ISSN:0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 8, Issue, 8, pp. 19269-19276, August, 2017

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

COMPARING ITEM SELECTION METHODS IN MULTIDIMENSIONAL COMPUTERIZED ADAPTIVE TESTING UNDER DIFFERENT TYPES OF MULTIDIMENSIONAL STRUCTURES

Halil Ibrahim Sari[1]*., Anne Corinne Huggins-Manley[2] and Cengiz Zopluoglu[3]

[1]Measurement and Evaluation in Education Program, Muallim Rifat Faculty of Education, Kilis 7 Aralik University, Kilis, Turkey
[2]Research and Evaluation Methodology Program, School of Human Development and Organizational Studies in Education, College of Education, University of Florida, Gainesville, Florida, USA
[3]Research, Measurement, and Evaluation Program, Department of Educational and Psychological Studies, University of Miami, Miami, Florida, USA

DOI: http://dx.doi.org/10.24327/ijrsr.2017.0808.0659

ARTICLE INFO

Article History:

Received 17th May, 2017
Received in revised form 12th June, 2017
Accepted 04th July, 2017
Published online 28th August, 2017

Key Words:

Multidimensionality, Adaptive Testing, Data Structure, Item Selection

ABSTRACT

The performance of item selection methods in multidimensional computerized adaptive testing has only been studied using an independent cluster multidimensional structure. The goal of this study is to examine the effect of four different item selection methods on test utilization and measurement accuracy under more complex multidimensional data structures. The Kullback-Leibler information method, the minimum angle method, the volume method, and a method that minimizes the error variance of the linear combination were included in the study as item selection methods. We simulated four two-dimensional factor structure conditions: (a) independent cluster, (b) approximate simple, (c) complex, and (d) general factor, while varying the magnitude of the correlation among the dimensions. In general, it was found that the type of data structure played a major role, the magnitude of correlation played a moderate role, and the type of item selection method played a minor role on the research outcomes. The results show the importance of considering more complex data structures in operational MCAT applications.

INTRODUCTION

Comparing Item Selection Methods in Multidimensional Computerized Adaptive Testing under Different Types of Multidimensional Structures Adaptive testing of complex multidimensional constructs is becoming more commonplace and desired with modern advances in technology and measurement theory (van der Linden &Glas, 2000). The simultaneous measurement of multiple related constructs often calls for individual test items that relate to the multiple constructs in varied and complex ways. For example, reading assessments are often composed of items that measure multiple related dimensions such as comprehension, fluency, critical reading, and vocabulary, and each item is related to these multiple dimensions in complex ways (Reckase, 2009). Data gathered from many operational test administrations such as Test of English as a Foreign Language (TOEFL), Law School Admission Test (LSAT), and American College Testing (ACT) often demonstrate more complex structures (Douglas, Kim, Roussos, Stout, & Zhang, 1999; Jang & Roussos, 2007; McDonald, 1999). To reap the benefits of adaptive testing in these situations, using the most appropriate and efficient methods of item selection is necessary to move examinees through a test. Previous research has examined item selection methods in multidimensional adaptive testing (MAT), but always under the assumption of an independent cluster multidimensional latent structure. However, many operational applications will not fit within this framework (Frey & Seitz, 2009; Reckase, 2009). As there has not yet been an attempt to compare the item selection methods under more complex multidimensional structures, little is known about how item selection methods operate under such conditions.

The purpose of the current study is to explore how complex multidimensional factor structures impact item utilization and measurement accuracy across four different item selection methods, and we achieve this by extending the work of Yao (2012, 2013) to more complex multidimensional structures.

Corresponding author:* **Halil Ibrahim Sari
Measurement and Evaluation in Education Program, MuallimRifat Faculty of Education, Kilis 7 Aralik University, Kilis, Turkey.

The following item selection methods were included in the study: (a) Kullback-Leibler information (KL; Veldkamp& van der Linden, 2002), (b) Minimum Angle (Ag; Reckase, 2009), (c) Volume- (V_m; Segall, 1996), and (d) a minimization of the error variance of the linear combination (V; van der Linden, 1999). The previous literature in multidimensional item response theory (MIRT) (e.g., Bolt, 2001; Bolt &Lall, 2003; Yao &Boughton, 2007) and MAT (e.g., Choi & Swartz; 2009; Lee, Ip, &Fuh, 2008) also indicated that the correlation of true abilities across multiple dimensions may affect several outcomes in multidimensional testing (Sass, 2010). Therefore, the current study also explores whether the impact of complex factor structures on item utilization and measurement accuracy depends on the magnitude of correlation among multiple dimensions. The study aims to provide evidence-based suggestions for choosing the most appropriate method of item selection under specific data designs within their MAT studies and applications for researchers and practitioners.

