Stochastic Curtailment of Questionnaires for Three-Level Classification: Shortening the CES-D for Assessing Low, Moderate, and High Risk of Depression

Applied Psychological Measurement
2016, Vol. 40(1) 22–36
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615592294
apm.sagepub.com

Niels Smits¹, Matthew D. Finkelman², and Henk Kelderman³

Abstract

In clinical assessment, efficient screeners are needed to ensure low respondent burden. In this article, Stochastic Curtailment (SC), a method for efficient computerized testing for classification into two classes for observable outcomes, was extended to three classes. In a post hoc simulation study using the item scores on the Center for Epidemiologic Studies–Depression Scale (CES-D) of a large sample, three versions of SC, SC via Empirical Proportions (SC-EP), SC via Simple Ordinal Regression (SC-SOR), and SC via Multiple Ordinal Regression (SC-MOR) were compared at both respondent burden and classification accuracy. All methods were applied under the regular item order of the CES-D and under an ordering that was optimal in terms of the predictive power of the items. Under the regular item ordering, the three methods were equally accurate, but SC-SOR and SC-MOR needed less items. Under the optimal ordering, additional gains in efficiency were found, but SC-MOR suffered from capitalization on chance substantially. It was concluded that SC-SOR is an efficient and accurate method for clinical screening. Strengths and weaknesses of the methods are discussed.

Keywords

stochastic curtailment, clinical assessment, computerized classification testing

Introduction

In clinical psychology, self-report questionnaires have become increasingly important, both in research settings and clinical practice. Examples of such questionnaires are the Center for

¹VU University Amsterdam, Amsterdam, The Netherlands

²Tufts University School of Dental Medicine, Boston, MA, USA

³Leiden University, The Netherlands

Corresponding Author:

Niels Smits, Department of Methods, Faculty of Psychology and Education, VU University Amsterdam, Van der Boerhorststraat 1, 1081 BT Amsterdam, The Netherlands.

Email: n.smits@vu.nl

Epidemiologic Studies–Depression Scale (CES-D; Radloff, 1977) and the Hamilton Anxiety Scale (Hamilton, 1959). In clinical practice, these questionnaires have been used to screen for subjects with a high probability of pathology. Respondents who screen positive are to be further assessed using a clinical interview to determine whether a diagnosis, such as major depression, is warranted and, if appropriate, treated (Mitchell & Coyne, 2009).

Answering a large number of items can be a burden to respondents, which may not only reduce the quality of responses provided (Herzog & Bachman, 1981) but may also decrease the willingness to fill out questionnaires (Mitchell & Coyne, 2009; Wacholder, Carroll, Pee, & Gail, 1994). Therefore, clinicians have endeavored to develop efficient screeners (Rogers, Wilson, Bungay, Cynn, & Adler, 2002). One approach is to convert an existing questionnaire into a short form (e.g., Burisch, 1997), but this usually means a decline in diagnostic accuracy (Emons, Sijtsma, & Meijer, 2007). A more principled method is computerized classification testing (CCT), which was developed in educational settings with the purpose of separating two or more groups, such as masters and non-masters (e.g., Parshall, Spray, Kalohn, & Davey, 2002). In CCT, the items are administered via a computer, and during the assessment, it is determined whether enough evidence has been collected in favor of one category, and if so, testing is terminated and the appropriate classification is given (Smits & Finkelman, 2013). Under this approach, diagnostic accuracy is fixed at an acceptable level, and the number of items administered is allowed to vary among respondents. CCT has been used for two types of variables: latent traits and observable outcomes (Finkelman, Kim, He, & Lai, 2013). It is uncertain whether CCT for latent variables is useful for developing efficient screeners, because it employs item response theory (IRT), but mental health instruments often fail to meet the requirements of IRT models (Fayers, 2007). Moreover, the implementation of its algorithm is said to be rather elaborate (e.g., Fokkema, Smits, Finkelman, Kelderman, & Cuijpers, 2014), and it may, therefore, be difficult to put it into action in clinical practice. Because the item responses to screeners are often totaled to obtain a sum score, which is an observable outcome, CCT for observable outcomes may be more appropriate for developing efficient screeners.

Recently, Finkelman, He, Kim, and Lai (2011) and Finkelman, Smits, Kim, and Riley (2012) examined a method for CCT for observable outcomes called stochastic curtailment (SC). SC is a sequential analysis method that stems from the field of clinical trials (e.g., Leung, Wang, & Amar, 2003) and is a method for computerized testing designed for classification into categories, such as “at risk” and “not at risk,” and it stops testing when it is improbable that not yet administered items would change the respondent’s classification. After each item, the probability that the full-length test will result in each classification (“at risk” and “not at risk”), given the current set of responses, is estimated. It stops when the probability of either classification becomes higher than some acceptable value. Although the construction phase of SC may be rather elaborate, the implementation of the algorithm is fairly easy as its rules for stopping can be written as a simple lookup table. The usefulness of SC as a CCT method for observable outcomes has been illustrated in several studies (see Finkelman et al., 2011; Finkelman, Kulich, Zoukhri, Smits, & Butler, 2013; Finkelman et al., 2012; Fokkema, Smits, Finkelman, et al., 2014; Fokkema, Smits, Kelderman, Carlier, & van Hemert, 2014).

