Psychiatry Research 188 (2011) 147–155

Contents lists available at ScienceDirect

Psychiatry Research

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

Applying computerized adaptive testing to the CES-D scale: A simulation study

Niels Smits *, Pim Cuijpers, Annemieke van Straten

Department of Clinical Psychology, Faculty of Psychology and Education, Vrije Universiteit, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 27 August 2009
Received in revised form 15 November 2010
Accepted 1 December 2010

Keywords:
Self assessment
Questionnaires
Psychometrics
Depressive disorder
Adolescent psychiatry

ABSTRACT

In this paper we studied the appropriateness of developing an adaptive version of the Center of Epidemiological Studies-Depression (CES-D, Radloff, 1977) scale. Computerized Adaptive Testing (CAT) involves the computerized administration of a test in which each item is dynamically selected from a pool of items until a pre-specified measurement precision is reached. Two types of analyses were performed using the CES-D responses of a large sample of adolescents ($N = 1392$). First, it was shown that the items met the psychometric requirements needed for CAT. Second, CATs were simulated by using the existing item responses as if they had been collected adaptively. CATs selecting only a small number of items gave results which, in terms of depression measurement and criterion validity, were only marginally different from the results of full CES-D assessment. It was concluded that CAT is a very fruitful way of improving the efficiency of the CES-D questionnaire. The discussion addresses the strengths and limitations of the application of CAT in mental health research.

1. Introduction

In the clinical field there is a high demand for mental health assessments which have both a short duration and good quality (e.g., Gardner et al., 2004; Cella et al., 2007; Smits et al., 2007). A methodology that offers substantial promise in this regard is Computerized Adaptive Testing (CAT). CAT involves the administration of a test or questionnaire via the computer. Each item is dynamically selected from a pool of items and is optimal for the responder in question. CAT relies on modern test theory, which is also known as Item Response Theory (IRT). It assumes that the responses to the items of a questionnaire are accounted for by a latent variable and characteristics of the items. IRT models have item parameters which quantify the relationship between the latent trait and the item score. In a CAT, after a response is provided by the responder, the CAT algorithm uses IRT to estimate the responder's provisional latent construct score, and selects a new item from the total set that is most informative for this estimate. (A more extensive description of CAT will be given in the next section.)

Initially, CAT was designed for cognitive testing (e.g., Wainer, 2000). More recently, various CAT procedures for attitude and personality assessment have been developed (see, e.g., Reise and Henson, 2000; Hol et al., 2001, 2005). Moreover, in the last decade, CAT has received a lot of attention in the field of quality of life research. For example, the Patient Reported Outcomes Measurement Information System (PROMIS, www.NIHpromis.org, Cella et al., 2007) project has as its goal the development of CATs for the measurement of physical and mental outcomes which allow for monitoring the health-related quality of life of medical patients. CATs have now been developed for depression (Fliege et al., 2005; Forkmann et al., 2009) and anxiety (Walter et al., 2007). By contrast, in the field of mental health, CATs are hardly, if ever, used (e.g., Gardner et al., 2004). For example, a CAT version of the Center for Epidemiologic Studies-Depression (CES-D) scale (Radloff, 1977), which is one of the most used depression screeners in the mental health field has not been developed yet. An adaptive version of the CES-D could potentially improve the efficiency of depression measurement, both in clinical and research settings.

This article has the following goals: (a) to assess whether the items of the CES-D meet the psychometric requirements needed for adaptive testing, (b) to study whether an adaptive version of the CES-D would yield inferences that are similar to those based on the full CES-D, and (c) to introduce IRT, adaptive testing, and the requirements for CAT to an audience which is unfamiliar with the topic. To that end we use the data of a sample of Dutch adolescents who filled out the full CES-D on the Internet. These data were used to (a) canvas the psychometric properties of the CES-D, and (b) as input for a CAT simulation: for each respondent, the actual responses of the full administration were used as input for a CAT algorithm. We first provide a short introduction to IRT and CAT for readers unfamiliar with adaptive testing.