Background

The ultimate goal of any test is to measure examinee skills as accurately and precisely as possible with a set of test items. Historically, this has been approached with traditional methods in which examinees are given a static set of items. That is, the item set is predetermined and does not change during the test. However, static tests have been extensively criticized for providing better precision for students at moderate ability levels rather than students at extreme ability levels (Weiss & Kingsbury, 1984). They are also associated with lengthy tests (Segall, 1996). Putting these two criticisms together, it is asserted that static tests have low measurement efficiency as defined by the ratio of measurement precision and test length (Frey & Seitz, 2009; Segall, 2005). In order to increase the measurement efficiency of a static test, either the test length must be shortened or measurement precision must be increased while holding the other constant, which is very difficult to do in practice. Assuming one is developing a test that is supposed to measure a wide range of abilities, the test must have a wide range of item difficulties (Dorans *et al*, 2000) which will automatically produce longer tests. This means that many items are wasteful for individual examinees because the item difficulty is not well-matched to the individual's ability. In this case, little information is gleaned about the individual's latent ability (Sands, Waters, & McBride, 1997). Furthermore, since many examinees respond to the same set of items, issues surrounding test security (e.g., cheating) often arise in static tests (Linacre, 1988). Lastly, if the static test is administered on paper, then delayed scoring and reporting is inevitable. These are some potential disadvantages of static tests.

Rather than administering a pre-determined set of items to examinees, a computerized adaptive test (CAT) administers different items to different examinees, and each administered item is dependent on the responses to previous items for each individual examinee (van der Linden & Glas, 2000). Therefore, the test is personalized. Although static tests are still dominantly used to measure student ability today, there is a remarkable increase in the use of adaptive tests. While the CAT version of the Armed Services Vocational Aptitude Test Battery (ASVAB) was the only adaptive test before the mid-1980s (van der Linden & Glas, 2000), today many operational CAT testing programs are available for use throughout both statewide (e.g., Student Assessment of Growth and Excellence-Utah's computer adaptive assessment) and nationwide testing programs (e.g., Graduate Management Admission Test-GMAT). This trend is expected to continue with the number of launched adaptive tests increasing at a fast rate (Doranset *al*, 2000).

Unidimensional CAT (UCAT) methodology has been extensively used in the past. In UCAT, test responses are modeled under the assumption that a single underlying trait is the only systematic cause of item responses. However, many educational and psychological traits are defined by several dimensions and the tests are built to capture all of them (Ackerman, Gierls& Walker, 2003). Hence, unidimensionality is considered a strong assumption that is either approximately met or violated to some degree (Reckase, 1979).

Multidimensional computerized adaptive testing (MCAT) (Segall, 1996; van der Linden, 1999; Veldkamp& van der Linden, 2002) allows for the simultaneous adaptive measurement of multiple and related subskills without the assumption of unidimensionality when theory and/or empirical findings do not support it.

MCAT can have great advantages over UCAT when applied to appropriate tests. First, when a test is measuring multiple related dimensions, UCAT often suffers from a confounding effect between content area and item difficulty (Ackerman, Gierl, & Walker, 2003). For instance, in a test of general science an examinee who is good at biology and poor at chemistry may receive easy biology items and difficult chemistry items. Neither set of items will provide useful information about his/her science ability proficiency, andthis situation will ultimately result in low measurement precision (Segall, 1996). MCAT algorithms can easily overcome this problem by specifying separate estimates for each domain. Second, research has found that MCAT yields greater precision of ability estimates and simultaneously reduces test lengths, which increases measurement efficiency (Segall, 1996; van der Linden, 1999). These advantages indicate the need for research that can continue to improve MCAT methodology for continued and increased use in practice.