Screeners are commonly used for classification into two categories, but sometimes, clinicians use more than two levels, such as “low risk,” “moderate risk,” and “high risk.” For example, the CES-D has been used to divide respondents into three or four of such risk groups (e.g., Bay, Hagerty, & Williams, 2007; Cole et al., 2007). Whereas CCT methods for three or more level classification were developed for latent traits (see Eggen, 1999, 2010; Eggen & Straetmans, 2000; Gnambs & Batinic, 2011; Wouda & Eggen, 2009), such methods have not yet been developed for observable outcomes. Therefore, a generalization of SC for classification into more than two levels is needed.

This article has the following goals: (a) show how screeners may be curtailed for three-level classification by generalizing the SC method for two-level classification to situations with three levels and (b) show how well these methods perform in a simulated administration of the CES-D in a general population.

Full Administration Scoring

Assume a screening instrument, which consists of K Likert-type items with $M + 1$ answer categories each, scored on a 0 to M rating scale. The respondent's total score X is obtained by summing the K item scores, and it varies between 0 and MK . Cut-score values C_1 and C_2 , $C_1 < C_2$, are used to transform X into three classes: If $X < C_1$, then the respondent is classified as "low risk" (coded as 0); if $C_1 \leq X < C_2$, then the respondent is classified as "moderate risk" (coded as 1); and if $X \geq C_2$, then a "high risk" classification (coded as 2) is given.

As an example, consider the CES-D, which consists of 20 items ($K = 20$) and is scored on a four category Likert-type scale ($M = 3$) that has been used for classification into three categories of the criterion variable (e.g., Cole et al., 2007; Cramer, Blum, Fanning, & Reed, 2004). The current study considers three-level classification using cutoffs $C_1 = 16$ and $C_2 = 27$, because this was the pair of cutoffs most often encountered in the literature.

Curtailment

Because SC is an extension of it, the idea of curtailment, which is sometimes referred to as the "countdown method" (e.g., Butcher, Keller, & Bacon, 1985), will be discussed first. During the assessment, suppose $k < K$ items have been administered, giving thus far a set of k item scores $\{x_1, \dots, x_k\}$. These scores are used to obtain a cumulative score X_k . The goal is to assign class membership on the basis of X_k . At stage $k < K$, the following rules are used:

1. If $X_k + M(K - k) < C_1$, the respondent is classified as "low risk" (0).
2. If $X_k \geq C_1$ and $X_k + M(K - k) < C_2$, a "moderate risk" (1) classification is given.
3. If $X_k \geq C_2$, the respondent is classified as "high risk" (2).
4. If testing is not halted in Steps 1 to 3, another item is presented.

These rules may be compiled in a lookup table: For each row k of the table, the stopping boundaries for each of the three classification levels are specified. This lookup table is easily implemented in a computer program. The left-hand part of Table 1 provides an example of a lookup table for curtailing the CES-D with cut scores $C_1 = 16$ and $C_2 = 27$.

SC

The idea of SC is the same as for curtailment: In both cases, the impact of future scores on the final classifications is evaluated. However, unlike curtailment, SC allows for early stopping if the probability of a correct classification is sufficiently high (Finkelman et al., 2012). This means that SC always stops, or will have stopped already, whenever curtailment stops. In this article, three ways of obtaining early stopping probabilities are applied. Like Finkelman et al. (Finkelman et al., 2011; Finkelman et al., 2012), the authors will use SC via empirical proportions (SC-EP) and SC via logistic regression, but because three instead of two categories were used, SC via logistic regression is extended to more than two categories and is called SC via simple ordinal regression (SC-SOR). The third method, SC via multiple ordinal regression (SC-MOR), is an extension of SC-SOR. Compared with curtailment, these methods require two

Table 1. Stopping Boundaries for Three-Level Classification of the CES-D Using Curtailment, SC-EP, and SC-SOR Under Regular Item Ordering.