2. Item response theory and computerized adaptive testing

2.1. IRT: the graded response model

IRT provides a much more powerful measurement framework for testing than does Classical Test Theory (CTT) (e.g., Edelen and Reeve, 2007). In contrast to CTT, IRT does not model the total score, but the

* Corresponding author. Department of Clinical Psychology, Faculty of Psychology and Education, Vrije Universiteit, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands.
E-mail address: n.smits@psy.vu.nl (N. Smits).

0165-1781/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.psychres.2010.12.001

pattern of item responses. This allows for a quantification of the quality of a single item. Consequently, IRT does, and CTT does not allow for the selection of items that are most appropriate for a given test taker, which is an important building block of CAT.

It is instructive to start a discussion of IRT with the two parameter logistic model (2PL) for cognitive ability tests with correct/false (dichotomous) outcomes. Typically, the 2PL employs a logit transformation of the linear equation: $w = a(\theta - b)$ to model the probability of a correct answer on the item. (The logit transformation of quantity w brings it on a probability, or 0 to 1, scale.) In this equation θ represents the subject's value on the latent trait scale. Commonly it is assumed that the distribution of θ over the subjects follows a standard normal distribution. Parameter a represents the extent to which the item discriminates between different ability levels. It may also be interpreted as the strength of association between the item and the construct being measured. The b represents the item threshold, i.e., the value on the latent trait scale above which a correct answer is expected (i.e., the probability of a correct answer is higher than of a false answer). The b parameters are often called 'difficulty' parameters, but when modeling mental health instruments they can better be thought of as 'difficulty to endorse'. Consider two CES-D items: (5) 'I had trouble keeping my mind on what I was doing' and (17) 'I had crying spells' that are to be answered by either yes or no. The second item would be more difficult to endorse because it presents a more extreme situation demanding a higher position on the latent depression variable to give an affirmative answer. Thus, it would have the higher estimated item difficulty.

The common version of the CES-D does not use a scale with two (yes/no) but with four categories (less than 1 day, 1–2 days, 3–4 days, 5–7 days, scored with 0, 1, 2, and 3, respectively). Therefore, IRT models for polytomous instead of dichotomous responses should be used. There are several IRT models for ordered polytomous items, such as the Graded Response Model (GRM, Samejima, 1969), and the Partial Credit Model (Muraki, 1992). Although these models will yield nearly identical estimates of the person parameters, there are at least two reasons to prefer the GRM. First, GRM has parameters which can be interpreted in terms of the responder behavior, i.e., filling out questionnaire items with a Likert rating scale, whereas others do not (Van Engelenburg, 1997; also see Mellenbergh, 1995). Second, GRM is easier to understand and illustrate to users than the other models (Reeve et al., 2007). The GRM is a generalized version of the 2PL. The 2PL can be interpreted as modelling the probability of 'stepping' from the lower ('no') to the higher item category ('yes'). Likewise, the GRM describes the probabilities of stepping from a lower category to higher categories; whereas the 2PL models one step, the GRM has a number of steps that is equal to the number of item categories minus one. For each of the steps, the GRM model employs a logit transformation of the linear equation $w_j = a(\theta - b_j)$. Again, a is the item discrimination parameter (which is identical for all steps within a single item) and b_j represents the threshold parameter of step j. The set of threshold parameters gives the boundaries on the latent variable scale above which one is expected to step from the lower to a higher category (i.e., for which this probability is higher than 50%). The order of the difficulties conforms to the order of the item categories: the value of the threshold between category 0 and 1 lies below the threshold between category 1 and 2, etcetera. Within this model, the item score equals the number of steps completed and is interpreted as a graded score. If a given step is completed, all steps which are less difficult are completed too. Alternatively, if a step is failed, all steps which are more difficult are failed too (Van Engelenburg, 1997). Once the discrimination and threshold parameters are estimated, these values can be used to obtain so-called Category Response Curves (CRCs), which describe the probability of choosing each response category as a function of the latent trait score (e.g., Embretson and Reise, 2000, chap. 5).

