complete questionnaire), classification decisions are made via Expression 2.

Item Ordering Methods

Background and Previous Work

As explained above, the effectiveness of stochastic curtailment might depend on the ordering of items within a questionnaire. Intuitively, for maximal gains in efficiency to be achieved, items should be ordered by their value in predicting the outcome of interest. Those with greatest predictive value should be presented first, so that a classification can be made quickly and less informative items can be eliminated by early stopping. Using this logic, Finkelman et al. (2011) proposed that items be placed in the same order that they would be selected by a forward stepwise logistic regression. If forward stepwise logistic regression was not involved in the initial model selection process, it can be run for the sole purpose of ordering items in the questionnaire,

using a lenient enough entry criterion that all items are entered into the model.

The above method uses early selection by the stepwise logistic regression as a proxy for an item's ability to assist in making an efficient classification. Other statistical criteria can also be defined and used to order items.

New Criteria for Ordering Items

p Value from the multiple logistic regression model. It could be argued that an independent variable's *p* value in the final logistic regression model is more important than when it was added by a stepwise procedure. After all, the final model is the one that is actually used to classify respondents. Therefore, items can be placed in ascending order of their *p* values in the final model. If ties occur due to the rounding of *p* values, then items can be ordered by the test statistic that produced these *p* values. Note that there are several candidate test statistics for logistic regression, including the Wald statistic, likelihood-ratio statistic, and efficient score statistic (Agresti, 1996).

Standardized logistic regression coefficient. An independent variable's influence in predicting the outcome can also be assessed by its coefficient in the final model. Standardized coefficients are more appropriate than unstandardized coefficients in this context because the former are independent of the unit of measurement (Menard, 2004). One definition of a standardized coefficient in logistic regression is (Menard, 2004)

$$\hat{\beta}_i^* = s_i \hat{\beta}_i, \tag{9}$$

where s_i is the standard deviation of independent variable *i*. Items in a questionnaire can be placed in descending order of their $|\hat{\beta}_i^*|$ statistics; absolute values are taken because an item might be highly associated with the outcome regardless of whether its coefficient is positive or negative.

A variation on this method would be to use the range, rather than the standard deviation, as a measure of spread. That is, for each item, the following index is computed:

$$\hat{\beta}_i^{**} = (\max_i - \min_i)\hat{\beta}_i. \tag{10}$$

Here, $\max_i$ is the item's maximum observed value and $\min_i$ is its minimum observed value. Items in a questionnaire are then placed in descending order of their $|\hat{\beta}_i^{**}|$ statistics.

Item Ordering With Constraints

Each of the above item ordering methods focuses solely on the statistical properties of the items. In an operational questionnaire, however, the ordering of items might be subject to certain practical constraints. For example, a test designer might specify that items with similar content be administered consecutively, so that the assessment is coherent from the respondent's perspective. Additionally, items that are general in nature (e.g., that ask about demographic information) might be presented before items relating to more sensitive material (e.g., that ask about emotional health or life-threatening illnesses), so that respondents can "warm up" in the initial stages of the test. Many more constraints are possible.

If the methods described above are applied in their pure forms, it is possible that some or all of the practical constraints will be violated. Therefore, if such constraints are considered an im-

portant part of the test design, the item selection methods must be altered for operational use. An obvious approach is to choose the item ordering that exhibits the best statistical characteristics, among the set of orderings that satisfy all practical constraints. For instance, suppose that the assessment can be partitioned into five domains, and the only constraint is that all items within a domain must be administered consecutively. In this case, items within a given domain can be ordered using the above criteria. The decision of how to order the domains themselves can be made based on comparisons between them. One approach would be to calculate the median *p* value or median absolute standardized coefficient of the items in each domain, and then place the domains in ascending or descending order of their medians.

As an illustration of the above, consider a test consisting of 15 items, with three items from each of five mutually exclusive domains: General Physical Health, Demographics, Depression, Recent Health Status, and Specific Medical Conditions. Table 1 provides the (hypothetical) absolute standardized logistic regression coefficient of each item. If no constraints are specified, and the ordering is to be done based on $|\hat{\beta}_i^*|$ values, then items are simply placed in descending order of these values ("No Constraints" column of Table 1). If items from the same domain must be presented consecutively, the median of each domain is computed. Domains with the largest median value are administered first; items within a given domain are ordered by their absolute values ("One Constraint" column). If a second constraint is added that all demographic items must be presented first, the ordering is adjusted accordingly ("Two Constraints" column).

In addition to the type of constraints described above, much research has considered the use of exposure control methods to enhance the security of computer-based tests (e.g., Chang, Qian, & Ying, 2001; McBride & Martin, 1983; Stocking & Lewis, 1998; Sympson & Hetter, 1985; van der Linden & Veldkamp, 2004). Exposure control methods are designed to achieve a prescribed balance of administration rates of items in a given item bank. They are most commonly used to ensure that the most popular items in the bank are not exposed at such a high rate that their security becomes compromised. When using questionnaires that are designed to assess respondents' health, however, there is generally no risk associated with allowing items to be publicly released. In fact, the release of items is common practice for such questionnaires, including the MHOS used in this study. As test security is thus not a concern for health questionnaires, these assessments do not generally require exposure control.

It is noted that all item orderings considered here are based on heuristic methods and are not necessarily optimal from a statistical perspective. More complicated tools such as genetic algorithms (Holland, 1968; Holland, 1973; Holland, 1975) could be used to find the ordering that minimizes the average test length, among all orderings that satisfy every constraint. Such procedures have the disadvantage of greater complexity, however, leading to the popularity of heuristic approaches in similar settings (Eggen, 1999; Spray & Reckase, 1994).

Simulation Design

A simulation study was performed to compare the different item ordering methods described above. The design of the study was similar to that of Finkelman et al. (2011); both involved post-hoc simulation using subjects' actual responses to the MHOS. Details about this survey are given elsewhere (Baker, Haffer, & Denniston, 2003; Cooper, Kohlmann, Michael, Haffer, & Stevic, 2001; Haffer & Bowen, 2004); briefly, it provides longitudinal information about the physical and mental statuses of Medicare beneficiaries over multiple years (Haffer & Bowen, 2004). The dataset consisted of responses from the first cohort taking the survey; this group was adminis

Table 1. Item Orderings in the Numerical Example

Item ID	Absolute Value of Standardized Coefficient	Order of Items: No Constraints	Order of Items: One Constraint[a]	Order of Items: Two Constraints[b]
General Physical Health: Median Absolute Value of Domain = 1.25				
A	1.41	4	7	10
B	0.83	12	9	12
C	1.25	8	8	11
Demographics: Median Absolute Value of Domain = 0.92				
D	0.56	15	12	3
E	0.92	11	11	2
F	1.81	1	10	1
Depression: Median Absolute Value of Domain = 1.31				
G	1.04	10	6	9
H	1.37	6	4	7
I	1.31	7	5	8
Recent Health Status: Median Absolute Value of Domain = 0.78				
J	0.59	14	15	15
K	1.13	9	13	13
L	0.78	13	14	14
Specific Medical Conditions: Median Absolute Value of Domain = 1.46				
M	1.58	2	1	4
N	1.38	5	3	6
O	1.46	3	2	5

[a]Constrained so that items from the same domain must be presented consecutively.
[b]Constrained so that items from the same domain must be presented consecutively, and demographic items must be presented first.

tered a baseline instrument in 1998 and followed up on two years later. Using the exclusion rules outlined in Finkelman et al. (2011), the final number of subjects in the study was 119,512. There were a total of 90 variables in the dataset, including demographic information (five items), questions about the subject's health at baseline (84 items), and vital status at follow-up (0 = alive, 1 = deceased). Because some of the variables exhibited missingness, a sequential regression approach had previously been undertaken to impute the missing data; see He, Zaslavsky, Harrington, Catalano, and Landrum (2010), Raghunathan, Lepkowski, Van Hoewyk, and Solenberger (2001), Schenker et al. (2006), and van Buuren, Boshuizen and Knook (1999) for details about the method of sequential regression, and for other applications of this procedure; see Finkelman et al. (2011) for information about the imputation of the particular dataset utilized herein.

