Psychometric Contributions to Improving the Efficiency and Fidelity of Clinical Assessment and Research

Fokkema, M.

2015

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl
Chapter 1

Curtailment of Self-Report Questionnaires

Abstract

Minimizing the respondent burden and maximizing the classification accuracy of tests is essential for efficacious screening for common mental health disorders. In previous studies, curtailment of tests has been shown to reduce average test length considerably, without loss of accuracy. In the current study, we simulate deterministic curtailment (DC) and stochastic curtailment (SC) for three self-report questionnaires for common mental health disorders, to study the potential gains in efficiency that can be obtained in screening for these disorders. The curtailment algorithms were applied to an existing dataset of item scores of 502 help-seeking participants. Results indicate that DC reduces test length by up to 37%, and SC reduces test length by up to 46%, with only very slight decreases in diagnostic accuracy. Compared to an item response theory based adaptive test with similar test length, SC provided better diagnostic accuracy. Consequently, curtailment may be useful in improving the efficiency of mental health self-report questionnaires.

1.1 Introduction

As noted by, for example, Gilbody, Sheldon en Wessely (2006) and the UK National Screening Committee (2003), tests used in screening for common mental health disorders should be simple, precise and acceptable to patients. Similarly, in discussing the costs and benefits of screening, Gray en Austoker (1998) noted “All screening programmes do harm; some also do good” (p. 983). Therefore, minimizing the respondent burden, and maximizing the accuracy of tests is essential for efficacious screening.

To reduce the respondent burden of screeners for common mental health disorders, many efforts have been aimed at creating fixed-length short forms of existing self-report questionnaires (e.g., Donker, van Straten, Marks & Cuijpers, 2011; Cuijpers, Smits, Donker, Ten Have & Graaf, 2010;
Rost, Burnam & Smith, 1993). However, for fixed-length short forms, the reduction in test length generally comes at the expense of diagnostic accuracy (Smith, McCarthy & Anderson, 2000; Mitchell & Coyne, 2007).

Over the past few decades, adaptive testing algorithms have been developed, aimed at reducing test length without reducing accuracy (e.g., Weiss, 1982; Van der Linden & Glas, 2010). In every stage of adaptive testing, earlier item responses are used to select the item which is most informative for the current respondent, and items that do not provide additional information are not administered. In general, this results in considerable test length reductions, while the diagnostic accuracy or measurement precision of the original full-length instrument is preserved (e.g., Fliege et al., 2005; Fries, Cell, Rose, Krishnan & Bruce, 2009; Gibbons et al., 2008; Smits, Cuijpers & van Straten, 2011; Walter et al., 2007). However, most adaptive testing algorithms are based on item response theory (IRT), and assume the data satisfies the conditions of a latent trait model. Often, a single latent trait underlying the data is assumed, which may be unrealistic for (mental) health self-report questionnaires (e.g., Fayers, 2007; Gardner, Kelleher & Pajer, 2002; Petersen et al., 2006). In addition, the purpose of classification is prediction of an external criterion, so methods that do not depend on a latent trait model may be preferable for classification (Smits & Finkelman, 2013).

Recently, Finkelman, He, Kim en Lai (2011) and Finkelman, Smits, Kim en Riley (2012) introduced curtailment as a method for reducing the respondent burden of mental health self-report questionnaires, which does not assume latent traits underlying the item scores. Earlier, the statistical properties of curtailment have been studied by Eisenberg en Simons (1978) and Eisenberg en Ghosh (1980), and curtailment has been used for early stopping in clinical trials (e.g., Lan, Simon & Halperin, 1982). The application of curtailment in psychological testing results in variable length tests, in which testing is halted when administration of the remaining items is unable or unlikely to change the final classification decision.
In other words, item administration is continued only as long as the resulting diagnostic outcome is amenable to change.