Items administered (k)	Curtailment			SC-EP			SC-SOR		
	Low risk stopping	Moderate risk stopping	High risk stopping	Low risk stopping	Moderate risk stopping	High risk stopping	Low risk stopping	Moderate risk stopping	High risk stopping
1	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	NA	NA	NA	NA	NA	NA	$X_k = 0$	NA	$X_k \geq 9$
5	NA	NA	NA	NA	NA	NA	$X_k \leq 1$	NA	$X_k \geq 10$
6	NA	NA	NA	NA	NA	NA	$X_k \leq 1$	NA	$X_k \geq 12$
7	NA	NA	NA	NA	NA	NA	$X_k \leq 2$	NA	$X_k \geq 14$
8	NA	NA	NA	NA	NA	$X_k \geq 23$	$X_k \leq 2$	NA	$X_k \geq 15$
9	NA	NA	$X_k = 27$	NA	NA	$X_k \geq 23$	$X_k \leq 2$	NA	$X_k \geq 16$
10	NA	NA	$X_k \geq 27$	NA	NA	$X_k \geq 23$	$X_k \leq 3$	NA	$X_k \geq 17$
11	NA	NA	$X_k \geq 27$	$X_k = 0$	NA	$X_k \geq 23$	$X_k \leq 3$	NA	$X_k \geq 20$
12	NA	NA	$X_k \geq 27$	$X_k \leq 2$	$X_k = 13$	$X_k \geq 24$	$X_k \leq 4$	NA	$X_k \geq 21$
13	NA	NA	$X_k \geq 27$	$X_k \leq 2$	$X_k = 13$	$X_k \geq 24$	$X_k \leq 4$	NA	$X_k \geq 21$
14	NA	NA	$X_k \geq 27$	$X_k \leq 4$	$X_k = 13$	$X_k \geq 24$	$X_k \leq 6$	NA	$X_k \geq 22$
15	$X_k = 0$	NA	$X_k \geq 27$	$X_k \leq 6$	$13 \leq X_k \leq 15$	$X_k \geq 25$	$X_k \leq 7$	NA	$X_k \geq 23$
16	$X_k \leq 3$	NA	$X_k \geq 27$	$X_k \leq 8$	$14 \leq X_k \leq 17$	$X_k \geq 25$	$X_k \leq 8$	$16 \leq X_k \leq 17$	$X_k \geq 25$
17	$X_k \leq 6$	$16 \leq X_k \leq 17$	$X_k \geq 27$	$X_k \leq 9$	$14 \leq X_k \leq 19$	$X_k \geq 25$	$X_k \leq 9$	$16 \leq X_k \leq 19$	$X_k \geq 26$
18	$X_k \leq 9$	$16 \leq X_k \leq 20$	$X_k \geq 27$	$X_k \leq 12$	$14 \leq X_k \leq 20$	$X_k \geq 27$	$X_k \leq 11$	$16 \leq X_k \leq 21$	$X_k \geq 26$
19	$X_k \leq 12$	$16 \leq X_k \leq 23$	$X_k \geq 27$	$X_k \leq 13$	$16 \leq X_k \leq 24$	$X_k \geq 27$	$X_k \leq 13$	$16 \leq X_k \leq 23$	$X_k \geq 27$
20	$X_k \leq 15$	$16 \leq X_k \leq 26$	$X_k \geq 27$	$X_k \leq 15$	$16 \leq X_k \leq 26$	$X_k \geq 27$	$X_k \leq 15$	$16 \leq X_k \leq 26$	$X_k \geq 27$

Note. Three-level classification used cutoffs $G_1 = 16$ and $G_2 = 27$. CES-D = Center for Epidemiologic Studies–Depression Scale; SC-EP = stochastic curtailment via empirical proportions; SC-SOR = stochastic curtailment via simple ordinal regression.

additional things. First, the test developer has to specify a confidence level for each of the classification levels (“low risk,” “moderate risk,” and “high risk”): γ_0 , γ_1 , and γ_2 , respectively. Higher values represent a higher classification accuracy and are associated with a higher number of administered items; typical values are .95 and .99. Second, they require a large sample from the population for which the procedure is constructed. Each of the three SC methods involves a sequence of models, and the analysis of this training sample is performed to determine (a) the largest value of X_k sufficient for early stopping at time k in favor of a “low risk” classification, (b) the range of values sufficient for a “moderate risk” classification, and (c) the lowest value of X_k sufficient for early stopping at time k in favor of a “high risk” classification.

SC-EP

SC-EP starts by splitting up the training set into three subsets according to the full-length total score: T^0 consisting of respondents scoring below the first threshold, $X < C_1$; T^1 consisting of respondents scoring in between the two thresholds, $C_1 \leq X < C_2$; and T^2 consisting of respondents scoring at or above the second threshold, $X \geq C_2$. Then, these subsets are used to determine the probability of each classification conditional on the cumulative scores X_k . To that end, for each $k < K$, for each possible $X_k = \{0, \dots, Mk\}$, the empirical rest scores on the $K - k$ remaining items and X_k are added up to obtain a distribution of synthetic total scores. Next, in each subset T^i , the proportion $\hat{P}_k^{ij}$ of cases who receive risk classification j based on the full-length test is determined. First, for low risk stopping, the maximal value of X_k , if any, for which the confidence level is attained in each of the three subsets of the training set is determined:

$$X_k^{0*} \text{ is the highest } X_k \in \{0, \dots, Mk\}, \text{ for which } \min(\hat{P}_k^{00}, \hat{P}_k^{10}, \hat{P}_k^{20}) \geq \gamma_0. \quad (1)$$

Second, for moderate risk stopping, the range of values, if any, for which the confidence level is attained in each of the three subsets of the training set is determined:

$$X_k^{1*} \text{ is the range of } X_k \in \{0, \dots, Mk\}, \text{ for which } \min(\hat{P}_k^{01}, \hat{P}_k^{11}, \hat{P}_k^{21}) \geq \gamma_1. \quad (2)$$

Third, for high risk stopping, the minimal value, if any, for which the confidence level is attained in each of the three subsets of the training set is determined:

$$X_k^{2*} \text{ is the lowest } X_k \in \{0, \dots, Mk\}, \text{ for which } \min(\hat{P}_k^{02}, \hat{P}_k^{12}, \hat{P}_k^{22}) \geq \gamma_2. \quad (3)$$

The rationale behind SC-EP is to first assume that a new respondent's future answers will be like those of past respondents who received a “low risk” classification. Next, it is assumed that the future answers will be like those of past respondents who received a “moderate risk” classification, and finally, it is assumed that future answers will be like those of past respondents with a “high risk” classification. Testing is only halted to give classification j if the probability of this classification is at least γ_j under *all three* assumptions. SC-EP is a conservative method (Finkelman et al., 2012), because for good screeners, the $\hat{P}_k^{ij}$ is ordered. For example, commonly $\hat{P}_k^{00} > \hat{P}_k^{10} > \hat{P}_k^{20}$, and therefore, the lowest probability, $\hat{P}_k^{20}$, is used for deciding whether the likelihood of a “low risk” classification is high enough for early stopping.