To compare the various item ordering methods under curtailment and stochastic curtailment, vital status at follow-up was used as the outcome of interest. A stepwise logistic regression was conducted to develop a predictive model for this outcome; all 89 candidate independent variables

(five items about demographics and 84 items about the subject's baseline health status) were input to the stepwise logistic regression. Wald's test was then used to decide which items would be entered into the model; this test is satisfactory when the sample size is large (Agresti, 1996). Specifically, a p value of 0.05 or less was required for a new item to be added to the model; a p value of 0.10 or greater was required for a previously entered item to be removed. Items selected by the stepwise logistic regression were defined as the *full-length test* upon which curtailment and stochastic curtailment could be performed. The Hosmer-Lemeshow test (Hosmer & Lemeshow, 1989) was used to assess the fit of the model that included all items on the full-length test. Note that the stepwise logistic regression was not performed using the complete set of 119,512 subjects, but approximately two-thirds of them (N = 79,675) who were selected at random as a training dataset. The remaining 39,837 subjects were used as a test dataset upon whom comparisons between item orderings were based.

Because post-hoc simulation (as opposed to monte-carlo simulation) was utilized in this study, subjects' actual and imputed responses were used in the comparison of item ordering methods. In particular, for a given item ordering method, the responses of all 39,837 test set subjects were rearranged as prescribed by the ordering. A FORTRAN 95 program was written and run to determine what each subject's predicted vital status and test length would have been if the item ordering and a given stopping rule (full-length test, curtailment, or stochastic curtailment via empirical proportions) had been used. Note that when the full-length test was used, results were identical for all item orderings; hence, the performance of the full-length test was evaluated only once using an arbitrary item ordering.

For the other stopping rules (curtailment and stochastic curtailment), the following item ordering methods were compared:

1. Placing items in ascending order of their p values in the final model that was fitted to the training data (hereafter *p value ordering*). Ties were broken by placing items with higher Wald statistics before items with lower Wald statistics.
2. Placing items in descending order of their $|\hat{\beta}_i^*|$ values from the final model ($|\hat{\beta}_i^*|$ *ordering*).
3. Placing items in descending order of their $|\hat{\beta}_i^{**}|$ values from the final model ($|\hat{\beta}_i^{**}|$ *ordering*).
4. Placing items in the same order that they were selected by the stepwise logistic regression (*stepwise ordering*).

The above four methods consider only an item's statistical properties when ordering the items within a questionnaire. As explained earlier, however, there might be scenarios where the coherence of a questionnaire is an integral part of its design. Therefore, a *constrained version* of each item ordering method was also defined. In the constrained versions, all items from a given domain were required to be presented consecutively, and all demographic items were required to be presented first. There were seven domains: Demographics, General Physical Health, Health Limitations/Difficulties, Health Status During the Past Four weeks, Non-Life-Threatening Conditions, Life-Threatening Conditions, and Depression. To create the constrained version of the p-value ordering, for example, the following steps were undertaken:

1. *Determine the order of the domains.* As explained above, the Demographics domain was automatically placed first in all constrained orderings. For every other domain, the medi-

an *p* value of items in that domain was calculated; then, domains were placed in ascending order of the medians.

2. *Determine the order of the items within each domain.* This step was simple: for all domains, including Demographics, items were placed in ascending order of their *p* values.

The procedures to create constrained versions of the other item ordering methods were analogous. In each case, the order of the domains themselves was determined by ordering the medians of an appropriate statistic—$|\hat{\beta}_i^*|$,$|\hat{\beta}_i^{**}|$—or the step at which an item was added in the stepwise logistic regression. The ordering of items within a domain was then based on the same statistic of interest.

In addition to the unconstrained and constrained item selection methods, 500 random item orderings were simulated as a baseline for comparison. All item orderings (unconstrained, constrained, and random) were evaluated at two cut points: $p^* = 0.10$ and $p^* = 0.20$. For stochastic curtailment, γ_1 and γ_0 were both set to 0.95, as in Finkelman et al. (2011). For each combination of item ordering method, stopping rule, and cut point, the following statistics were computed: sensitivity (true positive rate), specificity (true negative rate), positive predictive value (percentage of positive test results that are true positives), negative predictive value (percentage of negative test results that are true negatives), mean of the number of items administered, and standard deviation (SD) of the number of items administered. Note that the simulations were conducted solely to illustrate and compare the item ordering methods, not to create an operational questionnaire that would predict the follow-up vital statuses of actual respondents.

Results

Table 2 provides demographic information for the Training, Test, and Combined datasets. Most of the subjects were 65 to 74 years old (57.8% in the Combined dataset), white (88.5%), female (57.7%), married (58.6%), educated at the high school or GED level or less (66.4%), and alive after the two-year follow-up period (92.7%). The largest difference between any two datasets, for any demographic variable, was 0.4% (41.5% of subjects in the training dataset were not married, compared to 41.1% in the test dataset).

47 of the 89 candidate items were selected for the full-length test by the stepwise logistic regression procedure. Information about each of the 47 selected items is given in the Appendix. The *p* value of the Hosmer-Lemeshow test was 0.374; therefore, the fitted model did not exhibit significant misfit in the training dataset.

