Psychiatry Research 215 (2014) 477–482

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Curtailment: A method to reduce the length of self-report questionnaires while maintaining diagnostic accuracy

Marjolein Fokkema [a,*], Niels Smits [a], Matthew D. Finkelman [b], Henk Kelderman [a], Pim Cuijpers [a]

[a] Vrije Universiteit Amsterdam, the Netherlands
[b] Tufts University School of Dental Medicine, USA

ARTICLE INFO

Article history:
Received 12 April 2013
Received in revised form
24 October 2013
Accepted 4 November 2013
Available online 12 November 2013

Keywords:
Curtailment
Stochastic curtailment
Respondent burden
Efficiency
Common mental health disorders
Screening

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 (DC) and Stochastic (SC) Curtailment 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 in 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. Introduction

As noted by for example Gilbody et al. (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 and 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 et al., 2011; Cuijpers et al., 2010; Rost et al., 1993). However, for fixed-length short forms, the reduction in test length generally comes at the expense of diagnostic accuracy (Smith et al., 2000; Mitchell and 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 and 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 et al., 2009; Gibbons et al., 2008; Smits et al., 2011; Walter et al., 2007). However, most adaptive testing algorithms are based on Item Response Theory (IRT), and assume the data to satisfy 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 et al., 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 and Finkelman, 2013).

Recently, Finkelman et al. (2011) and Finkelman et al. (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 and Simons (1978) and Eisenberg and Ghosh (1980), and curtailment has been used for early stopping in clinical trials (e.g., Lan et al., 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,

* Correspondence to: Vrije Universiteit Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, the Netherlands. Tel.: +31 6 248 92 741.
E-mail address: m.fokkema@vu.nl (M. Fokkema).

0165-1781/$ - see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.psychres.2013.11.003

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 and Snaith, 1983) to screen for anxiety disorders, by using a cutoff score of eight (Bjelland et al., 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 option (an item score of three) on the first two items, it seems likely that their final test score will exceed the cutoff, and therefore further item administration may not be necessary. On the other hand, for respondents endorsing the least severe response option (an item score of zero) on the first two items, it seems likely that their final test score will not exceed the cutoff, 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 cutoff 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 cutoff value for classifying respondents as "at risk" is needed.[1] 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 cutoff value. Item administration is halted and a "not at risk" classification is made, when the remaining items can no longer result in a final test score below the cutoff value.

For application of Stochastic Curtailment (SC), a cutoff 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 γ: 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 cutoff value; the "not at risk" dataset contains item scores of all respondents with a test score below the cutoff 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 cutoff 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 cutoff 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).

[1] Note that the curtailment algorithms described here make two implicit assumptions: First, those respondents scoring above, or equal to, the cutoff value are classified as "at risk", and respondents scoring below the cutoff value are classified as "not at risk". Second, that all response options have values ≥ 0.

An illustration of DC and SC of administration of the HADS-A scale to two new respondents is provided in Appendix A. Although the curtailment algorithms can 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 Section 2, the dataset, algorithms and simulation design will be described. In Section 3, the findings in terms of test length reduction and accuracy will be presented. In Section 4, implications of the current study and directions for further research will be presented.

2. Method

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 et al., 2009, 2011). The total sample consisted of 502 participants, with a mean age of 43 (S.D. = 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.

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 past week. Acceptable sensitivity and specificity have been reported for a cutoff value of 16 (Beekman et al., 1997; Wada et al., 2007; Whooley et al., 1997).

Hospital anxiety and depression scale. The Anxiety subscale of the HADS (Zigmond and 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 α values ranging from 0.68 to 0.93 (mean 0.83). With a cutoff score of eight, sensitivity and specificity for the HADS-A were approximately 0.80.

Generalized anxiety disorder scale. The GAD scale (Spitzer et al., 2006) is a 7-item self-report scale. Items are scored 0–3, indicating increasing severity of symptoms over the last 2 weeks. Kroenke et al. (2007) reported Cronbach's α of 0.92, sensitivities for several anxiety disorders ranging from 0.66 to 0.89 and specificities ranging from 0.80 to 0.82, with a cutoff value of ten.

Composite international diagnostic interview. To assess the presence of DSM-IV disorders, the CIDI version 2.1 (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.

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