Item Selection in Polytomous CAT

Bernard P. Veldkamp*

Department of Educational Measurement and Data-Analysis, University of Twente, P.O.Box 217, 7500 AE Enschede, The Netherlands

Summary. In polytomous CAT items can be selected using Fisher Information, Maximum Interval Information, or Kullback-Leibler Information. In this paper, the different item selection criteria are described. In a simulation study the criteria are compared for a number of item pools. The process of deciding which one to choose is illustrated. Application of the three item selection criteria to polytomous CAT, constrained polytomous CAT, and CAT based on multi-peaked item pools is discussed.

Keywords. Fisher Information, Fisher Interval Information, Item selection, Kullback-Leibler Information, Polytomous CAT

1 Introduction

Computerized adaptive testing (CAT) is one of de major developments in educational measurement of the past decade. CAT stands in the long tradition of individualized testing. It can be compared with an oral exam where a computer program acts as the examiner. Like in oral exams, the difficulty of the items is adapted to the ability of the candidate. So, the examinees do not get bored or frustrated due to items that are too easy or too hard. Besides, the increased flexibility of CATs enables test developers to fullfil many of the examinee's wishes; for example shorter tests, and testing on demand.

Most research on CAT deals with dichotomously scored items. Only a few studies deal with polytomous CAT. One of the results from these studies is that polytomous CAT tends to need fewer items, since the items are more informative. But still, quite a number of questions need further attention.

One of the issues in polytomous CAT is the choose of item selection criterion. Fisher Information is commonly used, but this information measure is based on an estimate of the ability parameter, and the ability estimate is not very stable in the beginning of a CAT administration. Therefore, when the estimate is not close to the true value, using Fisher's Information criterion might result in inefficient item selection. The fact that item selection procedures may favor items with optimal properties at wrong ability values is generally known as the attenuation paradox in test theory (Lord and Novick, 1968, sect. 16.5). To overcome these problems some alternative criteria have been presented in the literature. In this paper the fo-

* Veldkamp, B.P. (2003). Item Selection in Polytomous CAT. In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.). *New Developments in Psychometrics* (pp.207-214). Tokyo, Japan: Springer Verlag.

cus is on selecting the most appropriate item selection criterion for polytomous CAT.

2 Item Selection Criteria

In the literature several item selection criteria have been proposed. In this section three of the most commonly used item selection criteria for polytomous CAT will be described.

2.1 Maximum Fisher Information

Item selection in polytomous CAT is mainly based on Fisher Information (Dodd, De Ayala, and Koch, 1995). For a single item, Fisher's Information function is defined by:

$$I_{ik}(\theta) = a_i^2 \left[\sum_{k=1}^{m} k^2 P_{ik}(\theta) - \left(\sum_{k=1}^{m} k P_{ik}(\theta) \right)^2 \right], \tag{1}$$

where m is the number of categories and $P_{ik}(\theta)$ is the probability that a candidate with ability θ will end up in category *k* of item *i*. When Fisher Information is used, the item is selected with maximum value of the information function at the estimated ability level of the examinee ($i = \arg\max_i I_{ik}(\hat{\theta})$).

2.2 Maximum Interval Information

Veerkamp and Berger (1997) introduced an interval information criterion for dichotomous CAT to overcome the problems of Fisher Information. Instead of maximizing Fisher's information function at an ability estimate, they proposed to integrate the function over a small interval around the estimate to compensate for the uncertanty in it.

In polytomous CAT there is an other reason to intergrate Fisher's Information function over an interval. Fisher's Information function might be multi-peaked, when items are calibrated with the GPCM (Muraki, 1993). In van Rijn, Eggen, Hemker, and Sanders (in press), it is demonstrated that a multipeaked item might contain more information for a small interval around the ability estimate than the item that contains maximum Fisher Information at the ability estimate.They propose to select the next item with a Maximum Interval Information criterion:

$$i = \arg\max_i \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} I_i(\theta) d\theta, \tag{2}$$

where i is the item to be selected and δ is a small constant defining the width of the interval.

