Eurasian Journal of Educational Research 69 (2017) 21-36

Eurasian Journal of Educational Research
www.ejer.com.tr

Effects of Content Balancing and Item Selection Method on Ability Estimation in Computerized Adaptive Tests

Alper SAHIN[1] Durmus OZBASI[2]

ARTICLE INFO

Article History:

Received: 14 January 2017

Received in revised form: 09 April 2017

Accepted: 23 April 2017

DOI: http://dx.doi.org/10.14689/ejer.2017.69.2

Keywords
likelihood weighted information
fisher's maximum information
Estimation accuracy

ABSTRACT

Purpose: This study aims to reveal effects of content balancing and item selection method on ability estimation in computerized adaptive tests by comparing Fisher's maximum information (FMI) and likelihood weighted information (LWI) methods. **Research Methods:** Four groups of examinees (250, 500, 750, 1000) and a bank of 500 items with 10 different content domains were generated through Monte Carlo simulations. Examinee ability was estimated by fixing all settings except for the item selection methods mentioned. True and estimated ability (θ) values were compared by dividing examinees into six subgroups. Moreover, the average number of items used was compared. **Findings:** The correlations decreased steadily as examinee θ level increased among all examinee groups when LWI was used. FMI had the same trend with the 250 and 500 examinees. Correlations for 750 examinees decreased as θ level increased as well, but they were somewhat steady with FMI. For 1000 examinees, FMI was not successful in estimating examinee θ accurately after θ subgroup 4. Moreover, when FMI was used, θ estimates had less error than LWI. The figures regarding the average items used indicated that LWI used fewer items in subgroups 1, 2, 3 and that FMI used less items in subgroups 4, 5, and 6. **Implications for Research and Practice:** The findings indicated that when content balancing is put into use, LWI is more suitable to estimate examinee θ for examinees between -3 and 0 and that FMI is more stable when examinee θ is above 0. An item selection algorithm combining these two item selection methods is recommended.

[1] Corresponding Author: Middle East Technical University Northern Cyprus Campus, School of Foreign Languages, Modern Languages Program, alpersahin2@yahoo.com

[2] Canakkale Onsekiz Mart University, TURKEY, dozbasi@comu.edu.tr

Introduction

Traditional paper and pencil tests are on the verge of being outdated due to recent technological advances that affect measurement and evaluation field. In the traditional paper and pencil tests, test takers take all items in a test and spend a considerable amount of time responding to items that are too easy or too difficult for them. Thanks to recent technology and advancements in educational measurement, test takers no longer have to take all items in a test. Rather, they only take the items aligned to their estimated ability (θ) level that is calculated while they are taking the test. This is possible with computerized adaptive tests (CATs). Typically, CATs have some advantages over traditional methods such as providing the test results immediately, reducing the number of items taken by each examinee dramatically, and being more reliable and valid than a conventional test while using fewer items (Hambleton & Swaminathan, 1985; Rudner, 1998; Weissman, 2006; Thompson & Weiss, 2011).

Having an estimate of examinee θ with less error highly depends on putting some sound criteria for item selection, test termination, and θ estimation together in the CAT environment. Item selection method is a very important component of CATs (Choi & Swartz, 2009), as the θ estimation in a CAT environment is conducted in real time according to the responses of the test takers to certain items with known item parameters. Therefore, ensuring that the computer makes the right decision in choosing which item to use next has the utmost influence on θ estimates, which are used for many high-stakes purposes. However, the selection of the appropriate item in the item pool is not an easy process in CATs. It has still been discussed in the literature (Chang & Ying, 1996; Veerkamp & Berger, 1997; van der Linden, 1998; Chen, Ankenmann & Chang, 2000; Cheng & Liou, 2003; Weissmann, 2006).

A successful CAT is based on an item bank composed of items that address a wide range of θ levels. This item bank has its own information function to which each item contributes with its own information function formed according to its item parameters. During a CAT session, items are mainly selected among the ones with the highest information and closest to the location of the estimated θ of the examinee taking the test. As expected, some item selection methods have been proposed by different authors (Kingsbury & Zara, 1989; Lord, 1980; Veerkamp & Berger, 1997; Chang & Ying, 1996) in order to optimize this procedure. However, selection of items in CAT is often dependent on Fisher's maximum information (FMI). FMI mostly uses the maximum likelihood estimate of the θ (Veerkamp & Berger, 1997; Barrada, Olea, Ponsoda & Abad, 2010).

FMI utilizes item information, the conversion of the item characteristic curve, to select items for CATs (Weiss, 1983). Selecting items from an item pool for a multiple-choice test, where the item characteristic curve is defined in three-parameter logistic model (3PLM; will be explained further later), FMI can be calculated using Equation 1 (Embretson and Reise, 2000):

$$I_i[\theta_{m-1}] = \frac{(D_{ai})^2(1-c_i)}{\left[c_i+e^{D_{ai}(\theta_{m-1}-b_i)}\right]\left[1+e^{D_{ai}(\theta_{m-1}-b_i)}\right]^2}, \quad (1)$$

in which,

m = examinee

a_i = item discrimination for item i;

b_i = item difficulty for item i;

c_i = pseudo-chance parameter of item i;

D = scaling constant (mostly used as 1.7)

and in which, c_i is set to 0.00 for two parameter model and a_i to 1.00 (and c_i to 0.00 as well) for one parameter model. The item information for each item in the item bank can be calculated with the formula above. With the help of equation 1, the total item information levels of the items given to one person reaches the maximum (Lord, 1980).