Although the topic of MCAT is a relatively new area, some work has been done on MCAT including the proposal of many new item selection methods (Reckase, 2009; Segall, 1996; van der Linden, 1999; Wang & Chang, 2011; Yao, 2010), item exposure methods (Finkelman, Nering, & Roussos, 2009; Lee, Ip, &Fuh, 2008), and stopping rules (Wang, Chang, and Boughton, 2012). Furthermore, many comparison studies have been conducted to explore the efficiency of each proposed method under varying conditions including sample size, item pool size, and test length (e.g., He, Diao, & Hauser, 2013; Riley, Dennis, & Conrad, 2010; Yao, 2012, 2013, 2014). Some computer programs such as the R packages "mat" (Choi & King 2015) and SimuMCAT (Yao, 2011b) have been developed for MCAT applications and simulations. The goal of the current study is to contribute to this relatively new literature by extending the work of Yao (2012, 2013) on item selection methods to more complex multidimensional latent structures.

Method

In this simulation study,three simulation factors were manipulated and fully crossed: (a) 4 item selection methods, (b) 4 types of dimensional structure, and (c) 2 different correlations among the latent factors, which resulted in 32 unique conditions. Other conditions such as ability estimation method and stopping rule were fixed across the conditions. The fixed and manipulated conditions are presented in this section.

Manipulated Factors

Type of dimensional structure. Four different multidimensional latent structureswere examined: independent cluster, approximate simple, complex, and general. Figure 1 displays the conceptual model of these four structures. In the independent cluster multidimensional structure, items have large loadings on a single dimension and zero loadings on the other dimension(s). There are many ways in which the independent cluster structure can be violated, and three particular ways are the approximate simple, complex, and general structures in this study. All relate to a situation in which individual items can load onto multiple dimensions, but it is often the case that each item has both a primary and secondary dimension on which it loads. Based on the size of the secondary loadings, one may have an approximate simplestructure or a complexstructure (see Finch, Stage, & Monahan, 2008; Gierl, Leighton & Tan, 2006; Zhang, 1996). Although there are no exact rules of thumb to distinguish the two structure types (Sass, 2010), approximate simple structures are defined by secondary standardized loadings that are smaller in size (e.g., between 0.1 and 0.4 or simply lower than primary loadings) compared to larger primary loadings while complex structures are defined by standardized loadings similar in magnitudes for both primary and secondary dimensions (see Finch, Stage, & Monahan, 2008;Gierl, Leighton & Tan, 2006; Sass, 2010; Zhang, 1996). The general factor structurehas been considered in other research (Jin, 2010; Sass & Schmitt, 2010) and refers to the situation in which the overall instrument measures a single dimension but all items also measure a secondary dimension to a lesser degree. For example, story problems in a mathematics test are predominately measuring math but they can also measure reading comprehension to some degree. It is possible to see examples of this type of structure in many educational and psychological test applications such as the GRE and TOEFL (see Beguin & Glas 2001; Hirsch & Miller, 1991 for empirical examples).

Correlations between the latent dimensions. The correlations between the dimensions were manipulated at two levels as ρ =.10 and ρ=.60 to represent low and moderate positive relationships, respectively.

Item selection rules. The most commonly used item selection methods were compared: Kullback-Leibler information method, the minimum angle method, the volume method, and a method that minimizes the error variance of the linear combination. A method to minimize the error variance of the composite score with the optimized weight (Yao, 2010), a modified version of the volume method, was also investigated. However, the results for this method are not discussed because they were indistinguishable with results of the volume method. A comprehensive explanation of these methods are not presented here as the technical details of them are discussed by Yao (2012, 2013) and Reckase (2009).

Fixed Factors

Stopping rule. The minimum standard error rule (SE; Weiss & Kingsbury, 1984) terminates CAT for a certain domain when the ability for that domain is estimated within desired limits, and it is a commonly used stopping rule in simulation studies. However, when an item bank no longer has informative items for improving upon the ability estimation, the stopping rule is never cued. Therefore, both test length and item exposure rate will be jeopardized. Due to this disadvantage of the minimum SE method, the rule of the predicted standard error reduction (PSER; Choi, Grady, & Dodd, 2011) was used in the current study. Under this method, one defines hyper (α) and hypo (β) parameters. The β parameter indicates a particular standard error of measurement for θ and the α parameter indicates the reduction in the standard error that must be achieved if an additional item is to be administered (Yao, 2014). For this study 0.05 and 0.01 were selected for β and α parameters, respectively. For example, the test continues for an examinee until the estimated theta for a certain domain is associated with a standard error of 0.05 or less, then the PSER is activated to scan the item pool for an item that could minimize the standard error for that domain at least by 0.01. If such an item is not available, item selection for that domain stops and then the process repeats for the next dimension of interest. The test is terminated when the criterion is achieved for all dimensions. One can refer to Choi, Grady, & Dodd (2011) and Yao (2013) for additional details.