Consider the example of using the seven-item anxiety symptom subscale of the Hospital Anxiety and Depression Scale (HADS-A; Zigmond & Snaith, 1983) to screen for anxiety disorders, by using a cut-off value for the test score of eight (Bjelland, Dahl, Haug & Neckelmann, 2002). Self-evidently, item administration can be ceased, whenever a respondent obtains a cumulative score of eight or higher, before all items have been administered. Likewise, item administration can be ceased, when it is no longer possible for a respondent to obtain a final score of eight or higher, with the remaining items. In addition, for respondents endorsing the most severe response options (an item score of three) on the first two items, it seems likely that their final test score will exceed the cut off, and therefore further item administration may not be necessary. On the other hand, for respondents endorsing the least severe response options (an item score of zero) on the first two items, it seems likely that their final test score will not exceed the cut off, and further item administration may not be necessary, either. However, for respondents endorsing response options representing more moderate levels of anxiety on the first two items, administration of subsequent items may be necessary to determine whether their final test score will or will not exceed the cut-off value.

Curtailment provides a formalization of this idea, and Finkelman et al. (2011) have developed an algorithm for application of curtailment to health questionnaires. Their method depends on observed scores only, and makes no assumptions about underlying latent traits. In addition, curtailment can be applied deterministically, or stochastically. For application of deterministic curtailment (DC), a cut-off value for classifying respondents as “at risk” is needed. During testing, item administration for a respondent is halted and an “at risk” classification is made, when the remaining items can no longer result in a final test score above or equal to the cut-off value. Item administration is halted and a “not at risk” classi-
1 Curtailment of Self-Report Questionnaires

Curtailment is made, when the remaining items can no longer result in a final test score below the cut-off value.

For application of stochastic curtailment (SC), a cut-off value for classifying respondents as “at risk” is required as well. In addition, for the stochastic part of the algorithm, an existing dataset of item scores is needed, and the user has to specify a value for $\gamma$: the threshold for the probability that the classification decision based on the stochastically curtailed version will match that of the full-length instrument. First, the algorithm is trained, by splitting the complete dataset of item scores in two parts: the “at risk” and the “not at risk” datasets. The “at risk” dataset contains item scores of all respondents with a test score meeting or exceeding the cut-off value; the “not at risk” dataset contains item scores of all respondents with a test score below the cut-off value (Finkelman et al., 2012).

Next, the algorithm is applied for shortening tests for new respondents: for every new respondent, after administration of every item, a cumulative score is calculated. In the “at risk” and “not at risk” datasets, all scores on the remaining items are appended to the cumulative score. When the proportion of resulting test scores meeting or exceeding the cut-off value is $\geq \gamma$ in both the “at risk” and “not at risk” datasets, item administration is halted, and an “at risk” classification is made for the new respondent. Similarly, when the proportion of resulting test scores meeting or exceeding the cut-off value is $\leq (1 - \gamma)$ in both the “at risk” and “not at risk” datasets, item administration is halted, and a “not at risk” classification is made for the new respondent (Finkelman et al., 2012).

An illustration of DC and SC of administration of the HADS-A scale to two new respondents is provided in the Appendix. Although the curtailment algorithms may seem elaborate, for practical application of curtailment, look-up tables with stopping criteria for every item can be created, which are easy to use and implement. It should be noted that application of SC with $\gamma = 1.00$ will generally yield the same results as application of DC. However, SC with $\gamma = 1.00$ may halt testing earlier than DC, when
some response patterns are theoretically possible, but not observed empirically. For example, when the highest response option for the last item is never observed in the training dataset, SC with $\gamma = 1.00$ may halt testing for some respondents before the last item, whereas testing for these respondents would be continued with DC.

The goal of the current article is to illustrate the potential gains in efficiency that may be obtained by application of curtailment in mental health care applications. In what follows, we will illustrate this with a post-hoc simulation of curtailment in an existing dataset of item responses on self-report questionnaires. To provide a benchmark for assessing the performance of curtailment, we will simulate an IRT-based adaptive test, as well. In the methods section, the dataset, algorithms and simulation design will be described. In the results section, the findings in terms of test length reduction and accuracy will be presented. In the discussion, implications of the current study and directions for further research will be presented.

### 1.2 Method

#### 1.2.1 Participants

The dataset used in the current study was collected for development of a fixed length, web-based screener for common mental disorders (Donker, Van Straten, Marks & Cuijpers, 2009; Donker et al., 2011). The total sample consisted of 502 participants, with a mean age of 43 (SD=13, range 18-80). A majority of the subjects (57%) was female. Detailed information about the sample is provided in (Donker et al., 2009). Questionnaires were completed by all 502 participants, and because of computerized questionnaire administration, no data were missing. Diagnoses on DSM-IV disorders were obtained from a subsample of 157 participants.
(Donker et al., 2009, 2011). Of these participants, 29.29% were diagnosed with major depressive disorder, 19.11% were diagnosed with generalized anxiety disorder, and 59.87% were diagnosed with an anxiety disorder.