In studies on item selection, FMI or an FMI-based method almost never changes as the performances of newly proposed methods are mostly compared to that of FMI. Although many studies were conducted to develop better alternative item selection methods, their results could yield slight differences or advantages over FMI. According to the current literature, especially when the CAT has more than 20 items, the difference in performance of a newly proposed method and FMI turns out to be trivial (Passos, Berger & Tan, 2007). For example, Chen, Ankenmann and Chang (2000) conducted a simulation study to compare item selection methods of FMI, Fisher interval information, Fisher information with a posterior distribution, Kullback Leibler information (KLI) and KLI with a posterior distribution in terms of test efficiency and ability estimation precision at the beginning of CAT session. In their results, they found that for CATs with more than 10 items, there is no difference between FMI and other selection methods in terms of θ estimation precision. Similarly, Chang and Ying (1996) compared the performance of KLI and FMI in two studies. In the first, they used an item bank of 800 items simulated from a pre-specified uniform distribution, and in the second one they used an item bank of 254 items whose parameters were taken from a National Assessment of Educational Progress reading test. They found that KLI performed slightly better when the test was short. Especially in the second study, the difference was trivial.

Additional studies have reached similar results with negligible differences between FMI and alternative methods for tests with more than 20 items (Barrada, Olea, Ponsoda & Abad, 2009; van Rijn, Eggen, Hemker & Sanders, 2002; Veldkamp, 2003). However, Veerkamp and Berger (1997) suggested a feasible alternative item selection criteria called likelihood weighted information (LWI). In LWI, which was suggested by Veerkamp and Berger (1997) as an alternative to FMI, the information function is formulated as a weighted mean of information function of all possible theta values. The LWI function is defined by Veerkamp and Berger (1997) as:

$$\max_{i \in In} \int_{-\infty}^{\infty} L_n(\theta\,; x_{n-1}) I_i(\theta) d\theta. \tag{2}$$

in which they define LWI as a product of $L_n(\theta; x_{n-1})$, the likelihood function (I) of the (n-1)th item with a response vector of x_{n-1}.

In their study, Veerkamp and Berger (1997) used two simulated item banks with 200 and 400 items generated in 3PLM. They compared FMI, interval information and LWI for up to 60-item tests. They found that LWI was a good alternative to the FMI. LWI was found to be the only alternative that outperformed FMI in tests over 20 at that time.

There is ample research on the comparison of item selection methods in CATs. However, the current literature lacks further studies considering the recent advances and practical needs of current CAT applications like content balancing. There was no study found in the literature that compared the performance of item selection methods when content balancing was put into use. Moreover, the current literature does not reveal how the examinees with different ability levels are affected from the changes in item selection method and content balancing. The present study addressed these issues by using FMI and LWI as the item selection methods together with content balancing in CAT and sought an answer to the following research question: Does the accuracy of the θ estimation change for examinees with different θ levels depending on the item selection method used when content balancing is put into use?

Method

Research Design

According to the International Council for Science (2004), basic research is defined as experiment- or theory-based research that aims to increase the current information on a topic with indirect concerns about its practicality. The present study is a basic research study, the data of which was generated through Monte-Carlo simulations using SimulCAT (Han, 2012).

Research Sample

As the first step of the item generation process, examinee samples of different sizes (250, 500, 750 and 1000) were generated with a standard normal distribution between -3 and +3. In this way, the true θ levels of these examinees were obtained.

Research Instruments and Procedures

After the generation of the examinee samples, the items in the item bank of the study were generated. For this purpose, a bank of 500 items with equally distributed items in 10 different content domains (with 50 items in each) were generated separately in 3PLM of item response theory (IRT). In 3PLM, each item has item discrimination (a), item difficulty (b) and pseudo-chance (c) parameters. The 3PLM can be shown with equation 3 (Hambleton, Swaminathan & Rogers, 1991):

$$P_{ij}(\theta_j) = c_i + (1-c_i)\frac{\exp\left[\mathrm{D}a_i(\theta_j - b_i)\right]}{1+\exp\left[\mathrm{D}a_i(\theta_j - b_i)\right]}, i = 1..., n \tag{3}$$

in which $P_{ij}(\theta_j)$ can be explained as the probability of a correct response of examinee j to item i on a specific θ level. Moreover, a_i corresponds to estimated a, b_i to estimated b, and c_i to the estimated c parameter for item i.

All item parameters were generated from a uniform distribution with the a parameters ranging between 0 and 1.5, the b parameters ranging between -3 and +3 and the c parameters ranging between 0 and 0.25. The item parameters were generated from a uniform distribution in order to obtain an item bank with more balanced capability of estimating θ in all areas of the θ continuum. The item bank information function of the item bank generated can be viewed in Figure 1.

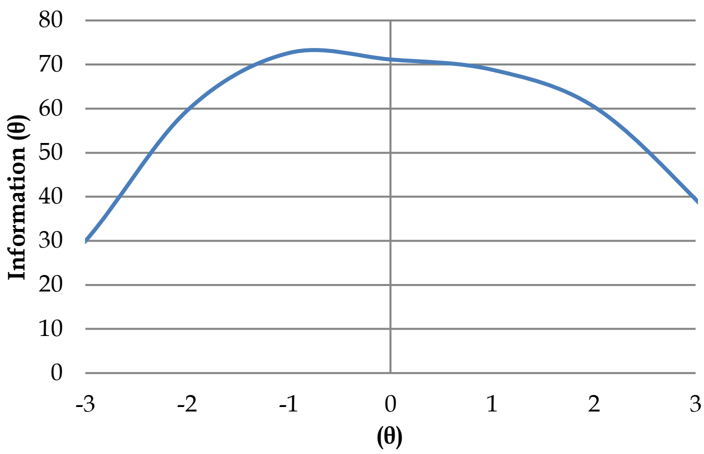

Figure 1. Item bank information function.