Appendix A of the online supplement discusses the monotonicity of stopping boundaries. Finally, if the boundaries meet this requirement, they are compiled in a lookup table, which can be implemented in a computer program for assessing new respondents.

SC via Simple Ordinal Regression

SC-SOR uses a statistical model calibrated in the training set to estimate the probabilities that the full-length instrument will make “low risk,” “moderate risk,” and “high risk” classifications. Finkelman et al. (2012) used binary logistic regression for two-level classification, and for three-ordered levels, both binary and ordinal logistic regression models are used.

The first step of SC-SOR is to deal with the issue of separability (e.g., Albert & Anderson, 1984). In log linear models, separability occurs when data points show little or no overlap such that the classes are (almost) perfectly separated, which may give non-unique infinite parameter estimates. For two-level classification, Finkelman et al. (2012) noted that including members of the training set whose cumulative scores are high or low enough so that their classification statuses are deterministic would be at odds with the notion of a probabilistic model. In the case of three or more ordered categories, however, the issue is broader: Even if classification statuses are not deterministic given X_k , some classes may have structural zero counts (e.g., Bishop, Fienberg, & Holland, 1975, chapter 5). Online Appendix B provides an illustration of this issue.

If one would fit a single model on a training set suffering from separation, this would have detrimental consequences for its parameter estimates (Albert & Anderson, 1984). To deal with the issue, similar to Finkelman et al. (2012), at each step $k < K$, the training set is split up into four potential subsets. The first subset consists of cumulative scores for which the final classification is certain (see the first three rules in the “Curtalement” section; note that this subset consists of three types of such scores). These deterministic scores are removed from the training set. The second subset has X_k values for which all three classifications are possible. The third subset consists of respondents for whom X_k is too low to result in a highest level classification: Only classification 0 or 1 can be obtained. The fourth subset consists of those respondents for whom X_k is too high to result in a lowest level classification: Only classification 1 or 2 can be obtained. For each of the last three subsets, a binary or ordinal regression model is fitted. The associated probabilities are then used to determine the boundaries for early stopping. Below, a description of the model fitting and stopping boundary determination in each subset is given.

1. In the subsets for which $X_k < C_1$ and $X_k + M(K - k) \geq C_2$ (all three levels, “low risk,” “moderate risk,” and “high risk,” are possible), an ordinal regression model is estimated. There are several models for ordinal outcomes, which differ in the type of logits used (Agresti, 2002, chapter 7). Here, the most popular model, the proportional odds model is used, which uses cumulative logits. Under this model, the probability $\hat{P}_k^{j*}$ of scoring class j or higher on the full-length test, conditional on X_k , is given by

$$\hat{P}_k^{j*} = [1 + \exp(\hat{\alpha}_j + \hat{\beta}_k X_k)]^{-1}, j = 0, 1, 2, \quad (4)$$

where $\hat{P}_k^{0*}$ is 1. The probability for class j is obtained by subtracting cumulative probabilities:

$$\hat{P}_k^j = \hat{P}_k^{j*} - \hat{P}_k^{j+1*}, j = 0, 1, 2, \quad (5)$$

where $\hat{P}_k^{j+1*}$ is 0 for $j = 2$. First, for low risk stopping, the maximal value of X_k , if any, for which the confidence level is attained is determined:

$$X_k^{0*} \text{ is the highest } X_k \in \{0, \dots, Mk\}, \text{ for which } \hat{P}_k^0 \geq \gamma_0, \quad (6)$$

where the prime is used to indicate that the boundary has a different value and interpretation than that of Equation 1. Second, for moderate risk stopping, the range of values, if any, for which the confidence level is attained is determined:

$$X_k'^{1*} \text{ is the range of } X_k \in \{0, \dots, Mk\}, \text{ for which } \hat{P}_k^1 \geq \gamma_1. \quad (7)$$

Third, for high risk stopping, the minimal value, if any, for which the confidence level is attained in each of the three subsets of the training set is determined:

$$X_k'^{2*} \text{ is the lowest } X_k \in \{0, \dots, Mk\}, \text{ for which } \hat{P}_k^2 \geq \gamma_2. \quad (8)$$

2. In the subset for which $X_k < C_1$ and $X_k + M(K - k) < C_2$ (only levels “low risk” and “moderate risk” are possible), a logistic regression model is fitted. Under this model, the probability of a “moderate risk” (coded as 1) classification is given by

$$\hat{P}_k^1 = \left[1 + \exp(\hat{\alpha}' + \hat{\beta}_k' X_k) \right]^{-1}, \quad (9)$$

where the prime is used to indicate that the estimates have different values and interpretations than those of Equation 4. First, for low risk stopping, the maximal value of X_k , if any, for which the confidence level is attained is determined:

$$X_k''^{0*} \text{ is the highest } X_k \in \{0, \dots, Mk\}, \text{ for which } (1 - \hat{P}_k^1) \geq \gamma_0, \quad (10)$$

where the double prime is used to indicate that the boundary has different values and interpretations than that of Equation 6. Second, for moderate risk stopping, the range of values, if any, for which the confidence level is attained is determined:

$$X_k''^{1*} \text{ is the lowest } X_k \in \{0, \dots, Mk\}, \text{ for which } \hat{P}_k^1 \geq \gamma_1. \quad (11)$$