Results for $p^* = 0.10$

Table 3 provides results for the cut point of $p^* = 0.10$. By definition, all curtailed methods made the same classifications as the full-length test (and one another), as indicated by the shaded area in Table 3; hence, their sensitivities, specificities, negative predictive values, and positive predictive values were identical. On the other hand, the methods varied in terms of the mean number of items that they administered to respondents. For unconstrained item ordering methods coupled with a curtailment stopping rule, the $|\hat{\beta}_i^{**}|$ ordering presented the fewest items on average (34.2), representing a 27.3% reduction in average respondent burden compared to the full-

Table 2. Demographic Characteristics of the Training, Test, and Combined Datasets

	Training Data (N = 79,675)		Test Data (N = 39,837)		Combined (N = 119,512)	
Characteristic	*N*	Percent	*N*	Percent	*N*	Percent
Age Category						
65-74	46,033	57.8	22,996	57.7	69,029	57.8
75+	33,642	42.2	16,841	42.3	50,483	42.2
Race						
White	70,435	88.4	35,348	88.7	105,783	88.5
Non-White	9,240	11.6	4,489	11.3	13,729	11.5
Gender						
Male	33,632	42.2	16,904	42.4	50,536	42.3
Female	46,043	57.8	22,933	57.6	68,976	57.7
Marital Status						
Not Married	33,098	41.5	16,378	41.1	49,476	41.4
Married	46,577	58.5	23,459	58.9	70,036	58.6
Education Level						
HS or GED or Less	52,850	66.3	26,488	66.5	79,338	66.4
Greater than HS or GED	26,825	33.7	13,349	33.5	40,174	33.6
Vital Status at Follow-Up						
Alive	73,869	92.7	36,903	92.6	110,772	92.7
Deceased	5,806	7.3	2,934	7.4	8,740	7.3

Note. All counts and percentages include both actual and imputed values. For the Combined dataset, the percentage missing was as follows: Age = 0.0%, Race = 1.6%, Gender = 0.5%, Marital Status = 0.5%, Education Level = 2.0%, Vital Status at Follow-up = 0.0%. All information other than Vital Status is based on baseline data. This table was adapted from Table 1 of Finkelman et al. (2011).

length test. This was followed by the stepwise, p value, and $|\hat{\beta}_i^*|$ orderings (24.1%, 22.8%, and 21.1% reductions in average respondent burden, respectively). For constrained item ordering methods coupled with a curtailment stopping rule, the p value ordering exhibited the greatest reduction in average respondent burden (21.7%), followed by the $|\hat{\beta}_i^{**}|$, stepwise, and $|\hat{\beta}_i^*|$ orderings (21.5%, 20.7%, and 20.7%, respectively). All methods, including both unconstrained and constrained orderings, exhibited lower mean test lengths than the average of the 500 random item ordering methods' mean test lengths (which was 41.1). The mean test lengths of the random orderings ranged from 35.8 to 44.8 under a curtailment stopping rule. Two methods, unconstrained $|\hat{\beta}_i^{**}|$ and unconstrained stepwise, exhibited greater reductions in average respondent burden than all 500 of the random orderings.

When stochastic curtailment was used as the stopping rule, measures of predictive accuracy (sensitivity, specificity, negative predictive value, and positive predictive value) varied among the four item ordering methods; however, differences in these statistics were never greater than 0.3%. Comparing the unconstrained methods' average levels of respondent burden, the stepwise ordering method exhibited the greatest reduction (57.6%), followed by the $|\hat{\beta}_i^*|$ ordering

Table 3. Results for $p^* = 0.10$

Termination Rule and Item Ordering Method	Sensitivity	Specificity	Negative Predictive Value	Positive Predictive Value	Number of Items	
					Mean	SD
Full-Length Termination						
Any Ordering	60.4	84.1	96.4	23.2	47.0	0.0
Curtailed Termination						
Stepwise (Unconstrained)	60.4	84.1	96.4	23.2	35.7	5.9
p Value (Unconstrained)	60.4	84.1	96.4	23.2	36.3	5.4
$\lvert \hat{\beta}_i^* \rvert$ (Unconstrained)	60.4	84.1	96.4	23.2	37.1	5.8
$\lvert \hat{\beta}_i^{**} \rvert$ (Unconstrained)	60.4	84.1	96.4	23.2	34.2	6.1
Stepwise (Constrained)	60.4	84.1	96.4	23.2	37.3	4.3
p Value (Constrained)	60.4	84.1	96.4	23.2	36.8	4.6
$\lvert \hat{\beta}_i^* \rvert$ (Constrained)	60.4	84.1	96.4	23.2	37.3	4.8
$\lvert \hat{\beta}_i^{**} \rvert$ (Constrained)	60.4	84.1	96.4	23.2	36.9	4.1
Average Value of 500 Random Item Orderings	60.4	84.1	96.4	23.2	41.1	3.4
Stochastically Curtailed Termination						
Stepwise (Unconstrained)	60.5	84.0	96.4	23.1	19.9	10.4
p Value (Unconstrained)	60.5	84.1	96.4	23.2	24.6	11.4
$\lvert \hat{\beta}_i^* \rvert$ (Unconstrained)	60.7	83.9	96.4	23.1	20.9	9.9
$\lvert \hat{\beta}_i^{**} \rvert$ (Unconstrained)	60.5	84.1	96.4	23.2	21.5	10.2
Stepwise (Constrained)	60.4	83.9	96.4	23.0	22.8	9.4
p Value (Constrained)	60.7	84.0	96.4	23.1	26.7	7.8
$\lvert \hat{\beta}_i^* \rvert$ (Constrained)	60.5	83.9	96.4	23.1	26.6	7.4
$\lvert \hat{\beta}_i^{**} \rvert$ (Constrained)	60.4	83.9	96.4	23.0	32.4	6.2
Average Value of 500 Random Item Orderings	60.3	84.1	96.4	23.1	34.7	6.9

(55.6%), the $|\hat{\beta}_i^{**}|$ ordering (54.2%), and the p value ordering (47.7%). Under constraints, the stepwise ordering again achieved the greatest reduction in average test length (51.5%), followed by the $|\hat{\beta}_i^{*}|$, p value, and $|\hat{\beta}_i^{**}|$ orderings (43.5%, 43.2%, and 31.2%, respectively). The mean test length of the $|\hat{\beta}_i^{**}|$ method was considerably higher than all other methods when constraints were applied, including an average difference of 9.6 items between the $|\hat{\beta}_i^{**}|$ and stepwise methods. The mean test lengths of the random orderings ranged from 25.4 to 41.8 items under a stochastic curtailment stopping rule, with an average of 34.7. All unconstrained and constrained item ordering methods based on statistical criteria exhibited lower mean test lengths than the average of the random item orderings. Every unconstrained method, as well as the constrained stepwise method, exhibited greater reductions in average respondent burden than all 500 of the random orderings.

Results for p = 0.20*

Table 4 provides results for the cut point of p^* = 0.20. For unconstrained item ordering methods coupled with a curtailment stopping rule, the relative levels of average respondent burden followed the same pattern as they had for p^* = 0.10: the $|\hat{\beta}_i^{**}|$ ordering exhibited the greatest reduction (38.4%), followed by the stepwise (34.1%), p value (32.6%), and $|\hat{\beta}_i^{*}|$ (31.4%) orderings. Under constraints and a curtailment stopping rule, the stepwise ordering exhibited the greatest reduction (29.4%), followed by the p value, $|\hat{\beta}_i^{**}|$, and $|\hat{\beta}_i^{*}|$ orderings (29.0%, 28.9%, and 28.0%, respectively). As in p^* = 0.10, all unconstrained and constrained methods exhibited lower mean test lengths than the average of the 500 random item ordering methods' mean test lengths, which was 38.2. The mean test lengths of the random orderings ranged from 31.5 to 43.7 under a curtailment stopping rule. The unconstrained $|\hat{\beta}_i^{**}|$ and unconstrained stepwise methods exhibited greater reductions in average respondent burden than all 500 of the random orderings.