1.2.2 Measures

Center for Epidemiological Studies - Depression Scale

The CES-D (Radloff, 1977) is a 20-item questionnaire about depressive symptomatology. Items are scored 0 to 3, indicating increasing frequency of symptom occurrence over the past week. Acceptable sensitivity and specificity have been reported for a cut-off value of 16 (Beekman et al., 1997; Wada et al., 2007; Whooley, Avins, Miranda & Browner, 1997).

Hospital Anxiety and Depression Scale

The Anxiety subscale of the HADS (Zigmond & Snaith, 1983) is a 7-item questionnaire about symptoms of anxiety. Items are scored on a four-point scale, ranging from 0 (not at all) to 3 (very often indeed). Bjelland et al. (2002) reported Cronbach’s $\alpha$ values ranging from .68 to .93 (mean .83). With a cut-off value of eight, sensitivity and specificity for the HADS-A were approximately .80.

Generalized Anxiety Disorder Scale

The GAD scale (Spitzer, Kroenke, Williams & Lowe, 2006) is a 7-item self-report scale. Items are scored 0 to 3, indicating increasing severity of symptoms over the last two weeks. Kroenke, Spitzer, Williams, Monahan en Löwe (2007) reported Cronbach’s $\alpha$ of .92, sensitivities for several
anxiety disorders ranging from .66 to .89 and specificities ranging from .80 to .82, with a cut-off value of ten.

**Composite International Diagnostic Interview**

To assess the presence of DSM-IV disorders, the Composite International Diagnostic Interview version 2.1 (CIDI; World Health Organization, 1997) was administered by telephone. CIDI diagnoses on depressive disorders were used as a 'gold standard' for assessment of the accuracy of the curtailed CES-D. Similarly, CIDI diagnoses on anxiety disorders were used for assessment of the accuracy of the curtailed HADS-A, and CIDI diagnoses on generalized anxiety disorder were used for assessment of the accuracy of the curtailed GAD scale.

**1.2.3 Simulation design**

DC and SC were simulated by application of the algorithms on the item score data of all 502 participants. The DC and SC algorithms were implemented in a custom function written in R (R Development Core Team, 2010), following the descriptions of Finkelman et al. (2012). The code used for application of curtailment can be obtained from the first author. SC was simulated with a $\gamma$ value of 0.95 (SC95), and a $\gamma$ value of 0.75 (SC75).

The performance of DC and SC was contrasted with that of an IRT-based computerized classification test (CTT), in which the item order was determined by maximum information at the cut-off value. The CTT algorithm as described in Smits en Finkelman (2013) was used. For training of the CCT, the graded response model (GRM; Samejima, 1979) was used, because all questionnaires consisted of polytomous items. The parameters of the GRM were estimated using marginal ML estimation, as implemented in the R package LTM (Rizopoulos, 2006). To determine the
cut-off value for $\gamma$, corresponding to the cut-off value for the test score, a test characteristic curve was computed using the R package \textit{lordif} (Choi, Gibbons & Crane, 2011).

For application of the CCT, items were selected by using maximum information at the cut-off value of $\theta$ (Equation 4, Thompson, 2009). After administration of every item, an estimate of the respondent’s $\theta$ value, and the corresponding standard error (SE), was obtained. These were used to calculate a confidence interval (CI) of the respondent’s estimated $\theta$ value (Equation 7, Thompson, 2009). When the CI did no longer include the cut-off value for $\theta$, testing was halted and a classification was made for the respondent. The value for the normal deviate corresponding to a 1 - $\epsilon$ confidence interval, was set to yield similar average test lengths for CCT and SC95. This approach allowed for straightforward comparison of both methods, in terms of accuracy (Thompson, 2011).