Post-Hoc Simulations. Following the generation of examinee and item parameters, five post-hoc simulations were conducted. During these post-hoc simulations, each exam session was set to have at least 10 items and 10% from each content domain. This was done to make sure that the sessions did not terminate with very few items and that there are approximately the same number of items from each content domain in each session. Maximum likelihood estimation was used to estimate examinee abilities in each research condition. Tests were terminated when the standard error of θ estimate was 0.25 and below. No exposure control method was utilized. Moreover, random values between -0.5 and 0.5 were taken as the initial θ estimates of the examinees.

As mentioned earlier, performance of two item selection methods, LWI and FMI methods were compared. This comparison was done with each of four examinee samples, and each research condition was replicated 10 times. In this way, 21 individual θ (including true θ) for each examinee in each examinee sample and a total of 84 scores were obtained. A brief overview of this can be seen in Table 1.

Table 1

A Brief Overview of Scores Obtained Through Simulations

	True ability score for each examinee	LWI (estimated score for each examinee with replications)	FMI (estimated score for each examinee with replications)	Total
250	1	10	10	21
500	1	10	10	21
750	1	10	10	21
1000	1	10	10	21
			Total	84

Data Analysis

Data analysis was handled by investigating the accuracy of θ estimates, conditional on ability subgroups.

The accuracy of θ estimates in each research condition was evaluated by calculating the correlation (r; Gao & Chen, 2005) between the true θ levels of the examinees that were obtained when the examinees were first generated and their estimated θ levels in each research condition and replication. Then, these correlations were averaged to obtain the average correlation of the estimated θ scores for each examinee. Moreover, the mean squared error (MSE; Veerkamp & Berger, 1997; Chang & Ying, 1996) between the true and estimated scores was also calculated using Equation 4:

$$\mathrm{MSE}\left(\widehat{\theta}\right) = \frac{\sum_{j=1}^{N}(\widehat{\theta}_j - \widehat{\theta}_{Ti})^2}{N}, \tag{4}$$

where $\hat{\theta}_j$ is the estimated θ, $\hat{\theta}_{Ti}$ is the true θ for the examinee j in each research condition, and N is the total number of examinees. Apart from the correlations and MSE values, the average numbers of items used in each research condition were also calculated conditional on examinee samples.

Ability Subgroups. Findings were analyzed conditional on examinees' θ level in pre-specified intervals rather than taking all examinees as a whole. This was done to have a deeper understanding of the effects of item selection on the θ estimation for examinees with various θ levels. It is known that examinees with different θ levels are affected differently from variations in CAT methodology (Sahin & Weiss, 2015).

Examinees were divided into subgroups according to their true θ levels with increments of 1.00 standard deviation. For example, examinees with θ levels higher than -2 were put into subgroup 1. Then, examinees between -2 and -1 were put into subgroup 2. Six subgroups were formed as a result of this procedure. Distribution of examinees in each θ group, conditional on examinee samples, can be seen in Figure 2.

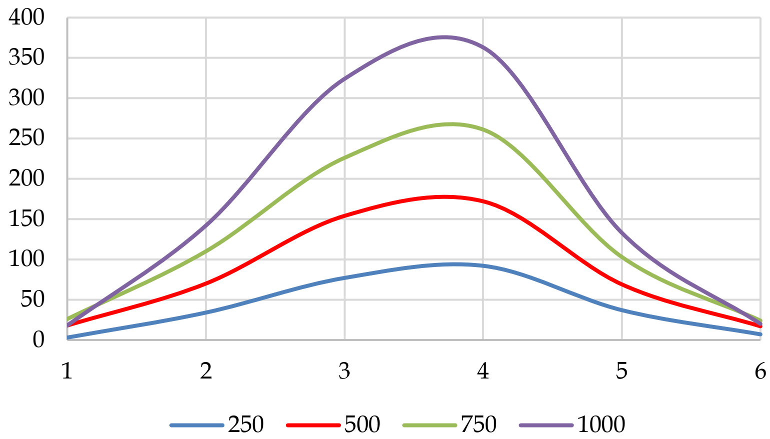

Figure **2.** Distribution of examinees in each θ group conditional on the examinee samples.

Results

The average correlation coefficients between true and estimated θ parameters, conditional on the number of test takers and θ groups, are presented in Figure 3. As can be seen in Figure 3, correlations are the highest in group 1, for the students with the lowest θ levels in all examinee samples. The highest correlation (r=0.94) that was obtained with 250 examinees was in group 1 when FMI was used as the item selection method. The lowest correlation obtained in the same group was r=0.26 when LWI was used in group 6.

The highest correlation obtained with 500 examinees was r=0.75 in group 1, when LWI was used. The lowest correlation obtained with the same examinees was r=0.24 in group 5, when LWI was used. When the examinee number increased to 750, the highest correlation was around the same value in group 1, when LWI (r=0.76) and FMI (r=0.75) were used. In addition, the lowest correlation (r=0.22) was obtained from group 6, when LWI was used. The highest correlation obtained when there were 1000 examinees who took the test was in group 1 again with similar values for FMI (r=0.74) and LWI(r=0.75), and the lowest correlation (r=0.19) was obtained in group 6 when the LWI was used.