Veldkamp, B.P. (2003). Item Selection in Polytomous CAT. In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.). *New Developments in Psychometrics* (pp.207-214). Tokyo, Japan: Springer Verlag.

2.3 Posterior Expected Kullback-Leibler Information.

The observation that item selection based on Fisher's Information criterion might be inefficient during the first stages of a CAT was also made by Chang and Ying (1996). They propose to select items based on global information rather than on local information. The global criterion they propose is based on Kullback-Leibler Information. Generally, Kullback-Leibler Information measures the distance between two likelihoods over the same parameter space (Lehmann and Casella, 1998, sect. 1.7). The purpose of CAT is to estimate the true ability of the examinee. For this purpose it is desirable to select items generating response vectors with a likelihood at the true ability differing maximally from those at any other value of the ability parameter

More precisely, Kullback-Leibler Information for a single item is defined as

$$K_i(\theta,\theta_0) \equiv \sum_{k=1}^{m} P_{ik}(\theta_0) \ln\left(\frac{P_{ik}(\theta_0)}{P_{ik}(\theta)}\right). \tag{3}$$

For an entire test of *n* items, the measure is equal to the sum of the information of the items. Because the true ability θ_0 of the examinee, is unknown and θ is unspecified, posterior expected information of θ (van der Linden, 1998) will be used. Actual item selection will be based on posterior expected Kullback-Leibler Information at the current ability estimate. Let $f(\theta \mid u_{i_1},...,u_{i_{k-1}})$ be the posterior density of θ after (k-1) items are administered and $\hat{\theta}^{k-1}$ the (EAP) estimate derived from this posterior. Posterior expected Kullback-Leibler Information in the response on the *i*th item in the pool after (k-1) items in the test is defined as

$$KL_i(\hat{\theta}^{k-1}) \equiv \int_{\theta} K_i(\theta,\hat{\theta}^{k-1}) f(\theta \mid u_{i_1},...,u_{i_{k-1}}) d\theta. \tag{4}$$

For application of posterior expected Kullback-Leibler Information to the problem of item selection see also Veldkamp and van der Linden (in press), where an application to the problem of assembling tests that measure multiple abilities is described.

3 How to Select a Criterion

In the previous section, three item selection criteria for polytomous CAT were introduced. The question remains, which one to choose. Due to the attenuation paradox, Maximum Interval Information and Kullback-Leibler Information seem to be preferable, at least during the early stage of a CAT. For dichotomous CAT, this statement was confirmed by the results of Chang and Ying (1996), and Berger and Veerkamp (1994). On the other hand, in van Rijn et al. (in press), no real differences in performance between Fisher Information and Maximum Interval Information were found for the polytomous case. A second argument exists that might favor Fisher Information over the others in the long run. In a frequentist framework, the asymptotic variance of the ability estimate is the reciprocal of the test information function (Lehmann, 1983, p. 465). This means that Fisher Information

Veldkamp, B.P. (2003). Item Selection in Polytomous CAT. In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.). *New Developments in Psychometrics* (pp.207-214). Tokyo, Japan: Springer Verlag.

is proportional to the error of measurement. A third, rather pragmatic argument is that Fisher Information is easier to calculate.

Based on these arguments, it is hard to make a conclusion. During the early stages of an adaptive test, Maximum Interval Information and Kullback-Leibler information are supposed to perform better than Fisher Information. But what are these early stages. Can some general recommendation be made? And how dependent are these recommendations on the characteristics of an item pool?

In general, polytomously scored items contain more information than dichotomously scored items. As a consequense, the ability estimation will be more precise in polytomous CAT. This might explain why van Rijn et al. did not find any differences in performance between Fisher Information and Maximum Interval Information for the polytomous case, where Berger and Veerkamp did find differences for the dichotomous case. The question remains, is the difference small enough to ignore?