3. In the subset for which $C_1 \leq X_k < C_2$ and $X_k + M(K - k) \geq C_2$ (only levels “moderate risk” and “high risk” are possible), a logistic regression model is estimated. Under this model, the probability of a “high risk” (coded as 2) classification is given by

$$\hat{P}_k^2 = \left[1 + \exp(\hat{\alpha}'' + \hat{\beta}_k'' X_k) \right]^{-1}, \quad (12)$$

where the double prime is used to indicate that the estimates have different values and interpretations than those of Equations 4 and 9. First, for moderate risk stopping, the maximal value of X_k , if any, for which the confidence level is attained is determined:

$$X_k'''^{0*} \text{ is the highest } X_k \in \{0, \dots, Mk\}, \text{ for which } (1 - \hat{P}_k^2) \geq \gamma_1, \quad (13)$$

where the triple prime is used to indicate that the boundary has different values and interpretations than that of previous equations. Second, for high risk stopping, the minimal value, if any, for which the confidence level is attained is determined:

$$X_k'''^{1*} \text{ is the lowest } X_k \in \{0, \dots, Mk\}, \text{ for which } \hat{P}_k^2 \geq \gamma_2. \quad (14)$$

Online Appendix C describes SC-MOR, a special case of SC-SOR with a regression weight for each item.

Method

Data

The scores to the items of the CES-D were taken from the Wisconsin Longitudinal Study (WLS; 2005), a study that has collected data of a group of high school students who graduated in 1957. The sample used consists of 6,126 people aged between 64 and 65 years (53% women) with complete data on all CES-D items. The CES-D was administered through either a telephone interview or mail survey in 2003–2005. A more thorough description of the data is provided on the study's website.

The CES-D (Radloff, 1977) is often used as a first-stage screener for depression. It contains 20 items, and each item asks the respondent to answer a question about his or her status during the past week in terms of the number of days a complaint was present. In the standard version, severity is expressed on a 4-point rating scale with categories 0 days (0), 1 to 2 days (1), 3 to 4 days (2), and 5 to 7 days (3). For grading the CES-D, the 4 negatively worded items are reverse-scored. Consequently, its sum score ranges from 0 (*no depressive complaints at all*) to 60 (*many depressive complaints*). In the WLS, however, an 8-point item scale (0 to 7 days) was used, and to be able to generalize the outcomes of the current study to other CES-D administrations, the items were recoded into the original 0 to 3 scale. Traditionally, when screening for clinical depression, a single cutoff of ≥ 16 has been used, but in the current simulation study, three-level classification was applied with cutoffs $C_1 = 16$ and $C_2 = 27$ (e.g., Cole et al., 2007; Cramer et al., 2004). The resulting classification levels were defined as “low risk,” “moderate risk,” and “high risk” of depression.

Simulation Design

The purpose of the current simulation study was to compare the four sequential methods in terms of respondent burden and classification properties under both the regular item ordering and an optimal item ordering (which is described in Online Appendix D).

The three SC methods need a large data set for calibration. Using these data for both calibration and validation in the current simulation study would in all likelihood yield results that are overly optimistic (Hastie, Tibshirani, & Friedman, 2009, chapter 7). To avoid such capitalization on chance, the data set was randomly split into two parts (3,063 observations each), a “training” set and a “validation” set. The training set was used to estimate the model parameters and to determine stopping boundaries that were converted into lookup tables. In the validation set, each procedure was applied using its lookup table(s), and the procedures were compared in terms of their classification performance and efficiency.

In each of the SC methods, confidence level constants for low, moderate, and high-risk classification were set equal: $\gamma_1 = \gamma_2 = \gamma_3$ (i.e., required accuracy is equal for each level).

The curtailment and SC methods were evaluated via two types of outcome measures. As measures of respondent burden, average test length (*SD*) and median test length (range) were used. Concordance with the full-length CES-D was assessed using two kinds of measures. First, classifications were treated as nominal, and concordance was assessed using the proportion of concordant classification and Cohen's kappa. Second, classifications were treated as ordinal, and concordance was assessed using Goodman and Kruskal's gamma and Spearman's rank correlation coefficient.

Obviously, respondent burden and accuracy are related: The more items are administered, the higher classification accuracy is. Consequently, methods using more items are expected to be more accurate than methods using less items, which makes a direct comparison troublesome.

An approach to deal with this issue is to match the different procedures based on concordance and compare them with respect to test length or vice versa (Thompson, 2011). In the current simulation study, it was chosen to select for each SC method those γ values in the training set which resulted in a proportion of concordant classification of 0.99. With accuracy level fixed, the efficiency of the SC methods was directly comparable. Note that curtailment always has a proportion of concordant classification of 1 and can, therefore, not be fixed at a lower value; it was added in the current study to provide a benchmark for respondent burden.

All simulations were conducted in the statistical environment R (R Core Team, 2013).

Results

Among the 6,126 respondents who took the full-length CES-D, the sum score ranged from 0 to 53 ($M = 8.49$, $SD = 7.81$). The average inter-item correlation was .28, and Cronbach's α for the whole scale was .88. Using cutoffs 16 and 27, 5,220 respondents (85.2%) were classified as low risk, 691 respondents (11.3%) were classified as moderate risk, and 215 respondents (3.5%) were classified as high risk.