### 1.2.4 Outcomes

Performance of SC and CTT was assessed by means of leave-one-out cross validation, because using the same data for training and evaluation of an algorithm may result in overly optimistic estimates of performance (Stone, 1974; Hastie, Tibshirani, Friedman, Hastie & Friedman, 2009). For every respondent, algorithms were trained using all data, minus the response pattern of the current respondent. Then, the algorithm was applied to the response pattern of the current respondent. For every respondent, the number of items administered and the final classification decision was recorded.

We evaluated the efficiency of curtailment by calculating means and standard deviations for test length distributions for DC, SC, and CTT. In addition, accuracy of the methods was evaluated by calculating the concordance rate (proportion of respondents with the same classification as the
full-length instrument), sensitivity (proportion of true positives / [proportion of true positives + proportion of false negatives]), and specificity (proportion of true negatives / [proportion of true negatives + proportion of false positives]). Sensitivity and specificity were calculated with respect to DMS-IV diagnoses, as assessed by the CIDI interview in a subgroup of 157 respondents.

1.3 Results

1.3.1 Deterministic curtailment

First, DC was applied to all three questionnaires. This resulted in substantial reductions in average test lengths, with no reductions in accuracy (Tables 1.1 and 1.2). The average test length of the CES-D was reduced from 20 to 12.59 items: a test length reduction of 37.05%, on average. For the shorter scales, test length reductions were somewhat smaller: average test length for the HADS-A scale was reduced by 24.00%, and for the GAD scale by 24.14% (Table 1.1). The test length distributions resulting from the application of DC for “at risk” respondents is shown in Figure 1.1, and for “not at risk” respondents in Figure 1.2. Overall, test length reductions were larger among the “at risk” respondents.

Table 1.1: Summary statistics for test length distributions of original and curtailed versions of several mental health self-report scales.

<table>
<thead>
<tr>
<th>scale</th>
<th>no. of items</th>
<th>DC M</th>
<th>DC SD</th>
<th>SC95 M</th>
<th>SC95 SD</th>
<th>SC75 M</th>
<th>SC75 SD</th>
<th>CTT M</th>
<th>CTT SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CES-D</td>
<td>20</td>
<td>12.59</td>
<td>4.42</td>
<td>12.47</td>
<td>4.31</td>
<td>10.74</td>
<td>4.87</td>
<td>12.50</td>
<td>7.47</td>
</tr>
<tr>
<td>HADS-A</td>
<td>7</td>
<td>5.32</td>
<td>1.33</td>
<td>5.19</td>
<td>1.27</td>
<td>4.62</td>
<td>1.57</td>
<td>5.20</td>
<td>2.15</td>
</tr>
<tr>
<td>GAD</td>
<td>7</td>
<td>5.31</td>
<td>1.13</td>
<td>5.31</td>
<td>1.13</td>
<td>4.49</td>
<td>1.33</td>
<td>5.30</td>
<td>2.05</td>
</tr>
</tbody>
</table>
1 Curtailment of Self-Report Questionnaires

1.3.2 Stochastic curtailment

Application of SC95 resulted in further reductions in average test length, compared to application of DC, without reducing accuracy (Tables 1.1 and 1.2). Application of SC to the CES-D yielded an average test length reduction of 37.65%. Application of SC to the HADS-A resulted in an average test length reduction of 25.86%. The average test length of the GAD scale was reduced by 24.14% (Table 1.1).

Table 1.2: Accuracies for curtailed versions of several mental health self-report scales.

<table>
<thead>
<tr>
<th>scale</th>
<th>DC</th>
<th>SC95</th>
<th>SC75</th>
<th>CCT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>conc</td>
<td>sens</td>
<td>spec</td>
<td>conc</td>
</tr>
<tr>
<td>CES-D</td>
<td>1.000</td>
<td>1.000</td>
<td>0.333</td>
<td>1.000</td>
</tr>
<tr>
<td>HADS-A</td>
<td>1.000</td>
<td>0.819</td>
<td>0.556</td>
<td>1.000</td>
</tr>
<tr>
<td>GAD</td>
<td>1.000</td>
<td>0.867</td>
<td>0.504</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Note. conc = concordance with classification of full length instruments (N=502); sens = sensitivity, calculated with respect to CIDI diagnoses (N=157); spec = specificity, calculated with respect to CIDI diagnoses (N=157).