A first step in deciding which criteria to choose, might be to find out how much overlap in items occurs for a test. If overlap is high (above 90 percent), almost the same items have been selected by the different criteria, and not much difference in performance is expected. Looking at item overlap to select a criterion is not new. In Sympson, Weiss and Ree (see Weiss, 1982, p. 478) Fisher Information and Owen's selection criterion were compared. In a real life application, they found an overlap of approximately 85 percent of the items. If overlap is low (below 75 percent) a second step might be to check whether Maximum Interval Information and Kullback-Leibler information outperform Fisher Information. If this is not the case, no gain in performance is expected by using Maximum Interval Information or Kullback-Leibler information, and Fisher Information seems a good choice. When on of them outperforms Fisher Information, it should be applied in polytomous CAT.

4 Numerical Examples

To illustrate the process of choosing an item selection criterion, a simulation study was carried out. The effects of item pool size and the amount of information in the items were taken into account. For a number of item pools that differed in size and in average item information, the percentage of overlapping items were recorded. In order to carry out the simulation study, an IRT model should be specified first.

4.1 IRT model

In this paper the focus is on the General Partial Credit model (GPCM) (Muraki, 1992). In the GPCM the probability of obtaining a score in category *k* above the adjacent category *k-1* is given by

$$P_{k|k,k-1} = \frac{\exp a(\theta - b_k)}{1 + \exp a(\theta - b_k)} \tag{5}$$

Veldkamp, B.P. (2003). Item Selection in Polytomous CAT. In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.). *New Developments in Psychometrics* (pp.207-214). Tokyo, Japan: Springer Verlag.

where a is the slope parameter, b_k is an item category parameter, k={0,1,2,..,m} is a category number, and θ represents the ability of the examinee. Rewriting this equation results in the probability of obtaining a score in category k:

$$P_k = \frac{\exp \sum_{v=0}^{k} a(\theta - b_v)}{\sum_{c=0}^{m} \exp \sum_{v=0}^{c} a(\theta - b_v)}. \quad (6)$$

4.2 Simulation study

Nine GPCM item pools were simulated to compare the different item selection criteria. The item pools differed in size (90, 180 and 360 items), and the item pools differed in value of the slope parameters (a in [0.4,1.0], a in [0.4,1.7], or a in [1.4,1.7]). Three items pools consist of easy items, three of medium items, and three of difficult items items (average category parameter in [-2,-1], in [-1,1], or in [1,2]).

To test the effect of size and average item information, the same study was carried out for all nine item pools. For several values of the ability parameter 100 examinees were simulated.

In an ordinary CAT procedure, the next item is selected by a criterion based on the estimated ability level of the examinees, and the selected item is presented to the examinee. However, because item overlap is between tests was measured, a slightly different approach was used. After every iteration of the CAT, all three item selection criteria were applied to propose the next item based on the estimated ability level. The proposed items were denoted in a file, and one of these items was presented to the examinee. In this way, we could make sure that the selection criteria were compared for identical situations. After doing the simulation study, the proposed items were compared, and the percentages of overlapping items are shown in Figure 1.

In the simulation study, EAP-estimates were used to estimate the ability level, the number of items in the test was equal to 20, and no item exposure control methods were applied.

The results for the different item banks are shown in Figure 1. As can be seen the overlap in items is between 85 and 100 percent. For a twenty-item test this means that on average the number of non-overlapping items is less or equal to three. The result that applying different item selection criteria will only result in three different items suggests that there will not be many differences in measurement precision. However, two general trends can be distinguished in Figure 1. When the number of items in the pool increases, the percentage of overlapping items decreases. When the average discrimination of the item pool increases, the percentage of overlapping items also decreases.

Veldkamp, B.P. (2003). Item Selection in Polytomous CAT. In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.). *New Developments in Psychometrics* (pp.207-214). Tokyo, Japan: Springer Verlag.