Ability estimation. Expected a posterior (EAP) estimation with a prior distribution of N (0,1)was used for ability estimation due to known advantages with respect to measurement precision in MCAT. Readers are referred to Segall (1996) for demonstrations of these advantages.

Item Bank Construction and Data Generation

Four different item banks with 480 items were created from the described four multidimensional latent structures. Two dimensions were considered under each type of structure. We used the item parameters gathered from a 2006 Florida Comprehensive Assessment Test (FCAT) 4th Grade 45-item Reading Test (FDOE, January 2007) taken by 192,480 students, and adopted its discrimination, difficulty and guessing parameters. The item statistics are given in Table 1. Aligned with the FCAT 4th Grade Reading Test (2006), the difficulty and guessing parameters were randomly selected from a uniform distribution with boundaries of [-3.34, .98] and [.07, .42], respectively.

Item discriminations were classified into three groups based on the information in Table 1 so as to be used across the independent cluster, approximate simple and complex latent structures. First, we assume that discriminations between 0.29 and 0.57 are small in size (e.g., first set), discriminations between 0.57 and 0.92 are medium in size (e.g., second set), and the discriminations between 0.92 and 1.39 are large in size (e.g., third set). By following this rule of thumb, the discriminations across the three latent structures were randomly selected and ranged from the given intervals. However, when manipulating the general factor structure, while the third set of discriminations was used for the general factor, the discriminations for the secondary factor ranged from a minimum of .29 to a maximum of 1.39. For demonstration purposes, the discrimination matrices for six items are given for the four types of latent structures in Figure 2, where the given intervals on the first and second columns represent the range of the uniform distributions from which the item discriminations were sampled for the first and second dimensions. Values of 0 indicate discriminations fixed at 0.

Table 1 2006 FCAT 4th Grade Item Summary

	Minimum	25th Percentile	Median	75th Percentile	Maximum
Discrimination (*a*)	0.29	0.57	0.72	0.92	1.39
Difficulty (*b*)	-3.34	-1.21	-0.51	-0.90	0.98
Guessing (*c*)	0.07	0.17	0.21	0.27	0.42