Application of SC75 again resulted in substantial reductions in average test length, compared to administration of the full-length questionnaires (Table 1.1). Application of SC to the CES-D yielded an average test length reduction from 20 to 10.74 items (46.30%). Application of SC to the HADS-A resulted in an average test length reduction of 34.00% (Table 1.1). The average test length of the GAD scale was reduced by 35.86%, to 4.49 items. Notably, no reductions in diagnostic accuracies were found for SC75, with only one exception: concordance with the full length instrument showed a small decrease for the GAD scale, from 1.000 to .998 (Table 1.2). For both SC95 and SC75, test length reductions were larger among “at risk” respondents (Figure 1.1), than among “not at risk” respondents (Figure 1.2).
Figure 1.1: Test length distributions for “at risk” (full-length test score equal to, or above, the cut-off value) respondents. From left to right, test lengths for the CES-D, HADS-A and GAD scales, resulting from application of four test-length reduction methods, are depicted.

### 1.3.3 Comparison of SC and CCT

The CCT, set to yield comparable test length to SC95 (Table 1.1), provided lower accordance with the classification of the full length instrument than SC95 and SC75 (Table 1.2). Similarly, the overall diagnostic accuracy with respect to CIDI diagnoses was higher for SC than for CCT, for the CES-D and HADS scales. For the GAD scale, however, CCT provided slightly higher sensitivity than SC (Table 1.2). Test length distributions resulting from application of CCT showed a pattern similar to that of DC and SC: test length reductions were higher for “at risk” (Figure 1.1), than for “not at risk” respondents (Figure 1.2).
1 Curtailment of Self-Report Questionnaires

Figure 1.2: Test length distributions for “not at risk” (i.e., full-length test score below cut-off value) respondents. From left to right, test lengths for the CES-D, HADS-A and GAD scales, resulting from application of four test-length reduction methods, are depicted.

1.4 Discussion

The results of our study indicate that DC provides substantial reductions in test length, while diagnostic accuracy remained identical to that of the full-length test. Furthermore, application of SC resulted in further reductions of test length, with identical diagnostic accuracy for SC95, and only a very slight change in accuracy for SC75, compared to the classification of the full-length test. Test length reductions of 24 to 37% were obtained by applying DC, whereas SC resulted in average test length reductions of 24 to 46%, compared to administration of the full length instrument. Compared to an IRT-based CCT with average test length comparable to SC95, the curtailment algorithms provided better diagnostic accuracy. Therefore, curtailment provides a powerful method for improving the efficiency
of self-report questionnaires used in screening for common mental health disorders.

Compared to adaptive methods for test length reduction, an advantage of curtailment is that it involves only minimal assumptions about the distribution of the data, in contrast to for example IRT-based methods (e.g., Van der Linden & Glas, 2010). Another advantage of curtailment is that the original item order can be retained. As a result, unforeseen item order effects (e.g., McFarland, 1981), which may hamper the performance of adaptive testing routines, can be ruled out. At the same time, practical implementation is expected to be less elaborate for curtailment than for adaptive testing routines, because look-up tables can be created, with stopping criteria for every item (see also Finkelman et al., 2012). However, like adaptive testing routines, in curtailment may require tests to be administered by computer, which can be an impediment in some applications. Although in case of oral test administration, curtailment algorithms may be used for creating look-up tables, which provide interviewers with rules for early stopping.

Another advantage of SC is its direct applicability to criterion-keyed testing. When the goal of classification is an external criterion, this external criterion may be incorporated in the training of SC directly. For example, for curtailment of a questionnaire used for diagnosis of depressive disorder, the “at risk” training dataset would then be composed of the item scores of respondents with a diagnosis of depressive disorder. Similarly, the “not at risk” training dataset would then be composed of item scores of respondents without a diagnosis of depressive disorder. In the current study, we have not taken this approach, because mental disorder diagnoses were available for only a subsample of the respondents.

In contrast to fixed-length short forms, curtailment provides the advantage of allowing the user to control the certainty of classification decisions. With fixed-length short forms, the certainty of the resulting classification decision may differ across respondents. With DC, testing can be contin-
ued as long as the certainty of the resulting classification decision can be improved, for the current respondent. With SC, the certainty of the resulting classification decision for every respondent can be controlled by setting the value of $\gamma$.