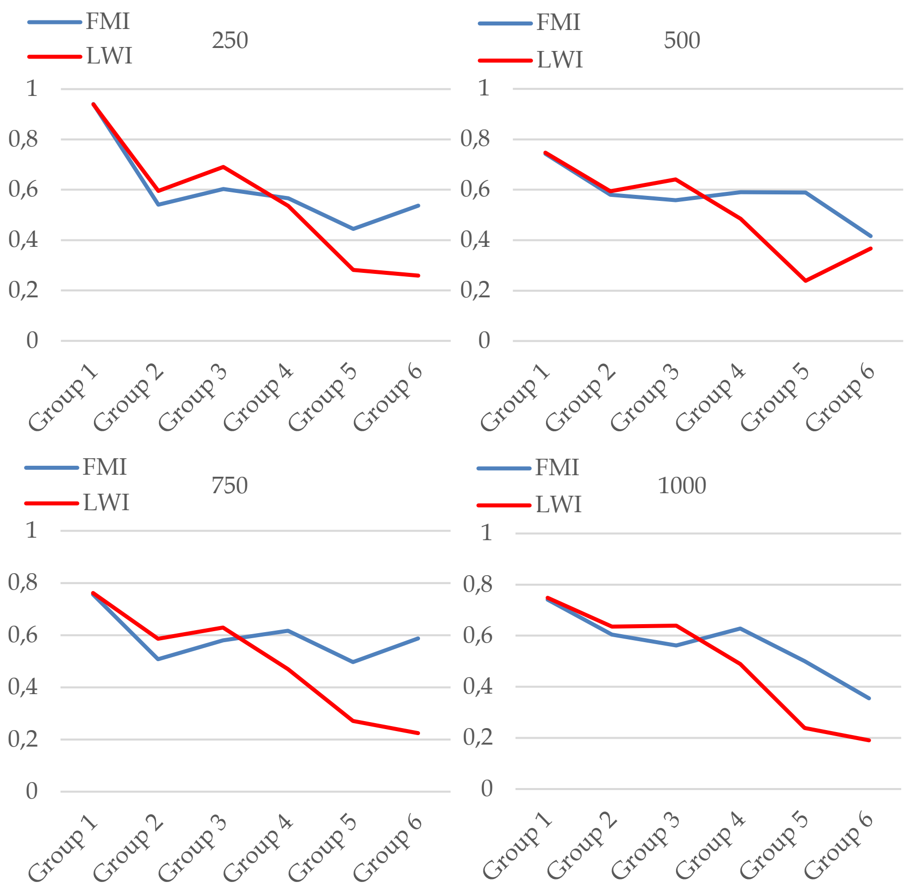

Figure 3. Correlations conditional on number of test takers, item selection method and θ groups.

MSE conditional on number of test takers, item selection method used and θ group of the examinees can be seen in Figure 4. The lowest MSE obtained with 250 examinees was in group 1 (MSE=0.10), when LWI was used. Moreover, MSE=1.11 was the highest MSE value obtained with 250 examinees in group 6, when LWI was used. When the examinee number increased to 500, the lowest MSE was obtained in group 2 (MSE=0.12), when LWI was used. In addition, the highest MSE was obtained in group 6 (MSE=1.22), when LWI was used. In the sample with 750 examinees, the lowest MSE was obtained in group 1 (MSE=0.11), when LWI was used as the item selection method. The highest MSE was in group 6 (MSE=1.35), when LWI was used. In the examinee sample with 1000 examinees, similar results were obtained. The lowest MSE (MSE=0.11) was obtained in group 1 when LWI was used. Group 6 was the one with the highest MSE (MSE=1.27), when LWI was used as the item selection method.