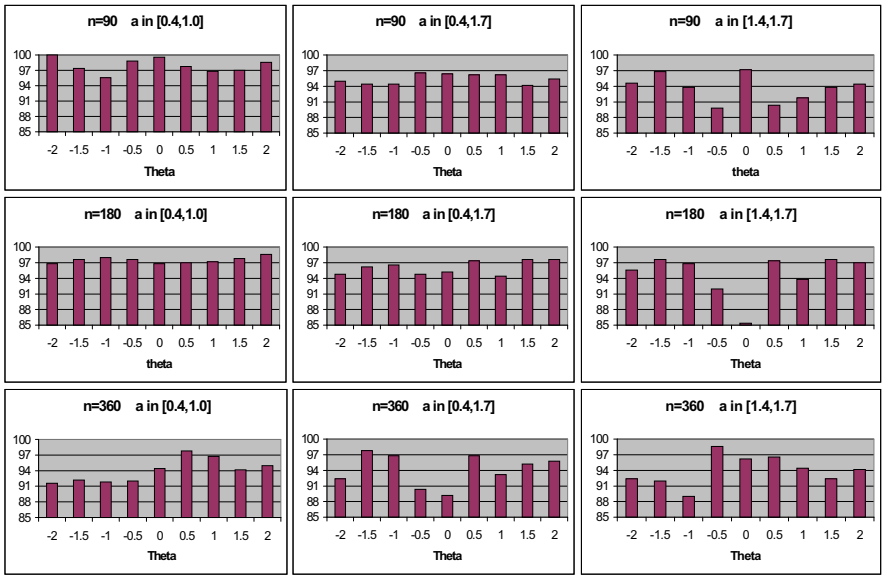

Fig. 1. For nine simulated item banks, the percentages of item overlap are shown for different theta values when different item selection criteria were used.

To check what the effect is of the different items that were choosen by the different criteria on the measurement precision, the resulting MSEs should be compared. In Figure 2, the MSEs of test assembled for the different item selection criteria are shown. Even for the item pool with minimum overlap (n=360, high slope parameters), only smal differences in measurement precision are found. After twenty items, the difference in MSE is smaller that 0.01.

5 Discussion

Several item selection criteria were discussed. Fisher Information, Maximum Interval Information, and Kullback-Leibler Information have been applied to dichotomous CAT before, and their pro's and cons have been discussed in the literature. But until now, no general recommendations about selecting one of these criteria could be given for polytomous CAT. In this study, the influences of several factors on the performance of the criteria were further investigated.

First of all, the performance of the criteria might be influenced by the quality of the items in the bank. In other words, differences between item banks might result in differences in performances. These differences can be caused by differences in the number of items in the bank, or by differences in item parameters. In a simulation study both of these factors were investigated. From the results, it can be concluded that when te number of items in an item bank increases, the differences

Veldkamp, B.P. (2003). Item Selection in Polytomous CAT. In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.). *New Developments in Psychometrics* (pp.207-214). Tokyo, Japan: Springer Verlag.

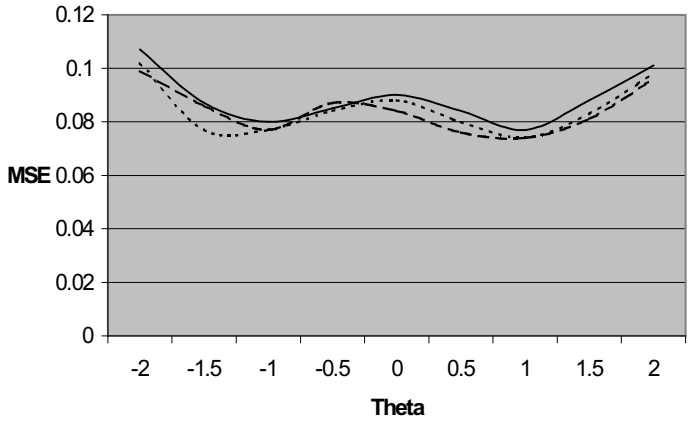

Fig. 2. MSE for polytomous CAT of 20 items when Fisher Information (line), Fisher Interval Information (dashed), or Kullback-Leibler Information (dotted) was applied.

also increase. It can also be concluded that differences in performance are higher for item banks with highly discriminating items. On the other hand, even for the largest differences in performance in this simulation study, the differences in MSE's were still small.