Compared to the study of Finkelman et al. (2012), test length reductions obtained by DC and SC of the CES-D in the current study were about 10% higher. This may be due to differences in test score distributions, as the prevalence of depressive disorder was notably higher in the current study. However, in terms of accuracy, the findings in both studies were the same: the reduction of test length provided by DC and SC did not reduce diagnostic accuracy.

The current study is the first to use a value for $\gamma$ of .75. For example, Finkelman et al. (2012) used values of .95 and .90. Although the results for $\gamma$ of .75 are encouraging, further research on diagnostic accuracy when $\gamma$-values are set lower than .95 is warranted. Based on current research, $\gamma$ values of .95 would be advised for practical application of SC. Or, when large training samples are available, a cross-validation approach may be taken to allow for determination of an optimal value for $\gamma$.

Lastly, it is important to note that application of SC within a population requires calibration of the algorithm in the same population. For example, application of SC in a clinical population requires an existing dataset of item scores from a representative clinical sample. Using an existing dataset of item scores from a non-clinical sample would reduce the accuracy of the resulting classification decisions. Also, clinicians should be aware that the accuracy of curtailed tests only relates to the classification problem at hand. If the assessment goal is measurement, instead of classification, adaptive testing algorithms aimed at measurement should be used.
1.5 Appendix: Curtailment of the HADS-A subscale

Suppose we want to administer a deterministically curtailed version of the HADS-A to two new respondents (respondents A and B). The full-length HADS-A scale has seven items (scored 0-3), and a cut-off value of eight is used for “at risk” classification. After administration of every item, the cumulative score is calculated for the current respondent.

Item responses and cumulative scores of new respondent A:

<table>
<thead>
<tr>
<th>item 1</th>
<th>item 2</th>
<th>item 3</th>
<th>item 4</th>
<th>item 5</th>
<th>item 6</th>
<th>item 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>item score</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>cumulative score</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Item responses and cumulative scores of new respondent B:

<table>
<thead>
<tr>
<th>item 1</th>
<th>item 2</th>
<th>item 3</th>
<th>item 4</th>
<th>item 5</th>
<th>item 6</th>
<th>item 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>item score</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>cumulative score</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
</tbody>
</table>

**Deterministic curtailment**

For new respondent A, testing can be halted after administration of item three, because then the cumulative score (8) already equals the cut-off value. For new respondent B, testing can be halted after administration of item five, because then the cumulative score (1) no longer allows for exceeding the cut-off value with the remaining item scores.

**Stochastic curtailment**

For stochastic curtailment, a training dataset is needed. In addition, we need to set a value for $\gamma$, which we will set to .95 in this example. We have
simulated ten training observations: five response patterns of “at risk” observations, and five response patterns of “not at risk” observations. For every training observation, a rest score (i.e., the sum of the item scores on the remaining items) is calculated for every item.

Item scores (and rest scores between brackets) in the “not at risk” training dataset:

<table>
<thead>
<tr>
<th>Observation</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
<th>Item 6</th>
<th>Item 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 (6)</td>
<td>0 (6)</td>
<td>1 (5)</td>
<td>2 (3)</td>
<td>1 (2)</td>
<td>1 (1)</td>
<td>1 (n/a)</td>
</tr>
<tr>
<td>2</td>
<td>2 (5)</td>
<td>0 (5)</td>
<td>1 (4)</td>
<td>2 (2)</td>
<td>1 (1)</td>
<td>1 (0)</td>
<td>0 (n/a)</td>
</tr>
<tr>
<td>3</td>
<td>1 (5)</td>
<td>0 (5)</td>
<td>1 (4)</td>
<td>2 (2)</td>
<td>1 (1)</td>
<td>1 (0)</td>
<td>0 (n/a)</td>
</tr>
<tr>
<td>4</td>
<td>1 (5)</td>
<td>1 (4)</td>
<td>1 (3)</td>
<td>2 (1)</td>
<td>0 (1)</td>
<td>1 (0)</td>
<td>0 (n/a)</td>
</tr>
<tr>
<td>5</td>
<td>1 (6)</td>
<td>2 (4)</td>
<td>1 (3)</td>
<td>2 (1)</td>
<td>0 (1)</td>
<td>1 (0)</td>
<td>0 (n/a)</td>
</tr>
</tbody>
</table>