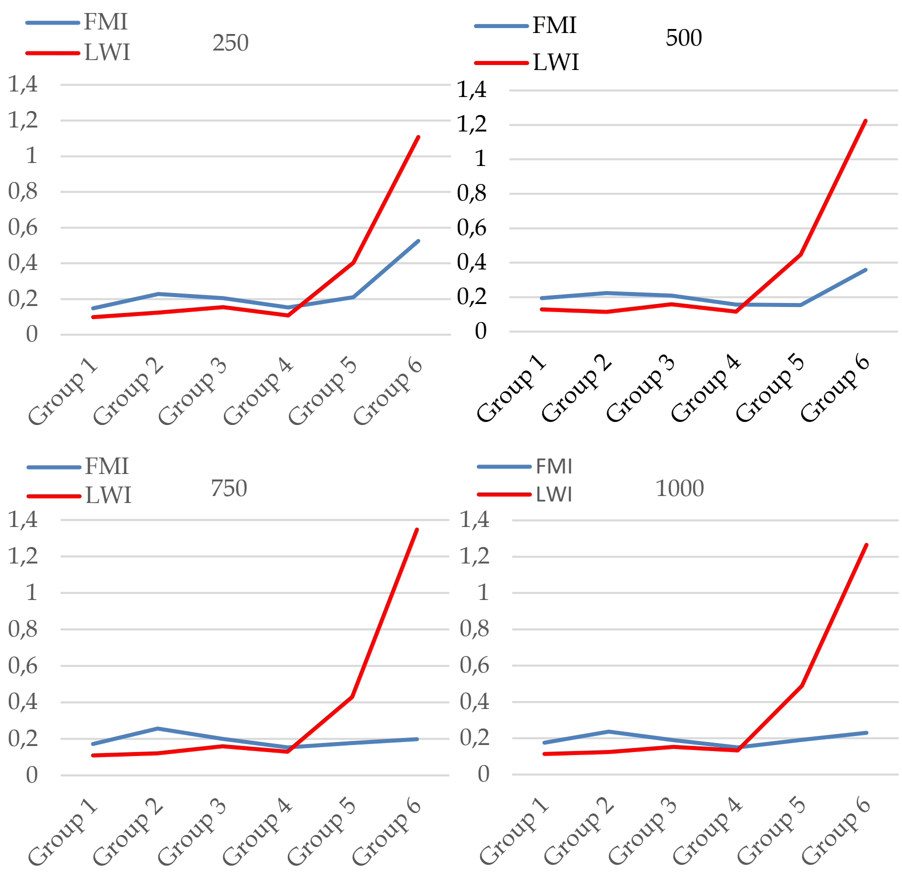

Figure 4. MSE conditional on number of test takers, item selection method and θ groups.

When Figure 5 was analyzed in terms of the average number of items used in each condition of the study, it was seen that 22.56 items were used for group 5, when FMI was used with 250 examinees. The highest average number of items used for the same 250 examinees was 41.77, when LWI was used for examinees in group 6. An average of 31.03 items were used for examinees in group 1 for this examinee sample as well. The highest average number of items used with 500 examinees was 44.55, when LWI was used for examinees in group 6. The lowest average number of items used was 22.78, for group 5 in 500 examinees, when FMI was used. The highest average number of items used for 750 examinees was 45.81, when LWI was used for examinees in group 6. The lowest average number of items used was 22.71 in group 5 of 500 examinees, when FMI was used. Among the 1000 examinees, group 6 got an average 44.1 items, and group 5 got an average of 22.65 items in their sessions.

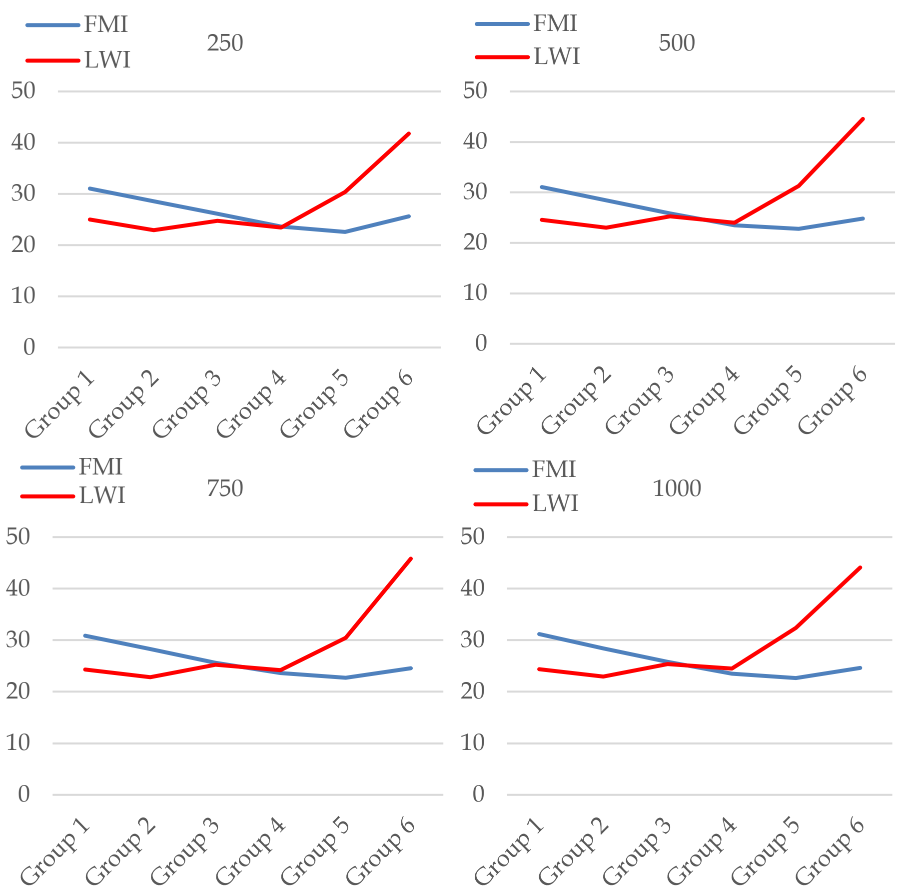

Figure 5. Average number of items used in each condition.

Discussion and Conclusion

The findings regarding the correlations indicated that correlation coefficients decreased steadily as examinee θ level increased from -3 to +3 in all examinee samples when LWI was used as the item selection method. FMI obtained decreasing correlations with 250 and 500 examinees as the examinee level increased. When 750 examinees took the test, correlations were somewhat steady in regard to FMI. When 1000 examinees took the test, FMI was not successful in estimating examinee θ accurately after Group 4. It is interesting to note that LWI is better in estimating the examinee θ levels in θ subgroups 1, 2, and 3. Similarly, FMI outperforms LWI in θ subgroups 4, 5 and 6.