Besides, other factors might cause differences in performance. For example, the number of items in the CAT that is assembled might play a role. Fisher Information selects the item with maximum Fisher Information at the ability estimate. However, during the first few stages of a CAT, the ability estimate is not very stable in the beginning of a CAT. This might cause problems. When the estimate is not close to the true value of the ability parameter, using Fisher Information will result in inefficient item selection. In the examples above, the number of items was equal to twenty. When the number of items is smaller, for example smaller than ten, the weaknesses of Fisher Information might be demonstrated and the other criteria might outperform Fisher Information.

A second topic for further research is the use of item pools where the information of the items is multi-peaked. In van Rijn et al. (in press), it was demonstrated that Fisher Information criterion can be outperformed by Maximum Interval information criterion when Fisher's Information function is multi-peaked. The nine simulated item pools in the example above consist of single-peaked items. So, more research is needed to check for this property.

A third topic for further research is decribed in Dodd, DeAyala and Koch (1995). They indicate that item selection procedures for polytomous CAT have not been studied under conditions in which it is necessary to ensure content balancing of the items presented during the CAT. Besides content balancing, it might be necessary to introduce other test specifications in practical testing situations. For dichotomous CAT the Weighted Deviation Model (Stocking and Swanson, 1993) and the Shadow Test Approach (van der Linden and Reese, 1998) have been developed to deal with such constraints. These methods can be modified and applied

Veldkamp, B.P. (2003). Item Selection in Polytomous CAT. In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.). *New Developments in Psychometrics* (pp.207-214). Tokyo, Japan: Springer Verlag.

to polytomous CAT. The effects of imposing constraints on the different item selection criteria are unknown. In general, imposing constraints reduces the number of available items, that can be selected during the next iteration of a CAT. Because imposing constraints reduces the number of available items, the percentage of overlapping items between the different criteria will probably increase.

Finally, in this study, number of items in the bank, and discriminating power of the items only caused only small differences in performance. Based on these results, no recommendations for selecting a criterion could be made. Further research on other factors is needed to reveal the strength and the weaknesses of the three item selection criteria.

6 References

Chang H-H, Ying Z (1996) A global information approach to computerized adaptive testing. Applied Psychological Measurement 20: 213-229

Dodd BG, De Ayala RJ, Koch WR (1995) Computerized adaptive testing with polytomous items. Applied Psychological Measurement 19: 5-22

Lehmann EL, Casella G (1998) Theory of point estimation. Springer-Verlag, New York

Lord FM, Novick MR (1968) Statistical theories of mental test scores. Addison-Wesley, Reading, MA

Muraki E (1992) A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement 19: 159-176

Stocking ML, Swanson L (1993) A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement 17: 277-292

van der Linden WJ (1998) Bayesian item selection criteria for adaptive testing. Psychometrika 63: 201-216

van der Linden WJ, Reese LM (1998) A model for optimal constrained adaptive testing. Applied Psychological Measurement 22: 259-270

van Rijn PW, Eggen TJHM, Hemker BT, Sanders PF (in press) A selection procedure for polytomous items in computerized adaptive testing. Applied Psychological Measurement

Veerkamp WJJ, Berger MPF (1997) Some new item selection criteria for adaptive testing. Journal of Educational and Behavioral Statistics 22: 203-226

Veldkamp BP, van der Linden WJ (in press) Multidimensional adaptive testing with constraints on test content. Psychometrika

Weiss DJ (1982) Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement 4: 473-485

Veldkamp, B.P. (2003). Item Selection in Polytomous CAT. In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.). *New Developments in Psychometrics* (pp.207-214). Tokyo, Japan: Springer Verlag.