Concordance With Full-Length CES-D

Although the SC methods were matched at their concordance with the full-length CES-D, concordance outcomes were studied for two reasons. First, a single measure, proportion of concordant classification, was used for matching, whereas four measures were used to assess concordance; the alternative measures may still yield different results. Second, the procedures were matched in the training set, and due to differences in generalization error (the extent to which capitalization on chance occurs; Hastie et al., 2009), methods may actually show differences in the validation set.

Table 2 shows concordance results; the left-hand part gives results under the regular item ordering and the right-hand part under the optimal item ordering. The table provides classification tables with the classification based on the sequential method in the columns and the full-length CES-D classification in the rows. In addition, it shows proportion (P) of concordant classification, Cohen's kappa (κ), and Spearman's rank correlation (ρ). Because under both item orderings Goodman and Kruskal's gamma showed no differences among the methods, it was left from the table. The first row of the table shows outcomes of curtailment, which confirms that this method always makes identical classifications as the full-length CES-D.

Turning to a comparison of the three stochastic methods, under the regular item ordering, all three had a proportion of concordant classification of about 99%, showing no (or very little) generalization error. Cohen's kappa and Spearman's rank correlation were very high for all methods. In addition, these measures showed small differences among the methods, with SC-EP giving marginally better and SC-MOR marginally worse outcomes. Under the optimal item ordering, differences were more apparent, with SC-SOR giving marginally better results than SC-EP, and with SC-MOR giving substantially worse results. SC-MOR suffered from generalization error, giving a worse proportion of concordant classification in the validation set than in the training set. For SC-SOR, using the optimal item ordering was an improvement over using the regular ordering, whereas for the other two methods, it meant a decline.

Respondent Burden

Table 3 shows the number of items used for all sequential methods under the two item orderings. Under both item orderings, curtailment needed on average about 85% of the full item set,

Table 2. Classification Results for Curtailment, SC-EP, SC-SOR, and SC-MOR Under Regular and Optimal Item Ordering.

Method	Regular item ordering			Optimal item ordering				
	Classification table			Classification table				
	P	κ	ρ	P	κ	ρ		
Curtailment	Early stopping							
	Full	0	1	2	Early stopping			
	0	2,603	0	0	0	2,603	0	0
	1	0	353	0	1	0	353	0
	2	0	0	107	2	0	0	107
SC-EP	Early stopping							
	Full	0	1	2	Early stopping			
	0	2,590	13	0	0	2,594	9	0
	1	4	348	1	1	14	339	0
	2	0	10	97	2	0	15	92
SC-SOR	Early stopping							
	Full	0	1	2	Early stopping			
	0	2,601	0	2	0	2,603	0	0
	1	26	325	2	1	24	326	3
	2	0	3	104	2	0	2	105
SC-MOR	Early stopping							
	Full	0	1	2	Early stopping			
	0	2,601	0	2	0	2,555	48	0
	1	25	318	10	1	35	312	6
	2	0	7	100	2	0	38	69
	0	2,601	0	2	0	2,555	48	0
	1	25	318	10	1	35	312	6
	2	0	7	100	2	0	38	69

Note. P is proportion of concordant classifications, κ is Cohen's kappa, and ρ is Spearman's rank correlation. SC-EP = stochastic curtailment via empirical proportions; SC-SOR = stochastic curtailment via simple ordinal regression; SC-MOR = stochastic curtailment via multiple ordinal regression.

Table 3. Respondent Burden for Curtailment, SC-EP, SC-SOR, and SC-MOR Under Regular and Optimal Item Ordering.

Method	Regular item ordering				Optimal item ordering			
	<i>M</i>	<i>SD</i>	Median	Range	<i>M</i>	<i>SD</i>	Median	Range
Curtailment	17.02	1.59	17	11-20	17.13	1.65	17	9-20
SC-EP	13.76	2.64	14	9-20	11.65	2.84	11	8-20
SC-SOR	8.40	5.74	5	2-20	6.48	5.98	2	2-20
SC-MOR	8.46	5.45	6	2-20	5.83	5.07	2	2-19

Note. SC-EP = stochastic curtailment via empirical proportions; SC-SOR = stochastic curtailment via simple ordinal regression; SC-MOR = stochastic curtailment via multiple ordinal regression.

which was substantially more than the SC methods. In addition, using the optimal item ordering instead of the regular item ordering gave a reduction in items for the SC methods but not for curtailment. Under the regular item ordering, SC-EP needed substantially more items (68.8% of the total set) than the two regression-based methods (which had very similar outcomes, each using about 42% of the total set). Likewise, under the optimal item ordering, SC-EP used substantially more items (58% of the total item set) than the regression-based methods (about 30% of the item set). SC-SOR and SC-MOR, however, showed a somewhat larger difference in average number of items used, with SC-MOR using less items.

Discussion

In the present study, the SC methods by Finkelman et al. (2012) for two-level classification were generalized to the case of three-level classification. The generalization of the method that used logistic regression proved to be extensive in terms of the construction steps needed. In a post hoc simulation using CES-D item scores, it was shown that SC also works in the case of three-level classification: Assessments were both efficient and accurate. In addition, using an optimal item ordering in the sense that the most predictive items were used first, as suggested by Finkelman, Kim, et al. (2013), showed that efficiency could even be improved. SC-SOR showed the best results; under the optimal item ordering on average, about 30% of the item set was needed and the classifications showed a very high congruency with the classifications based on the full-length CES-D. SC-EP was not as efficient but had comparable classification accuracy. By contrast, SC-MOR, a special case of SC-SOR with a regression weight for each item (see Online Appendix C), performed somewhat worse than SC-SOR. Although it exhibited similar respondent burden, it had worse classification accuracy, and moreover, it suffered from capitalization on chance.