When the figures regarding the MSE are analyzed, parallel conclusions can be drawn. From the figure for MSE, it is visible that there is a dramatic increase in MSE values in subgroup 6 when LWI was used in all conditions. There is also an increase in MSE when FMI was used, but it is somewhat limited compared to LWI. As indicators of estimation accuracy, MSE values indicate that when FMI is used as the item selection method, θ estimates are estimated with less error compared to LWI. Moreover, it is important to note that when the examinee number reached 750, the increase in MSE values when FMI was used became nearly invisible. According to the findings in this regard, as in correlation coefficients, LWI outperforms FMI in θ subgroups 1, 2, 3 and FMI outperforms LWI by having less MSE in θ subgroups 4, 5 and 6. The same rule applies when the average number of items used in all conditions are analyzed.

When all these are put together and interpreted as a whole to answer our research question, it can be said that LWI is more suitable to estimate examinee θ for examinees between -3 and 0 when content balancing is put into use. Moreover, our results also suggest that FMI is more stable when examinee θ is above 0, but it is less accurate in estimating examinee θ when the examinee level is below 0. This is somewhat conflicting with Veerkamp and Berger (1997), who found that LWI might be a sound alternative to FMI. LWI may be a good alternative to FMI when θ estimates are compared as a whole and when content balancing is not put into use; however, when the content balancing is in use and when examinees are divided into θ groups, LWI outperforms FMI only for certain θ subgroups. Therefore, a new item selection algorithm using the LWI method for the examinees with θ levels below 0 and using FMI for examinees with θ levels above 0 might be more beneficial and more robust against possible difficulties that both of these item selection methods experience for certain groups of examinees during CAT administration.

The current study has some limitations. First of all, the current findings are limited to the uses of LWI and FMI when content was balanced in 10 different content areas that comprised 10% of each CAT session. Secondly, although the data analyses were replicated 10 times, because of the nature of the study, the findings may be rather limited to the data generated in this study. Moreover, the item bank generated in the present study had high information in nearly all areas of the θ continuum, so the results may be limited to CATs with similar item banks. Finally, SE(θ) <= 0.25 was used as the test termination criteria. This may be a rather stringent termination criteria for real CAT administrations, and current findings may be limited as such.

The results of this study have caused some questions to emerge, and it is suggested that they be investigated in further detail by follow-up studies. A possible follow-upstudy would investigate the feasibility of using a mixed-method item selection algorithm, as suggested by the findings of the present study, that uses LWI when the examinee level is below 0 and FMI when it is above 0. Moreover, a similar study with real or simulated data that compares the accuracy of the θ estimates when content balancing is and is not used would also be beneficial. Last but not least, a study comparing the performances of LWI and FMI with item banks of different sizes would be highly valuable.

References

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2009). Item selection rules in computerized adaptive testing: Accuracy and security. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *5*, 7-17.

https://dx.doi.org/10.1027/1614-2241.5.1.7

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement, 34,* 438-452.

https://dx.doi.org/10.1177/0146621610370152

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

http://dx.doi.org/10.1177/014662169602000303

Chen, S.-Y., Ankenmann, R. D., & Chang, H. H.(2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*, 241-255.

http://dx.doi.org/10.1177/01466210022031705

Cheng, E.P. & Liou, M. (2003). Computerized adaptive testing using the nearest-neighbors criterion. *Applied Psychological Measurement*, *27*(3), 204-216.

http://dx.doi.org/10.1177/0146621603027003002

Choi, S. W.& Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement*, *33*, 419-440.

http://dx.doi.org/10.1177/0146621608327801

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three parameter logistic model. *Applied Measurement in Education, 18,* 351–380.

http://dx.doi.org/10.1207/s15324818ame1804_2

Hambleton, R. K., & Swaminathan H. (1985). *Item response theory: Principals and applications*. Norwell, MA: Kluwer Nijhof.

Hambleton, R. K., Swaminathan, H., & Rogers, H.J. (1991) *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Han, K. T. (2012). SimulCAT: Windows software for simulating computerized adaptive test administration. *Applied Psychological Measurement, 36*(1), 64-66.

http://dx.doi.org/10.1177/0146621611414407

International Council for Science. (2004). *The value of basic scientific research.* Retrieved March 10, 2017 from http://www.icsu.org/publications/icsu-position-statements/value-scientific-research

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359–375.

http://dx.doi.org/10.1207/s15324818ame0204_6

Lord, F. M. (1980) *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Passos, V. L., Berger, M. P. F., & Tan, F. E. (2007). Test design optimization in CAT early stage with the nominal response model. *Applied Psychological Measurement*, *31*, 213-232.

http://dx.doi.org/10.1177/ 0146621606291571

Rudner, L. M. (1998). *An on-line, interactive, computer adaptive testing tutorial.* Retrieved February 25, 2017 from

http://echo.edres.org:8080/scripts/cat/catdemo.htm

Sahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences: Theory & Practice, 15,* 1585-1595.

http://dx.doi.org/10.12738/estp.2015.6.0102

Thompson, N. A., & Weiss, D.J. (2011). A Framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16*(1), 1-9. Retrieved from:

http://pareonline.net/getvn.asp?v=16&n=1

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*, 201-216.

http://dx.doi.org/10.1007/BF02294775

van Rijn, P., Eggen, T. J., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26*, 393-411. http://dx.doi.org/10.1177/014662102237796

Veerkamp, W. J .J. & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*(2), 203-226. Retrieved from

http://www.jstor.org/stable/1165378

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.