Under stochastic curtailment, the four item ordering methods were all within 0.7% of one another for each measure of predictive accuracy. For the unconstrained methods' average levels of respondent burden, the stepwise method resulted in the greatest reduction (74.3%), followed by the $|\hat{\beta}_i^{**}|$ordering (72.3%), the $|\hat{\beta}_i^{*}|$ ordering (70.1%), and the p value ordering (68.8%). When constraints were applied, the stepwise ordering again achieved the largest reduction (65.1%), followed by the $|\hat{\beta}_i^{*}|$, p value, and $|\hat{\beta}_i^{**}|$ orderings (58.7%, 57.9%, and 48.7%, respectively). As in p^* = 0.10, there was greater variation in the mean test lengths under constraints than under no constraints. Specifically, when p^* = 0.20, stopping was performed via stochastic curtailment, and constraints were imposed, the $|\hat{\beta}_i^{**}|$ ordering administered an average of 7.7 more items than the stepwise ordering. By contrast, the difference in means of any two unconstrained methods was no more than 2.6 items when stochastic curtailment was applied and p^* = 0.20. The mean test lengths of the random orderings ranged from 18.2 to 37.1 items under a stochastic curtailment stopping rule, with an average of 27.2. All unconstrained and constrained item ordering methods based on statistical criteria exhibited lower mean test lengths than the average of the random item orderings. Every unconstrained method, as well as the constrained stepwise meth-

Table 4. Results for $p^* = 0.20$

Termination Rule and Item Ordering Method	Sensitivity	Specificity	Negative Predictive Value	Positive Predictive Value	Number of Items	
					Mean	SD
Full-Length Termination						
Any Ordering	37.3	94.5	95.0	35.1	47.0	0.0
Curtailed Termination						
Stepwise (Unconstrained)	37.3	94.5	95.0	35.1	31.0	6.2
p Value (Unconstrained)	37.3	94.5	95.0	35.1	31.7	5.9
$\lvert \hat{\beta}_i^* \rvert$ (Unconstrained)	37.3	94.5	95.0	35.1	32.2	7.0
$\lvert \hat{\beta}_i^{**} \rvert$ (Unconstrained)	37.3	94.5	95.0	35.1	28.9	7.1
Stepwise (Constrained)	37.3	94.5	95.0	35.1	33.2	5.9
p Value (Constrained)	37.3	94.5	95.0	35.1	33.3	4.9
$\lvert \hat{\beta}_i^* \rvert$ (Constrained)	37.3	94.5	95.0	35.1	33.9	5.0
$\lvert \hat{\beta}_i^{**} \rvert$ (Constrained)	37.3	94.5	95.0	35.1	33.4	4.5
Average Value of 500 Random Item Orderings	37.3	94.5	95.0	35.1	38.2	3.9
Stochastically Curtailed Termination						
Stepwise (Unconstrained)	36.7	94.6	95.0	35.3	12.1	9.5
p Value (Unconstrained)	36.9	94.6	95.0	35.4	14.7	10.8
$\lvert \hat{\beta}_i^* \rvert$ (Unconstrained)	36.9	94.6	95.0	35.2	14.0	9.2
$\lvert \hat{\beta}_i^{**} \rvert$ (Unconstrained)	36.7	94.7	95.0	35.6	13.0	10.0
Stepwise (Constrained)	37.1	94.6	95.0	35.3	16.4	8.7
p Value (Constrained)	37.1	94.6	95.0	35.2	19.8	8.5
$\lvert \hat{\beta}_i^* \rvert$ (Constrained)	37.1	94.5	95.0	35.0	19.4	9.1
$\lvert \hat{\beta}_i^{**} \rvert$ (Constrained)	37.4	94.5	95.0	35.0	24.1	9.0
Average Value of 500 Random Item Orderings	36.8	94.6	95.0	35.2	27.2	8.4

od, exhibited greater reductions in average respondent burden than all 500 of the random orderings.

Further Analysis

Correlations among item orderings. To gain further insight into the differences between methods, the item orderings themselves were compared. A new variable was created for each method, indicating where that method placed the items within the questionnaire. The first item presented by a given method was coded "1" in this new variable; the second item presented was coded "2," and so forth, so that every item had a value between 1 and 47. For example, in the new variable created for the unconstrained stepwise method, the item coded "1" related to difficulties with bathing, the item coded "2" related to general health, the item coded "3" related to age, and so on (see Appendix). The Spearman rank correlation was then computed for each pair of new variables to assess the degree of concordance between item orderings.

Results are presented in Table 5. The correlation was 0.65 or higher between each pair of unconstrained item ordering methods. Pairwise correlations between the constrained stepwise, constrained p value, and constrained $|\hat{\beta}_i^*|$ methods all exceeded 0.90; however, the correlation between the constrained $|\hat{\beta}_i^{**}|$ method and other constrained methods never exceeded 0.21. In fact, the correlation between the constrained $|\hat{\beta}_i^{**}|$ method and any other method, constrained or unconstrained, was never higher than 0.50. This finding shows that the ordering based on the constrained $|\hat{\beta}_i^{**}|$ method was non-trivially different from all other orderings, and might partially explain why it exhibited lower reductions in average respondent burden than other methods in some simulation conditions.

Ordering of domains by constrained methods. As stated previously, all items from a domain had to be administered consecutively when simulations under constraints were performed. To investigate why the constrained $|\hat{\beta}_i^{**}|$ method exhibited low Spearman correlations with the other methods, the ordering in which this method presented the different domains was examined. This ordering is shown in Table 6, along with analogous results for the other three constrained procedures. The table indicates that the constrained stepwise, p value, and $|\hat{\beta}_i^*|$ orderings all placed the Past Four Weeks domain in the sixth position, near the end of the questionnaire. The constrained $|\hat{\beta}_i^{**}|$ method, on the other hand, placed this domain in the third position, near the start of the questionnaire. As this domain included more items than any other domain (16 items), it had a large influence on the overall ordering of items. Additionally, the constrained $|\hat{\beta}_i^{**}|$ method placed the General Health and Non-Life-Threatening Conditions domains later than the other methods. A deeper look at the constrained $|\hat{\beta}_i^{**}|$ method revealed that several domains exhibited similar median $|\hat{\beta}_i^{**}|$ values to one another (0.274 for Past Four Weeks, 0.264 for Health Limitations/Difficulties, and 0.261 for General Health), suggesting that the ordering of domains was based on minor deviations for this method. These minor deviations resulted in a different domain ordering from the other constrained methods, leading to the low Spearman correlations described above.

Table 5. Spearman Correlations Between the Item Orderings of Different Methods

Item Ordering	Stepwise Unconstrained	*p* Value Unconstrained	$\lvert\hat{\beta}_i^{*}\rvert$ Unconstrained	$\lvert\hat{\beta}_i^{**}\rvert$ Unconstrained	Stepwise Constrained	*p* Value Constrained	$\lvert\hat{\beta}_i^{*}\rvert$ Constrained	$\lvert\hat{\beta}_i^{**}\rvert$ Constrained
Stepwise Unconstrained	1.0	0.80**	0.78**	0.65**	0.59**	0.50**	0.49**	0.24
p Value Unconstrained		1.0	0.84**	0.78**	0.52**	0.60**	0.56**	0.32*
$\lvert\hat{\beta}_i^{*}\rvert$ Unconstrained			1.0	0.75**	0.54**	0.51**	0.57**	0.25
$\lvert\hat{\beta}_i^{**}\rvert$ Unconstrained				1.0	0.37*	0.35*	0.32*	0.50**
Stepwise Constrained					1.0	0.92**	0.92**	0.18
p Value Constrained						1.0	0.97**	0.21
$\lvert\hat{\beta}_i^{*}\rvert$ Constrained							1.0	0.12
$\lvert\hat{\beta}_i^{**}\rvert$ Constrained								1.0

*Statistically significant at $p \leq 0.05$ (two-tailed).
**Statistically significant at $p \leq 0.01$ (two-tailed).