The estimated GRM parameters of the CES-D data used in the current study (details will be given in the sections that follow) are shown in Table 1; the category response curves for items 5 and 17 are displayed in Fig. 1. The discrimination parameters indicate that item 5 ($a = 1.35$) has a somewhat lower ability to demarcate fine gradations among persons with similar levels of depression than item 17 ($a = 1.84$). This also becomes apparent in the category response curves: for item 17, the curves are somewhat steeper (for the highest and lowest category) and more narrow and peaked (for the middle categories) than for item 5. In addition, the curves of item 5 suggest that subjects with a latent trait value lower than −0.87 have the highest probability of choosing category 0; subjects with values between −0.87 and 0.27 are more likely to choose category 1; subjects with values between 0.27 and 1.73 are more likely choose category 2, and subjects with values of 1.73 and above have the greatest likelihood of choosing category 3. In addition, when comparing the category response curves of the two items, it can be seen that item 5 is generally more easily endorsed because its curves are located more to the left. This becomes even more apparent when focusing, for example, on subjects with a latent depression score half a standard deviation above

Table 1
Estimated GRM parameters of the items of the CES-D ($N = 1392$).

	Item	Item parameters							
		a	(SE)	b_1	(SE)	b_2	(SE)	b_3	(SE)
1	I was bothered by things that usually don't bother me	1.73	(0.11)	0.07	(0.05)	1.50	(0.09)	3.19	(0.21)
2	I did not feel like eating; my appetite was poor	0.93	(0.08)	0.54	(0.09)	2.14	(0.18)	3.55	(0.30)
3	I felt that I could not shake off the blues	2.26	(0.14)	0.73	(0.05)	1.55	(0.08)	2.42	(0.12)
4	I felt that I was just as good as other people	0.99	(0.08)	−0.37	(0.09)	0.97	(0.12)	2.02	(0.19)
5	I had trouble keeping my mind on what I was doing	1.35	(0.09)	−1.04	(0.09)	0.38	(0.07)	1.95	(0.13)
6	I felt depressed	2.63	(0.15)	0.38	(0.04)	1.34	(0.07)	2.23	(0.10)
7	I felt that everything I did was an effort	1.49	(0.10)	−0.02	(0.06)	1.32	(0.09)	2.53	(0.16)
8	I felt hopeful about the future	1.19	(0.09)	−0.68	(0.08)	0.96	(0.10)	2.18	(0.17)
9	I thought my life had been a failure	2.40	(0.16)	0.72	(0.05)	1.52	(0.07)	2.23	(0.11)
10	I felt fearful	1.51	(0.11)	0.73	(0.07)	2.08	(0.13)	3.28	(0.23)
11	My sleep was restless	1.16	(0.09)	−0.05	(0.07)	1.36	(0.11)	2.60	(0.19)
12	I was happy	2.04	(0.11)	−0.22	(0.05)	1.02	(0.07)	2.04	(0.11)
13	I talked less than usual	1.31	(0.09)	0.10	(0.06)	1.65	(0.11)	2.96	(0.20)
14	I felt lonely	2.40	(0.14)	0.22	(0.04)	1.20	(0.06)	2.07	(0.10)
15	People were unfriendly	1.26	(0.09)	0.36	(0.07)	2.13	(0.14)	3.75	(0.29)
16	I enjoyed life	1.97	(0.11)	−0.14	(0.05)	1.16	(0.07)	2.10	(0.12)
17	I had crying spells	1.84	(0.13)	0.78	(0.06)	1.66	(0.09)	2.58	(0.15)
18	I felt sad	2.90	(0.16)	0.12	(0.04)	1.24	(0.06)	2.18	(0.10)
19	I felt that people disliked me	1.76	(0.11)	0.23	(0.05)	1.41	(0.09)	2.77	(0.16)
20	I could not get going	1.52	(0.10)	−0.03	(0.06)	1.29	(0.09)	2.50	(0.16)

Note. a is the discrimination parameter; the bs are location parameters; SE is standard error of the parameter estimate.