Weiss, D. J. (1983). *Final Report: Computer-Based Measurement of Intellectual Capabilities*. University of Minnesota, Department of Psychology. Retrieved from the University of Minnesota Digital Conservancy,

http://hdl.handle.net/11299/135261

Weissman, A. (2006). A Feedback control strategy for enhancing item selection efficiency in computerized adaptive testing. *Applied Psychological Measurement*, *30*(2), 87-99.

http://dx.doi.org /10.1177/0146621605282774

Bilgisayar Ortamında Bireye Uyarlanmış Testlerde İçerik Dengeleme ve Madde Seçme Yönteminin Yetenek Düzeyi Kestirimine Etkileri

Atıf :

Sahin, A. & Ozbasi, D. (2017). Effects of content balancing and item selection method on ability estimation in computerized adaptive testing. *Eurasian Journal of Educational Research*, *69*, 21-36. http://dx.doi.org/10.14689/ejer.2017.69.2

Özet

Problem Durumu: Son yıllardaki teknolojik gelişmelerin ölçme ve değerlendirme alanına katkılarıyla birlikte geleneksel anlamda kağıt kalem testleri artık eski popülerliğini yitirmeye başlamıştır. Gelişen bilgisayar teknolojisi, hem ölçme işleminin süresinin kısalmasını hem de daha geçerli ve güvenilir testlerin işekoşulmasını mümkün hale getirmiştir. Özellikle bireyin yetenek düzeyine uygun sınav sorularıyla karşılaşması zaman ve kullanılan süre açısından önemli bir tasarruf sağlamaktadır. Bu, ancak bilgisayar ortamında bireye uyarlanmış test (BOBUT) uygulaması ile mümkün olabilmektedir. BOBUT uygulaması, başlatma kuralı, madde seçim yöntemi, yetenek kestirimi, içerik dengeleme ve test sonlandırma gibi önemli süreçlerden oluşmaktadır. Bu süreçlerin belki de en önemlisi madde seçim yöntemidir. Bu çalışmada BOBUT uygulamasının en önemli aşamalarından olan madde seçim yöntemleri ele alınmıştır. Alanyazındaki madde seçimine yönelik çalışmalar incelendiğinde madde seçim yöntemlerinin içerik dengeleme (content balancing) kullanıldığında farklı yetenek düzeylerindeki bireylerin örtük puanları üzerinde nasıl bir etki gösterdiğinin halihazırda henüz incelenmediği görülmüştür.

Araştırmanın Amacı: Bu araştırmanın amacı BOBUT uygulamalarında içerik dengeleme kullanıldığında madde seçim yöntemindeki değişikliğin yetenek kestirimine etkisini yaygın olarak kullanılan Fisher'ın en yüksek bilgi (Fisher's maximum information) ve onun önemli bir alternatifi olduğu daha önceki

araştırmalarda tespit edilen ağırlıklandırılmış bilgi oranı (Likelihood weighted information) yöntemlerini kullanmak suretiyle belirlemek ve içerik dengeleme üzerine sonraki dönemlerde yapılacak çalışmalara ışık tutmaktır.

Araştırmanın Yöntemi: Araştırmada kullanılan veriler Monte-Carlo simülasyon yöntemi ile elde edilmiştir. Bu bağlamda, araştırmanın verileri için yetenek düzeyleri -3 ile +3 arasında normal dağılım gösteren 4 farklı büyüklükte 250, 500, 750 ve 1000 birey grupları oluşturulmuştur. Yetenek kestirimlerinde en yüksek olabilirlik kestirim (Maximum likelihood estimation) yöntemi kullanılmıştır. Benzetim ile oluşturulan bireyler bu aşamada elde edilen gerçek yetenek düzeylerine göre altı alt yetenek grubuna ayrılmıştır (Örn. $-3 < \theta < -2$ = grup 1, $-2 < \theta < -1$ = grup 2, … vb.).

Madde havuzu için her birine yönelik 50'şer madde bulunan 10 farklı konu alanında toplam 500 madde benzetim yöntemiyle üretilmiştir. Madde parametreleri a parametresi için 0 ile 1.5, b için -3 ile +3 ve c için ise 0 ile 0.25 arasında sabit (uniform) dağılım gösterecek şekilde üretilmiştir. Birey ve maddelerin elde edilmesi sonrası bir dizi Post-hoc benzetim çalışması gerçekleştirilmiştir. Bu çalışmalar, birey yetenek başlangıç düzeyi -0.5 ile +0.5 aralığında olacak, en kısa test uzunluğu her bir konu alanından %10 oranında madde içerecek şekilde en az 10 madde kullanılacak ve yetenek düzeyi kestirimi standart hata değeri 0.25'ten küçük olduğunda testi sonlandıracak şekilde ayarlanmıştır. Post-hoc benzetimler 10 kez tekrarlanmıştır.