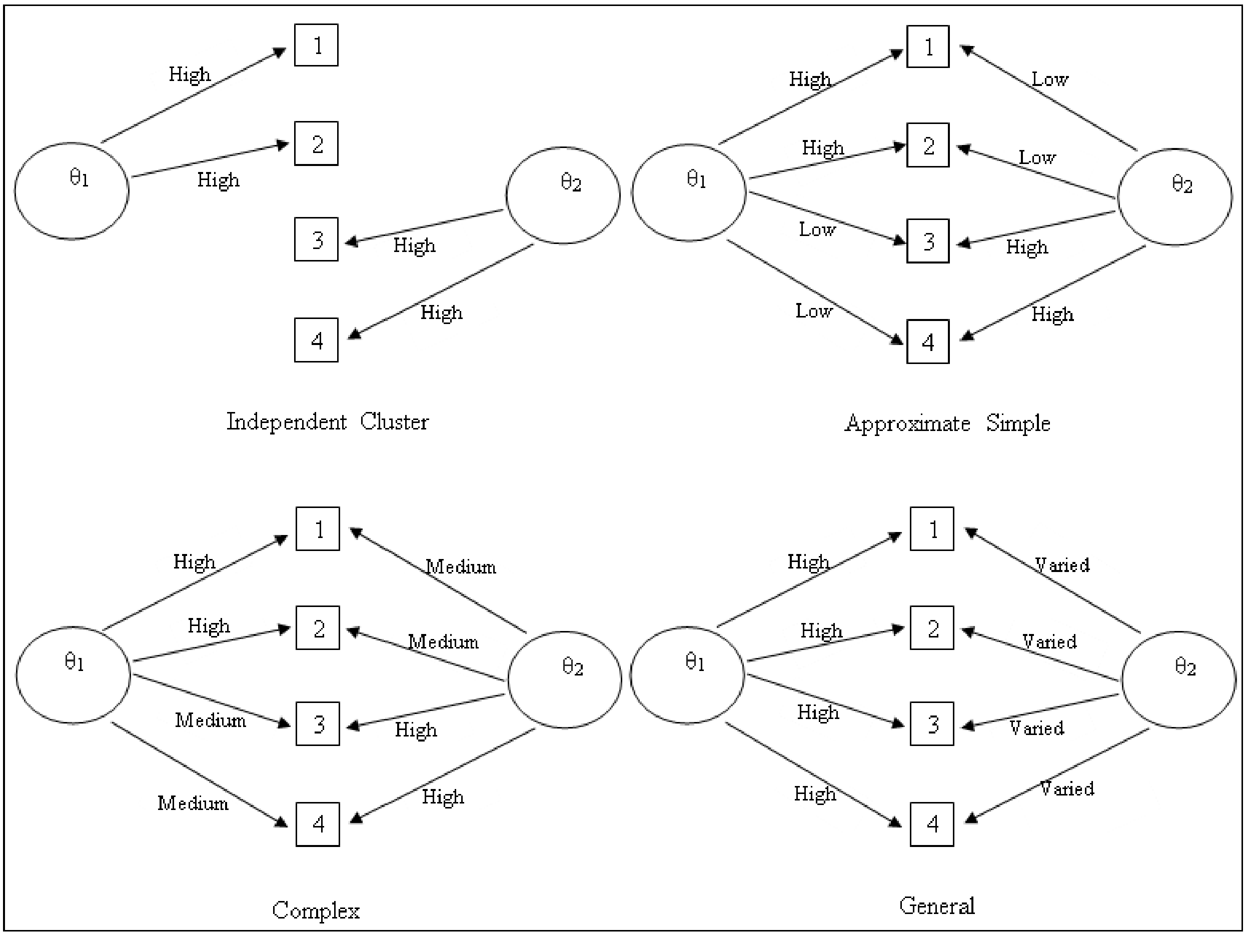

Figure 1 Conceptual models for the four types of data structures.

Note: High=The loading is high in size, medium= The loading is medium in size, low=The loading is low in size, varied=The loading is varied from low to high.

The ability levels for a total of 1,000 examinees were generated from a multivariate normal distribution with a mean vector of 0 and a correlation matrix of $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ where ρ is the correlation between two dimensions. Once item banks and ability levels were generated, the item response data were simulated for each of the 32 corresponding simulation conditions using the multidimensional three parameter logistic (M3PL) model (Reckase, 1997), which defines the probability of correct response for person *j* to item *i* as

$$P_{ji1} = P\left(x_{ji} = 1 \middle| \boldsymbol{\theta}\right) = c_i + \frac{1-c_i}{1+\exp\left[-\mathbf{a}_i'(\boldsymbol{\theta}-b_i\mathbf{1})\right]}, \qquad (1)$$

where $\boldsymbol{\theta}$ is a vector indicating a person's location in a multidimensional latent space, b_i is the difficulty parameter, $\mathbf{a}'_i$ is a vector of discrimination parameters**,** and c_i is the guessing parameter for the *i*th item. Fifty replications were performed for each unique condition. After data are generated for each condition, MCAT analysis was conducted using the Simu MCAT software (Yao, 2011b).

Evaluation Criteria

To evaluate the performance of the item selection methods under various latent structures, we examined five sets of statistics representing two outcomes of interest: (a) item pool utilization as indicated by item pool usage and test length, and (b) measurement precision as indicated by absolute bias, mean square error and the fidelity coefficient. For item pool utilization, we calculated item pool usage as the percentage of items used in one replication across 1,000 examinees, and then averaged over the fifty replications. We calculated test lengths as the average number of administered items in one replication across 1,000 examinees, and then averaged over the fifty replications. For the accuracy of ability estimates, we computed the overall absolute bias (ABSBIAS), mean square error (MSE), and fidelity coefficient (*r*) for the composite dimension. These accuracy indices are defined as

$$ABSBIAS = \frac{1}{N}\sum_{j=1}^{n} |\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_{jTRUE}|, \tag{2}$$

$$MSE = \frac{1}{N}\sum_{j=1}^{n} (\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_{jTRUE})^2, \tag{3}$$

and

$$r_{\theta,\theta_{TRUE}} = \frac{cov(\theta,\theta_{TRUE})}{\sigma_\theta \sigma_{\theta_{TRUE}}}, \tag{4}$$