In the current study, performance and gains in efficiency for three-category classification were somewhat lower than for two-category classification (see Appendices E and F of the online supplement), which is not surprising because the former is generally a more difficult task than the latter. However, the three-category procedures performed similarly to the two-category classification methods of Finkelman et al. (2012) and Fokkema, Smits, Finkelman, et al. (2014). This may be due to differences in test score distributions; the previous studies were performed in clinical populations, whereas the current study used data from a general population. This also shows that the methods for efficient assessment are population dependent, which means that the early stopping rules may only be valid for respondents for whom the training sample is representative. Because classification models may be dissimilar between populations, applying a method

that is trained in one population in another population may lead to misleading results. Another simulation on the current data, the results of which are not shown herein, suggests that if SC is developed using a training set that mainly consists of higher scores, the method also works well in a validation set consisting of lower scores, whereas using a training set consisting of lower scores gives inferior results in a validation set consisting of higher scores. This result may originate from psychopathology items being informative on only one side of the scale (Reise & Waller, 2009).

In the current study, the effect of item ordering was assessed (see Appendices D and F). Whether the ordering used was truly optimal may be hard to tell because it was assessed post hoc, that is, after collecting the data in the regular item ordering. It is not unlikely that the predictive power of an item is dependent upon the position of the item in the questionnaire. For example, answering may become sloppy at the end of the questionnaire, or learning may take place while answering the test (e.g., McFarland, 1981). Therefore, the predictive power of the items may change under a different item ordering. To validate the alleged optimal item ordering, the predictive power of the items should be studied in a new sample where the items are administered in the proposed order.

A related issue is that when SC-MOR (see Online Appendix C) was constructed under the optimal item ordering, in fact, two kinds of parameters were estimated: regression parameters and item ordering. Because the same training set was used twice, chance may have played a bigger role than for the other SC methods, which may explain the inferior results. Further inspection showed that the optimal item ordering was somewhat different in the training and validation set, so that it is very likely that SC-MOR suffered from capitalization on chance on both kinds of parameters. Based on these outcomes, it is advised to assess item ordering in another training set than the training set used for model estimation.

Table B1 (see online supplement) shows that for SC-SOR under the optimal item ordering, classification is already possible after administering two items. Some readers may feel that this number of items is too low for a valid assessment. It should be noted, however, that such an outcome is typical for CCT methods in general (see, for example, Weiss, 1982), because for individuals with extreme scores, it is rather easy to determine on what side of the cut score they will fall. A related issue is that when the respondent makes a mistake in selecting the right answer category (e.g., a “0” instead of “1”), using only a few items might result in an incorrect classification. However, the current study shows empirically that when the test stops after two items, the full-length test has a high probability of making the same decision as the shorter version.

For the best method in this study, SC-SOR, stopping rules are presented in Table 1 and Online Appendix Table B1. It is uncertain whether these tables should be used by clinicians, because in practice, cut scores other than 16 and 27 may be preferred. In addition, more or less efficiency than the quantities presented in this study may be required by the tester, and in the current study, accuracy was fixed at an arbitrary level to be able to compare methods. Moreover, classification accuracy was set equal for all classes. The interested clinician is advised to use a large training set from the right population and use his or her own specific set of requirements when building an SC method. To that end, the R code used in this study can be freely obtained from the first author.

The procedures for three-level classification presented here may be extended toward situations with more than three levels. However, with increasing number of classification levels, the number of *subsets* of classification levels that are theoretically possible given X_k (see Online Appendix Table A1) increases rapidly. This number is a triangular number (Wells, 1986, pp. 47-48): With n ordered classes, the number of potential subsets of classification levels is $\frac{1}{2}n(n+1)$. In the case of three classes, this number equals 6; for four classes, it equals 10; and for five classes, it equals 15, and so on. As a result, using more than three levels means a

substantial increase in the construction steps (estimating models for subsets, etc.) needed. It is probably best to not extend the methods beyond four of five class levels. If one wishes to use SC for more levels, it may be more appropriate to use a recently developed SC method for predicting total scores (Smits & Finkelman, 2014).

In this study, to deal with the separability issue, the authors followed Finkelman et al. (2012) by fitting models on subsets of data, which has several disadvantages. First, it is cumbersome to fit multiple models for each k . Second, for each k with adjacent models, monotonicity should be checked, and if it is violated, ad hoc solutions are needed. It may be better to use other methods for dealing with separability, such as those suggested by Gelman, Jakulin, Pittau, and Su (2008) or Rousseeuw and Christmann (2003). Alternatively, a method should be developed that allows for simultaneously estimating all models for step k and in which parameters of adjacent logistic regression models are set equal, similar to what is done in structural equation modeling.