Araştırmanın Bulguları: Farklı madde seçme yöntemleri kullanıldığında, gerçek ve kestirilen yetenek düzeyleri arasındaki korelasyonlar (r) 4 farklı büyüklükteki grup ve bu grupların her birinde 6 farklı yetenek aralığındaki bireyler için ayrı ayrı incelenmiştir. Buna göre 250 kişilik grup için Fisher'ın en yüksek bilgi yöntemi kullanıldığında, gerçek ve kestirilen yetenek düzeyleri arasında en yüksek korelasyon r=0.94 olarak bulunmuştur. En düşük korelasyon (r=0.26) ise madde seçme kuralı olarak ağırlıklandırılmış bilgi fonksiyonu kullanıldığında elde edilmiştir. Sınavı alan kişi sayısı 500'e çıktığında ise en yüksek korelasyon madde seçme kuralı olarak ağırlıklandırılmış bilgi oranı kullanıldığında elde edilmiştir (r=0.75). Kişi sayısı 750'ye çıktığında en yüksek korelasyon katsayıları her iki yöntem için de çok yakın bulunmuştur (r_{fisher}=0.75; $r_{ağırlıklandırılmış}$=0.76). Benzer bir durum, örneklem sayısı 1000'e çıktığında da geçerli olmuş ve benzer en yüksek korelasyonlar elde edilmiştir (r_{fisher}=0.74; $r_{ağırlıklandırılmış}$=0.75).

Farklı birey gruplarında her alt yetenek düzeyi için iki madde seçme kuralı ayrı ayrı uygulandığında elde edilen tahmini yetenek düzeyleri ile bireylerin gerçek yetenek düzeyleri arasındaki ortalama karesel hata (MSE; Mean Squared Error) değerleri karşılaştırılmıştır. Buna göre, en düşük MSE değeri 250 kişilik grupta ağırlıklandırılmış bilgi oranı yöntemi kullanıldığında 1. alt yetenek grubunda elde edilmiştir (MSE=0.10). Yine aynı madde seçme kuralında alt yetenek grubu 6'da ise MSE=1.11 ile diğer yetenek gruplarına göre daha yüksek bir değer almıştır. Birey sayısı 500'e çıktığında, ağırlıklandırılmış bilgi oranı yöntemi kullanıldığında alt yetenek grubu 1 MSE=0.12 ile en düşük değer almıştır. En yüksek MSE ise alt grup 6'da MSE=1.22 olarak hesaplanmıştır. Birey sayısı 750'ye çıktığında ise ağırlıklandırılmış bilgi yöntemi kullanıldığında MSE değeri en düşük alt yetenek grubu 1'de (MSE=0.11) elde edilmiştir. En yüksek MSE (1.35) ise yine alt grup 6'da elde edilmiştir. Birey sayısı 1000'e çıktığında da benzer sonuçlar elde edilmiştir. En düşük MSE değeri grup 1'de, en yüksek MSE değeri ise yine grup 6'dan elde edilmiştir.

Her iki madde seçme yönteminin kestirim kalitesi kullanılan ortalama madde sayıları açısından da karşılaştırılmıştır. 250 kişinin sınavı aldığı durumda, en fazla sayıda madde, madde seçme kuralı olarak ağırlıklandırılmış bilgi oranı yöntemi kullanıldığında alt yetenek grubu 6'da ortaya çıkmıştır (kullanılan madde sayısı 41.77). En düşük ortalama madde sayısı (31.03) ise alt yetenek grubu 1'den elde edilmiştir. Sınavı alan birey sayısı 500'e çıktığında ise, en yüksek ortalama madde sayısı madde seçme kuralı olarak ağırlıklandırılmış bilgi yöntemi kullanıldığında grup 6'da elde edilirken, en düşük madde sayısı Fisher'ın en yüksek bilgi yöntemi kullanıldığında 5. alt yetenek grubundan elde edilmiştir (22.78). Bu durum sınavı alan birey sayısı 750 ve 1000 olduğunda da değişmemiş, en yüksek ve en düşük ortalama madde uygulanan yetenek aralıkları ve bunlara ait madde seçme kuralları değişmemiştir. Bir başka ifade ile sınavı alan birey grubu 750 ve 1000 olduğunda en yüksek madde kullanımı her iki birey grubunda da madde seçme kuralı olarak ağırlıklandırılmış bilgi oranı yöntemi kullanıldığında grup 6'da sırasıyla ortalama 45.81 ve 44.1 şeklinde elde edilmiştir. En düşük ortalama madde kullanımı ise madde seçme kuralı olarak Fisher'ın en yüksek bilgi yöntemi kullanıldığında grup 5'te sırasıyla 22.71 ve 22.65 şeklinde elde edilmiştir.

Araştırmanın Sonuçları ve Önerileri: Çalışmada elde edilen tüm bulgular göz önüne alındığında, içerik dengeleme kullanıldığında, ağırlıklandırılmış bilgi oranı yönteminin literatürde geçtiği şekliyle Fisher'ın en yüksek bilgi yöntemine aslında tamamen üstünlük sağlamadığı, bu üstünlüğün yetenek değeri -3 ile 0 aralığında olan bireyler için geçerliği olduğu, yetenek düzeyi 0'ın üzerine çıktığı durumlarda ise Fisher'ın en yüksek bilgi yönteminin yetenek kestiriminde daha başarılı olduğu sonucuna varılmıştır. Bu durum 0'dan küçük yetenek düzeylerinde ağırlıklandırılmış bilgi oranı yönteminin, 0'dan büyük yetenek düzeylerinde Fisher'ın en yüksek bilgi yönteminin kullanılmasını sağlayacak bir madde seçme algoritmasının her iki yöntemin de eksiklerini giderebileceğinden hareketle her durumda BOBUT uygulamalarında daha başarılı yetenek düzeyi kestirimleri elde edilmesini sağlayacak böyle bir algoritmanın geliştirilmesi önerilmektedir.

Anahtar Kelimeler: Ağırlıklandırılmış bilgi oranı, fisher'ın en yüksek bilgi yöntemi, kestirim keskinliği.