Table 6. Ordering of Domains by Constrained Methods

Domains Presented	Constrained Stepwise	Constrained p Value	Constrained $\lvert \hat{\beta}_i^{*} \rvert$	Constrained $\lvert \hat{\beta}_i^{**} \rvert$
First	Demographics	Demographics	Demographics	Demographics
Second	General Health	Life-Threatening Conditions	General Health	Life-Threatening Conditions
Third	Life-Threatening Conditions	General Health	Life-Threatening Conditions	Past Four Weeks
Fourth	Health Limitations/Difficulties	Non-Life-Threatening Conditions	Non-Life-Threatening Conditions	Health Limitations/ Difficulties
Fifth	Non-Life-Threatening Conditions	Health Limitations/ Difficulties	Health Limitations/ Difficulties	General Health
Sixth	Past Four Weeks	Past Four Weeks	Past Four Weeks	Non-Life-Threatening Conditions
Seventh	Depression	Depression	Depression	Depression

Discussion and Conclusions

The results indicated that the statistical methods did exhibit lower average test lengths than random item orderings while maintaining comparable levels of predictive accuracy. Although no method was superior under all simulation conditions, the stepwise ordering achieved the greatest average reduction in test length under the majority of conditions. The $\lvert \hat{\beta}_i^{**} \rvert$ ordering exhibited substantially higher average test lengths than the other methods when constraints were applied and stochastic curtailment was used; this ordering also had low Spearman correlations with better-performing methods. When selecting a constrained item ordering for an operational questionnaire, practitioners might favor those orderings that exhibit high Spearman correlations with the unconstrained methods, which should be expected to have better statistical properties and in particular, lower average test lengths.

The data used in simulation came from the same cohort of the MHOS as was examined in Finkelman et al. (2011); however, several aspects of the simulation design differed between the two studies. In Finkelman et al. (2011), a random subsample of 20,000 subjects was selected from the 119,512 subjects who met the criteria for inclusion. Half of the 20,000 subjects were assigned to the training set, and half were assigned to the test set. By contrast, all 119,512 subjects were used in the current study: two-thirds were randomly selected for the training set and one-third were selected for the test set. By using all available data, the precision of the results was enhanced. Another consequence of using a larger sample size in the training set was that standard errors of $\hat{\beta}_i$ values were reduced, and more items with modest effect sizes were entered into the model [a total of 47 items were entered, as opposed to 23 items in Finkelman et al. (2011)]. This feature of the current study allowed the examination of the sequential stopping rules' respective performances when applied to a longer questionnaire. Using curtailment and stochastic curtailment with a 47-item questionnaire resulted in slightly higher reductions in per-

cent respondent burden than had been observed for the 23-item questionnaire of Finkelman et al. (2011). This finding was likely due to the aforementioned presence of items with modest effect sizes in the 47-item questionnaire. Because these items lacked predictive power, they were unlikely to change a respondent's classification, and thus they were often eliminated by early stopping.

Several other features of the simulation results were consistent with either expectations, previous studies, or both. For instance, unconstrained item orderings achieved lower average test lengths than constrained orderings, an intuitive finding that agrees with research in other CCT settings (Bartroff et al., 2008; Lau & Wang, 1999). Additionally, greater percent reductions in respondent burden were found when the cut point was further from the proportion of subjects who died before follow-up (which was approximately 7.3%; see Table 2). This pattern was also found in Finkelman et al. (2011). It is explained by the fact that when the cut point is close to the proportion of deaths, the starting point for the probability of each classification is close to 50% and substantial evidence must be found in one direction or another before early stopping can be invoked. However, when the cut point is far from the proportion of deaths, the starting probability of one classification is higher than 50%, and less evidence is required for that classification to be made via early stopping (Finkelman et al., 2011).

All simulations with the MHOS dataset were performed solely to compare the item ordering methods. It is not intended that the resulting questionnaire be used operationally in predicting respondents' two-year vital statuses. Practical implementation of a questionnaire would require "skip patterns" (Al-Tayyib, Rogers, Gribble, Villarroel, & Turner, 2002; Des Jarlais et al., 1999) to avoid presentation of redundant or irrelevant items. For instance, respondents who identified themselves as female, or who claimed never to have been diagnosed with any cancer, would be skipped past the item about prostate cancer. Additionally, the survey sometimes asked if a respondent had "any of the following problems," and subsequently provided a set of items related to possible medical issues. An operational ordering method would need to present these items consecutively, or change the stem of the items to make them independent of one another, so that each would make sense to the respondent. Further constraints would have to be considered, such as presenting items related to sensitive material near the end of the questionnaire. Kingsbury and Zara (1989) and Veldkamp and van der Linden (2002) provided examples of constrained assessment in other computerized adaptive testing contexts. Finally, the full-length test defined herein was based on a stepwise logistic regression and was fairly long (47 items). Questionnaires with many items are not uncommon in the health field (Quittner et al., 2005; Slade & Spencer, 1994); however, if the purpose of the study had not been purely illustrative, other candidate full-length tests would have been considered. Aday (1996) provides information about the process of designing and conducting a health survey.

It is worth noting that computer-based health questionnaires have gained increased attention recently due to their prominence in the Patient-Reported Outcomes Measurement Information System (PROMIS). PROMIS instruments have been developed to measure depression, physical function, pain, and fatigue, among many other domains; see Choi, Reise, Pilkonis, Hays, and Cella (2010), Fries, Cella, Rose, Krishnan, and Bruce (2009), and Reeve et al. (2007) for information about computer-based assessment within this system.

It is also notable that the methods utilized in this study are different from most previous CCT research. In particular, logistic regression modeling has been used in lieu of IRT models, and new approaches to item ordering have been examined in lieu of traditional criteria based on Fisher information (e.g., Spray & Reckase, 1994) or Kullback-Leibler information (e.g., Eggen, 1999). These differences are due to the fact that the current study's methodology was designed to

predict an observable outcome rather than to measure a latent trait. Specifically, when predicting an observable variable that is dichotomous, logistic regression is a more appropriate statistical modeling tool than IRT. It was therefore necessary to develop item ordering methods that are suited to logistic regression and other predictive models of observable outcomes. In addition to the illustrative example presented in this study (using responses from the MHOS to predict vital status at follow-up), there are a number of other instruments that are used to predict observable outcomes. These include, but are not limited to, questionnaires that predict deliberate self-harm and referral for mental health service (Ferdinand & Verhulst, 1994), the results of a skin test for tuberculosis (Froehlich, Ackerson, Morozumi, & The Pediatric Tuberculosis Study Group of Kaiser Permanente, Northern California, 2001), and the future abuse of medication by chronic pain patients (Butler, Fernandez, Benoit, Budman, & Jamison, 2008). The goal of the current study has been to improve such questionnaires by coupling them with efficient item ordering methods using computer-based, variable-length testing procedures. Therefore, although the statistical tools utilized herein are unlike most previous CCT approaches, the goal is consistent with the fundamental objectives of CCT.