Authors' Note

This research uses data from the Wisconsin Longitudinal Study (WLS) of the University of Wisconsin–Madison. A public use file of data from the WLS is available from the Wisconsin Longitudinal Study (University of Wisconsin–Madison) and at <http://www.ssc.wisc.edu/wlsresearch/data/>. The opinions expressed herein are those of the authors. Since 1991, the Wisconsin Longitudinal Study (WLS) has been supported principally by the National Institute on Aging (AG-9775 AG-21079 and AG-033285), with additional support from the Vilas Estate Trust, the National Science Foundation, the Spencer Foundation, and the Graduate School of the University of Wisconsin–Madison.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The online appendices are available at <http://apm.sagepub.com/supplemental>

References

- Agresti, A. (2002). *Categorical data analyses* (2nd ed.). New York, NY: Wiley.
- Albert, A., & Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71, 1-10.
- Bay, E., Hagerty, B., & Williams, R. A. (2007). Depressive symptomatology after mild-to-moderate traumatic brain injury: A comparison of three measures. *Archives of Psychiatric Nursing*, 21, 2-11.
- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.
- Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality*, 11, 303-315.
- Butcher, J. N., Keller, L. S., & Bacon, S. F. (1985). Current developments and future directions in computerized personality assessment. *Journal of Consulting and Clinical Psychology*, 53, 803-815.
- Cole, A. L., Lee, P. J., Hughes, D. A., Deegan, P. B., Waldek, S., & Lachmann, R. H. (2007). Depression in adults with fabry disease: A common and under-diagnosed problem. *Journal of Inherited Metabolic Disease*, 30, 943-951.

- Cramer, J. A., Blum, D., Fanning, K., & Reed, M. (2004). The impact of comorbid depression on health resource utilization in a community sample of people with epilepsy. *Epilepsy & Behavior*, 5, 337-342.
- Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, 23, 249-261.
- Eggen, T. J. H. M. (2010). Three-category adaptive classification testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 373-387). New York, NY: Springer.
- 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, 713-734. doi: 10.1177/00131640021970862
- Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, 12, 105-120.
- Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: The challenges for health outcomes assessment. *Quality of Life Research*, 16, 187-194.
- 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, 1989-2004.
- Finkelman, M. D., Kim, W., He, Y., & Lai, A. (2013). Item ordering in stochastically curtailed health questionnaires with an observable outcome. *Journal of Computerized Adaptive Testing*, 1, 38-66.
- Finkelman, M. D., Kulich, R. J., Zoukhri, D., Smits, N., & Butler, S. F. (2013). Shortening the current opioid misuse measure via computer-based testing: A retrospective proof-of-concept study. *BMC Medical Research Methodology*, 13(1), Article 126.
- Finkelman, M. D., Smits, N., Kim, W., & Riley, B. (2012). Curtailment and stochastic curtailment to shorten the CES-D. *Applied Psychological Measurement*, 36, 632-658.
- Fokkema, M., Smits, N., Finkelman, M. D., Kelderman, H., & Cuijpers, P. (2014). Curtailment: A method to reduce the length of self-report questionnaires while maintaining diagnostic accuracy. *Psychiatry Research*, 215, 477-482.
- Fokkema, M., Smits, N., Kelderman, H., Carlier, I. V. E., & van Hemert, A. M. (2014). Combining decision trees and stochastic curtailment for assessment length reduction of test batteries used for classification. *Applied Psychological Measurement*, 38, 3-17.
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2, 1360-1383.
- Gnambs, T., & Batinic, B. (2011). Polytomous adaptive classification testing: Effects of item pool size, test termination criterion, and number of cutscores. *Educational and Psychological Measurement*, 71, 1006-1022.
- Hamilton, M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology*, 32, 50-55.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). New York, NY: Springer.
- Herzog, A., & Bachman, J. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, 45, 549-559.
- Leung, D. H.-Y., Wang, Y.-G., & Amar, D. (2003). Early stopping by using stochastic curtailment in a three-arm sequential trial. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52, 139-152.
- McFarland, S. (1981). Effects of question order on survey responses. *Public Opinion Quarterly*, 45, 208-215.
- Mitchell, A. J., & Coyne, J. C. (2009). *Screening for depression in clinical practice: An evidence-based guide*. Oxford, UK: Oxford University Press.
- Parshall, C., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.
- Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1, 385-401. doi:10.1177/014662167700100306
- R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from <http://www.R-project.org/>

- Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Review of Clinical Psychology, 5*, 27-48.
- 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*, 164-175.
- Rousseeuw, P. J., & Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis, 43*, 315-332.
- Smits, N., & Finkelman, M. D. (2013). A comparison of computerized classification testing and computerized adaptive testing in clinical psychology. *Journal of Computerized Adaptive Testing, 1*, 19-37.
- Smits, N., & Finkelman, M. D. (2014). Variable length testing using the ordinal regression model. *Statistics in Medicine, 33*, 488-499.
- Thompson, N. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation, 16*(4), 2.
- Wacholder, S., Carroll, R. J., Pee, D., & Gail, M. H. (1994). The partial questionnaire design for case-control studies. *Statistics in Medicine, 13*, 623-634.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.
- Wells, D. (1986). *The penguin dictionary of curious and interesting numbers*. London, England: Penguin.
- Wisconsin Longitudinal Study. (2005). *Wisconsin Longitudinal Study: Graduates, siblings, and spouses* [Data file]. Madison: University of Wisconsin-Madison. Retrieved from <http://www.ssc.wisc.edu/wlsresearch/documentation>
- Wouda, J. T., & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved from <http://publicdocs.iacat.org/cat2010/cat09wouda.pdf>