Item scores (and rest scores between brackets) in the “at risk” training dataset:

<table>
<thead>
<tr>
<th>Observation</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
<th>Item 6</th>
<th>Item 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2 (9)</td>
<td>1 (8)</td>
<td>1 (7)</td>
<td>3 (4)</td>
<td>0 (4)</td>
<td>2 (2)</td>
<td>2 (n/a)</td>
</tr>
<tr>
<td>7</td>
<td>0 (8)</td>
<td>1 (7)</td>
<td>2 (5)</td>
<td>2 (3)</td>
<td>1 (2)</td>
<td>1 (1)</td>
<td>1 (n/a)</td>
</tr>
<tr>
<td>8</td>
<td>2 (7)</td>
<td>2 (5)</td>
<td>1 (4)</td>
<td>3 (1)</td>
<td>0 (1)</td>
<td>1 (0)</td>
<td>0 (n/a)</td>
</tr>
<tr>
<td>9</td>
<td>2 (8)</td>
<td>1 (7)</td>
<td>2 (5)</td>
<td>2 (3)</td>
<td>1 (2)</td>
<td>2 (0)</td>
<td>0 (n/a)</td>
</tr>
<tr>
<td>10</td>
<td>3 (5)</td>
<td>1 (4)</td>
<td>1 (3)</td>
<td>0 (3)</td>
<td>0 (3)</td>
<td>3 (0)</td>
<td>0 (n/a)</td>
</tr>
</tbody>
</table>

For administration of the test to new respondents, the rest scores are used to calculate two empirical proportions, after administration of every item: \( \hat{P}^+ \) and \( \hat{P}^- \). \( \hat{P}^+ \) is the proportion of rest scores in the “at risk” training dataset, which, when added to the new respondent’s cumulative score, will result in a test score equal to, or exceeding, the cut-off value. Similarly, \( \hat{P}^- \) is the proportion of rest scores in the “not at risk” training dataset, which, when added to the new respondent’s cumulative score, will result in a test score equal to, or exceeding the cut-off value.

For new respondent A, the score on the first item was 3. Addition of the rest scores for item 1, yields a test score equal to, or exceeding 8, in
Empirical proportions for new respondent A:

<table>
<thead>
<tr>
<th>item 1</th>
<th>item 2</th>
<th>item 3</th>
<th>item 4</th>
<th>item 5</th>
<th>item 6</th>
<th>item 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P^-$</td>
<td>1.0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$P^+$</td>
<td>1.0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

five out of five “at risk” training observations. Therefore, $\hat{P}^+ = 1.0$ after administration of item 1. Similarly, addition of the rest scores for item 1, yields a test score equal to, or exceeding 8, in five out of five “not at risk” training observations. Therefore, $\hat{P}^- = 1.0$ after administration of item 1. Because $\hat{P}^- > \gamma$ and $\hat{P}^+ > \gamma$, testing is halted after administration of the first item, and an “at risk” classification is made for respondent A.

Empirical proportions for new respondent B:

<table>
<thead>
<tr>
<th>item 1</th>
<th>item 2</th>
<th>item 3</th>
<th>item 4</th>
<th>item 5</th>
<th>item 6</th>
<th>item 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P^-$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$P^+$</td>
<td>0.8</td>
<td>0.6</td>
<td>0.2</td>
<td>0.0</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

For new respondent B, the score on the first item was 1. Addition of the rest scores for item 1 to this item score, yields a test score equal to, or exceeding 8, in four out of five “at risk” training observations. Therefore, $\hat{P}^+ = 0.8$ after administration of item 1. Similarly, addition of the rest scores for item 1 to this item score, yields a test score equal to, or exceeding 8, in zero out of five “not at risk” training observations. Therefore, $\hat{P}^- = 0.0$ after administration of item 1. Because $\hat{P}^- < (1 - \gamma)$, but $\hat{P}^+ > (1 - \gamma)$, testing continues. In a similar manner, testing continues after administration of item 2 and item 3. After administration of the fourth item, $\hat{P}^- < (1 - \gamma)$ and $\hat{P}^+ < (1 - \gamma)$. Therefore, testing is halted, and a “not at risk” classification is made for respondent B, after administration of item 4.