Although the current study has laid the groundwork for ordering items in a stochastically curtailed questionnaire that predicts an observable outcome, further examination is needed. Item ordering methods should be compared using additional datasets and under different conditions. More specifically, the constraints, outcome variable to be predicted, and number of items in the full-length test should be varied. Item ordering methods in the presence of statistical interactions between items should be developed. The case of predicting multiple outcomes should be studied, as should the impact of context effects when items are arranged using statistical criteria. All of these topics will be undertaken in future work.

References

Adams, L.M., & Gale, D. (1982). Solving the quandary between questionnaire length and response rate in educational research. *Research in Higher Education, 17*(3), 231-240. CrossRef

Aday, L.A. (1996). *Designing and conducting health surveys, second edition*. San Francisco, CA: Jossey-Bass.

Agresti, A. (1996). *An introduction to categorical data analysis*. New York, NY: Wiley.

Al-Tayyib, A.A., Rogers, S.M., Gribble, J.N., Villarroel, M., & Turner, C.F. (2002). Effect of low medical literacy on health survey measurements. *American Journal of Public Health, 92*(9), 1478-1480. CrossRef

Baker, F., Haffer, S.C., & Denniston, M. (2003). Health-related quality of life of cancer and noncancer patients in Medicare managed care. *Cancer, 97*(3), 674-681. CrossRef

Bartroff, J., Finkelman, M., & Lai, T.L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika, 73*(3), 473-486. CrossRef

Betensky, R.A. (1997). Early stopping to accept H(o) based on conditional power: Approximations and comparisons. *Biometrics, 53*(3), 794-806. CrossRef

Butler, S.F., Fernandez, K., Benoit, C., Budman, S.H., & Jamison, R.N. (2008). Validation of the revised Screener and Opioid Assessment for Patients with Pain (SOAPP-R). *Journal of Pain, 9*(4), 360-372. CrossRef

Chang, H.-H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing with *b* blocking. *Applied Psychological Measurement, 25*(4), 333-341. CrossRef

Choi, S.W., Reise, S.P., Pilkonis, P.A., Hays, R.D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research, 19*(1), 125-136. CrossRef

Chung, F., Yegneswaran, B., Liao, P., Chung, S.A., Vairavanathan, S., Islam, S., Khajehdehi, A., & Shapiro, C.M. (2008). STOP questionnaire: A tool to screen patients for obstructive sleep apnea. *Anesthesiology, 108*(5), 812-821. CrossRef

Cooper, J.K., Kohlmann, T., Michael, J.A., Haffer, S.C., & Stevic, M. (2001). Health outcomes. New quality measure for Medicare. *International Journal of Quality Health Care, 13*(1), 9-16. CrossRef

Davis, B.R., & Hardy, R.J. (1994). Data monitoring in clinical trials: The case for stochastic curtailment. *Journal of Clinical Epidemiology, 47*(9), 1033-1042. CrossRef

Des Jarlais, D.C., Paone, D., Milliken, J., Turner, C.F., Miller, H., Gribble, J., Shi, Q., Hagan, H., & Friedman, S.R. (1999). Audio-computer interviewing to measure risk behaviour for HIV among injecting drug users: A quasi-randomised trial. *Lancet, 353*(9165), 1657-1661. CrossRef

Eggen, T.J.H.M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*(3), 249-261. CrossRef

Eggen, T.J.H.M., & Straetmans, G.J.J.M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*(5), 713-734. CrossRef

Eisenberg, B., & Ghosh, B.K. (1980). Curtailed and uniformly most powerful sequential tests. *The Annals of Statistics, 8*(5), 1123-1131. CrossRef

Eisenberg, B., & Simons, G. (1978). On weak admissibility of tests. *The Annals of Statistics, 6*(2), 319-332. CrossRef

Ferdinand, R.F., & Verhulst, F.C. (1994). The prediction of poor outcome in young adults: Comparison of the Young Adult Self-Report, the General Health Questionnaire and the Symptom Checklist. *Acta Psychiatrica Scandinavica, 89*(6), 405-410. CrossRef

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*(4), 442-463. CrossRef

Finkelman, M.D. (2010). Variations on stochastic curtailment in sequential mastery testing. *Applied Psychological Measurement, 34*(1), 27-45. CrossRef

Finkelman, M.D., He, Y., Kim, W., & Lai, A.M. (2011). Stochastic curtailment of health questionnaires: A method to reduce respondent burden. *Statistics in Medicine, 30*(16), 1989-2004. CrossRef

Fries, J.F., Cella, D., Rose, M., Krishnan, E., & Bruce, B. (2009). Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. *Journal of Rheumatology, 36*(9), 2061-2066. CrossRef

Froehlich, H., Ackerson, L.M., Morozumi, P.A., & the Pediatric Tuberculosis Study Group of Kaiser Permanente, Northern California (2001). Targeted testing of children for tuberculosis: Validation of a risk assessment questionnaire. *Pediatrics, 107*(4), E54.

Haffer, S.C., & Bowen, S.E. (2004). Measuring and improving health outcomes in Medicare: The Medicare HOS program. *Health Care Finance Review, 25*(4), 1-3.

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

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE.

He, Y., Zaslavsky, A.M., Harrington, D.P., Catalano, P., & Landrum, M.B. (2010). Multiple

imputation in a large-scale complex survey: A practical guide. *Statistical Methods in Medical Research, 19*(6), 653-670. CrossRef

Herzog, A.R., & Bachman, J.G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly, 45*(4), 549-559. CrossRef

Holland, J. (1968). *Hierarchical descriptions, universal spaces and adaptive systems: Technical report, ORA Projects 01252 and 08226*. Ann Arbor, MI: University of Michigan.

Holland, J. (1973). Genetic algorithms and the optimal allocation of trials. *SIAM Journal of Computing, 2*(2), 88-105. CrossRef

Holland, J. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor, MI: University of Michigan.

Hosmer, D.W., & Lemeshow, S. (1989). *Applied logistic regression*. New York, NY: Wiley.

Huebner, A. (2012). Item overexposure in computerized classification tests using sequential item selection. *Practical Assessment, Research & Evaluation, 17*(12). Retrieved from http://pareonline.net/pdf/v17n12.pdf

Huebner, A., & Li, Z. (2012). A stochastic method for balancing item exposure rates in computerized classification tests. *Applied Psychological Measurement, 36*(3), 181-188. CrossRef

Kingsbury, G.G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375. CrossRef

Lan, K.K.G., Simon, R., & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics-Sequential Analysis, 1*(3), 207-219. CrossRef

Lau, C.A., & Wang, T. (1999). *Computerized classification testing under practical constraints with a polytomous model.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Leung, D.H.-Y., Wang, Y.-G., & Amar, D. (2003). Early stopping by using stochastic curtailment in a three-arm sequential trial. *Applied Statistics, 52*(2), 139-152. CrossRef

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*(4), 367-386. CrossRef

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

McBride, J.R., & Martin, J.T. (1983). Reliability and validity of adaptive verbal ability tests in a military setting. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223-236). New York: Academic Press.

Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *The American Statistician, 58*(3), 218-223. CrossRef

Quittner, A.L., Buu, A., Messer, M.A., Modi, A.C., & Watrous, M. (2005). Development and validation of The Cystic Fibrosis Questionnaire in the United States: A health-related quality-of-life measure for cystic fibrosis. *Chest, 128*(4), 2347-2354. CrossRef

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401. CrossRef

Raghunathan, T., Lepkowski, J.M., Van Hoewyk, J., & Solenberger, P.W. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85-95.

Reckase, M.D. (1983). A procedure for decision making using tailored testing. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237-255). New York: Academic Press.

Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., Thissen, D., Revicki, D.A., Weiss, D.J., Hambleton, R.K., Liu, H., Gershon, R., Reise, S.P., Lai, J.S., & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*(5 Suppl. 1), S22-31. CrossRef

Rogers, W.H., Wilson, I.B., Bungay, K.M., Cynn, D.J., & Adler, D.A. (2002). Assessing the performance of a new depression screener for primary care (PC-SAD). *Journal of Clinical Epidemiology, 55*(2), 164-175. CrossRef

Rudner, L.M. (2009). Scoring and classifying examinees using measurement decision theory. *Practical Assessment, Research & Evaluation, 14*(8). Retrieved from http://pareonlne.net/getvn.asp?v=14&n=8

Ruige, J.B., de Neeling, J.N., Kostense, P.J., Bouter, L.M., & Heine, R.J. (1997). Performance of an NIDDM screening questionnaire based on symptoms and risk factors. *Diabetes Care, 20*(4), 491-496. CrossRef

Schenker, N., Raghunathan, T.E., Chiu, P.L., Makuc, D.M., Zhang, G., & Cohen, A.J. (2006). Multiple imputation for missing income data in the National Health Interview Survey. *Journal of the American Statistical Association, 101*(475), 924-933. CrossRef

Scientific Advisory Committee of the Medical Outcomes Trust (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research, 11*(3), 193-205. CrossRef

Slade, G.D., & Spencer, A.J. (1994). Development and evaluation of the Oral Health Impact Profile. *Community Dental Health, 11*(1), 3-11.

Spray, J.A., & Reckase, M.D. (1994). *The selection of test items for decision making with a computer adaptive test.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Stocking, M.L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Applied Psychological Measurement, 23*(1), 57-75. CrossRef

Sympson, J.B., & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thompson, N.A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research & Evaluation, 12*(1). Retrieved from http://pareonline.net/getvn.asp?v=12&n=1

Thompson, N.A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation, 16*(4). Retrieved from http://pareonline.net/getvn.asp?v=16&n=4

van Buuren, S., Boshuizen, H.C., & Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18*(6), 681-694. CrossRef

van der Linden, W.J., & Veldkamp, B.P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*(3), 273-291. CrossRef

Veldkamp, B.P., & van der Linden, W.J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*(4), 575-588. CrossRef

Vos, H.J. (2000). A Bayesian procedure in the context of sequential mastery testing. *Psicológica, 21,* 191-211. Retrieved from http://www.uv.es/revispsi/articulos1y2.00/vos.pdf

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Walter, O.B. (2010). Adaptive tests for measuring anxiety and depression. In W.J. van der

Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing* (pp. 123-136). New York, NY: Springer.

Ware, J.E., & 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.

Weiss, D.J., & Kingsbury, G.G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375. CrossRef

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement, 67*(1), 41-58. CrossRef

Acknowledgments

The authors thank two anonymous reviewers, as well as the Associate Editor and Editor, for their comments.

Author Address

Matthew D. Finkelman, 1 Kneeland St., Boston, MA 02111. U.S.A.
Email Matthew.Finkelman@tufts.edu.

Appendix

Information About Items Selected by Stepwise Logistic Regression*

Topic of Item	Step Added	Domain	Description/Wording	Codes	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$
Bathing	1	Health Limitations/ Difficulties	Because of a health or physical problem, do you have any difficulty doing the following activities? Bathing**	1 = I am unable to do this activity 2 = Yes, I have difficulty 3 = No, I do not have difficulty	-.321	.036
General Health	2	General Health	In general, would you say your health is:	1 = Excellent 2 = Very Good 3 = Good 4 = Fair 5 = Poor	.220	.024
Age	3	Demographics	Age category of the subject	0 = 65-74 years 1 = 75 years or more	.674	.032
Lung Cancer Treatment	4	Life Threatening Conditions	Are you currently under treatment for: Lung cancer**	1 = Yes 2 = No	-1.243	.101
Congestive Heart Failure	5	Life Threatening Conditions	Has a doctor ever told you that you had: Congestive heart failure**	1 = Yes 2 = No	-.548	.042
Walking One Block	6	Health Limitations/ Difficulties	Does your health limit you in these activities? If so, how much: Walking one block**	1 = Yes, limited a lot 2 = Yes, limited a little 3 = No, not limited at all	-.099	.034
Gender	7	Demographics	Gender of the subject	0 = Male 1 = Female	-.450	.035
Arthritis Pain	8	Past Four Weeks	During the past 4 weeks, how would you describe the arthritis pain you usually had?	1 = None 2 = Very mild 3 = Mild 4 = Moderate 5 = Severe	-.043	.017
Any Cancer	9	Life Threatening Conditions	Has a doctor ever told you that you had: Any cancer (other than skin cancer)**	1 = Yes 2 = No	-.488	.041
Time Interfered With Social Activities	10	Past Four Weeks	During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities?	1 = All of the time 2 = Most of the time 3 = Some of the time 4 = A little of the time 5 = None of the time	-.091	.018
Smoked 100 Cigarettes	11	General Health	Have you ever smoked at least 100 cigarettes in your entire life?	0 = No or Don't Know 1 = Yes	.250	.032

Lifting/Carrying Groceries	12	Health Limitations/ Difficulties	Does your health limit you in these activities? If so, how much? Lifting or Carrying Groceries**	1 = Yes, limited a lot 2 = Yes, limited a little 3 = No, not limited at all	-.207	.030
Sciatica	13	Non-Life Threatening Conditions	Has a doctor ever told you that you had: Sciatica**	1 = Yes 2 = No	.192	.039
Marital Status	14	Demographics	Current marital status of the subject	0 = Married 1 = Non-married	.207	.032
Low Back Pain	15	Past Four Weeks	In the past 4 weeks, how often has low back pain interfered with your usual daily activities?	1 = All of the time 2 = Most of the time 3 = Some of the time 4 = A little of the time 5 = None of the time	.063	.014
Dyspnea when Walking	16	Past Four Weeks	During the past 4 weeks, how often have you felt short of breath under the following conditions? When walking less than one block**	1 = All of the time 2 = Most of the time 3 = Some of the time 4 = A little of the time 5 = None of the time	-.072	.015
Acid Indigestion/Heartburn	17	Non-Life Threatening Conditions	Do you now have acid indigestion or heartburn?	1 = Yes 2 = No	.158	.033
Excellent Health	18	General Health	How true or false is each of the following statements for you? My health is excellent**	1 = Definitely true 2 = Mostly true 3 = Don't know 4 = Mostly false 5 = Definitely false	.068	.018
Vision	19	Health Limitations/ Difficulties	Can you see well enough to read newspaper print (with your glasses or contacts if that's how you see best)?	1 = Yes 2 = No	.211	.047
Bending, Kneeling, and Stooping	20	Health Limitations/ Difficulties	Does your health limit you in these activities? If so, how much: Bending, kneeling, or stooping**	1 = Yes, limited a lot 2 = Yes, limited a little 3 = No, not limited at all	.186	.029
Walking More Than a Mile	21	Health Limitations/ Difficulties	Does your health limit you in these activities? If so, how much: Walking more than a mile**	1 = Yes, limited a lot 2 = Yes, limited a little 3 = No, not limited at all	-.146	.035
Bodily Pain	22	Past Four Weeks	How much bodily pain have you had during the past 4 weeks?	1 = None 2 = Very mild 3 = Mild 4 = Moderate 5 = Severe 6 = Very severe	-.083	.016

Health Interfering With Social Activities	23	Past Four Weeks	During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities…?	1 = Not at all 2 = Slightly 3 = Moderately 4 = Quite a bit 5 = Extremely	.070	.017
Bathing or Dressing	24	Health Limitations/ Difficulties	Does your health limit you in these activities? If so, how much? Bathing or dressing yourself**	1 = Yes, limited a lot 2 = Yes, limited a little 3 = No, not limited at all	-.143	.031
Walking	25	Health Limitations/ Difficulties	Because of a health or physical problem, do you have any difficulty doing the following activities? Walking**	1 = I am unable to do this activity 2 = Yes, I have difficulty 3 = No, I do not have difficulty	-.123	.036
Arthritis of Hip or Knee	26	Non-Life Threatening Conditions	Has a doctor ever told you that you had: Arthritis of the hip or knee**	1 = Yes 2 = No	.160	.037
Depression Most of the Time	27	Depression	Have you ever had 2 years or more in your life when you felt depressed or sad most days, even if you felt okay sometimes?	1 = Yes 2 = No	.189	.046
Sores/Wounds on Feet	28	Past Four Weeks	During the past 4 weeks, how much of the time have you had any of the following problems with your legs and feet? Sores or wounds on your feet that did not heal**	1 = All of the time 2 = Most of the time 3 = Some of the time 4 = A little of the time 5 = None of the time	-.088	.023
Energy	29	Past Four Weeks	How much of the time during the past 4 weeks…Did you have a lot of energy?**	1 = All of the time 2 = Most of the time 3 = A good bit of the time 4 = Some of the time 5 = A little of the time 6 = None of the time	.057	.017
Prostate Cancer	30	Life Threatening Conditions	Are you currently under treatment for: Prostate cancer**	1 = Yes 2 = No	.276	.079
Orthopnea	31	Past Four Weeks	During the past 4 weeks, how often have you felt short of breath under the following conditions? When lying down flat**	1 = All of the time 2 = Most of the time 3 = Some of the time 4 = A little of the time 5 = None of the time	-.086	.019

Chest Pain/Pressure on Exertion	32	Past Four Weeks	During the past 4 weeks, how often have you had any of the following problems: Chest pain or pressure when you exercise**	1 = All of the time 2 = Most of the time 3 = Some of the time 4 = A little of the time 5 = None of the time	.067	.018
Calm and Peaceful	33	Past Four Weeks	How much of the time during the past 4 weeks…Have you felt calm and peaceful?**	1 = All of the time 2 = Most of the time 3 = A good bit of the time 4 = Some of the time 5 = A little of the time 6 = None of the time	-.056	.013
Vigorous Activities	34	Health Limitations/ Difficulties	Does your health limit you in these activities? If so, how much: Vigorous activities**	1 = Yes, limited a lot 2 = Yes, limited a little 3 = No, not limited at all	.086	.029
Eating	35	Health Limitations/ Difficulties	Because of a health or physical problem, do you have any difficulty doing the following activities? Eating**	1 = I am unable to do this activity 2 = Yes, I have difficulty 3 = No, I do not have difficulty	-.132	.042
Hemiparalysis/Weakness	36	Non-Life Threatening Conditions	Have you ever had paralysis or weakness on one side of the body?	1 = Yes, I have it 2 = Yes, but it went away 3 = No	.145	.032
Stroke	37	Life Threatening Conditions	Has a doctor ever told you that you had: Stroke**	1 = Yes 2 = No	-.191	.049
Urination	38	Non-Life Threatening Conditions	Do you have difficulty controlling urination?	1 = Yes 2 = No	.097	.034
Diabetes	39	Life Threatening Conditions	Has a doctor ever told you that you had: Diabetes, high blood sugar, or sugar in the urine**	1 = Yes 2 = No	-.105	.037
Foot Tingling/Burning	40	Past Four Weeks	During the past 4 weeks, how much of the time have you had any of the following problems with your legs and feet? Tingling or burning in your feet especially at night**	1 = All of the time 2 = Most of the time 3 = Some of the time 4 = A little of the time 5 = None of the time	.061	.017

Emotional Problems Limiting Time on Activities	41	Past Four Weeks	During the past 4 weeks, have you had problems with your work or other regular daily activities as a result of any emotional problems? Cut down on the amount of time you spent on work or other activities**	1 = Yes 2 = No	-.085	.038
Comparative Health	42	General Health	How true or false is each of the following statements for you? I am as healthy as anybody I know**	1 = Definitely true 2 = Mostly true 3 = Don't know 4 = Mostly false 5 = Definitely false	.037	.016
Numbness in Feet	43	Past Four Weeks	During the past 4 weeks, how much of the time have you had any of the following problems with your legs and feet? Numbness or loss of feeling in your feet**	1 = All of the time 2 = Most of the time 3 = Some of the time 4 = A little of the time 5 = None of the time	-.038	.017
Feeling Worn Out	44	Past Four Weeks	How much of the time during the past 4 weeks…Did you feel worn out?**	1 = All of the time 2 = Most of the time 3 = A good bit of the time 4 = Some of the time 5 = A little of the time 6 = None of the time	.035	.013
Depression Much of the Time	45	Depression	In the past year, have you felt depressed or sad much of the time?	1 = Yes 2 = No	-.108	.046
Walking Several Blocks	46	Health Limitations/ Difficulties	Does your health limit you in these activities? If so, how much: Walking several blocks**	1 = Yes, limited a lot 2 = Yes, limited a little 3 = No, not limited at all	-.087	.039
Pep	47	Past Four Weeks	How much of the time during the past 4 weeks…Did you feel full of pep?**	1 = All of the time 2 = Most of the time 3 = A good bit of the time 4 = Some of the time 5 = A little of the time 6 = None of the time	.037	.016
Intercept	N/A	N/A	N/A	N/A	2.045	.378

* This appendix was adapted from the appendix in Finkelman et al. (2011).
** Item was among a set of questions with the same stem. Emphasis was not added but was part of the item originally.