Where n is the total number of examinees and N is the number of replications, $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_{jTRUE}$ are the estimated and true theta for examinee j, respectively, and $cov(\theta, \theta_{TRUE})$ is the population covariance, σ_θ and $\sigma_{\theta_{TRUE}}$ are the respective standard deviations. We used the same strategy when we calculated final absolute bias, MSE and the fidelity coefficient. Then, factorial ANOVA procedures were conducted to examine the effect of the three simulation factors on the five outcomes. SPSS version 19 was used for the ANOVA analyses.

Independent ClusterApproximate SimpleComplexGeneral

$$\begin{bmatrix} .92-1.39 & 0 \\ .92-1.39 & 0 \\ .92-1.39 & 0 \\ 0 & .92-1.39 \\ 0 & .92-1.39 \\ 0 & .92-1.39 \end{bmatrix} \begin{bmatrix} .92-1.39 & .29-.57 \\ .92-1.39 & .29-.57 \\ .92-1.39 & .29-.57 \\ .29-.57 & .92-1.39 \\ .29-.57 & .92-1.39 \\ .29-.57 & .92-1.39 \end{bmatrix} \begin{bmatrix} .92-1.39 & .57-.92 \\ .92-1.39 & .57-.92 \\ .92-1.39 & .57-.92 \\ .57-.92 & .92-1.39 \\ .57-.92 & .92-1.39 \\ .57-.92 & .92-1.39 \end{bmatrix} \begin{bmatrix} .92-1.39 & .29-1.39 \\ .92-1.39 & .29-1.39 \\ .92-1.39 & .29-1.39 \\ .92-1.39 & .29-1.39 \\ .92-1.39 & .29-1.39 \\ .92-1.39 & .29-1.39 \end{bmatrix}$$

Figure 2. Discrimination matrices across the four type of data structures.

RESULTS

The results for each outcome across the four item selection methods are reported under different data structures and different correlation settings (see Figures3-7). To assess for statistically significant patterns, a factorial ANOVA was conducted separately for each outcome with the three study factors as independent variables. The results are presented in Table 2.

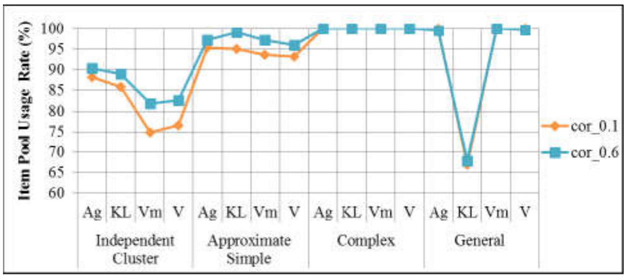

Figure 3 Item pool usage rates across the conditions.

Table 2 ANOVA Results Across The Dependent Variables

	Pool Usage		**Test Length**		**Absolute Bias**		**Mean Square Error**		**Fidelity**	
Source of Variance	***p***	**η²**	***p***	**η²**	***p***	**η²**	***p***	**η²**	***p***	**η²**
IS*DS*TC Structure*Theta Correlation	**.68**	**.00**	**.00***	**.00**	**.00***	**.00**	**.02***	**.00**	**.00***	**.00**
IS*DS	.00*	.44	.00*	.01	.00*	.01	.00*	.01	.00*	.02
IS*TC	.42	.00	.00*	.00	.22	.00	.23	.00	.00*	.00
DS*TC	.00*	.00	.00*	.20	.00*	.02	.00*	.01	.00*	.02
IS	.00*	.08	.00*	.00	.00*	.01	.00*	.00	.00*	.00
DS	.00*	.35	.00*	.65	.00*	.91	.00*	.93	.00*	.91
TC	.00*	.00	.00*	.11	.00*	.03	.00*	.02	.00*	.00

Note: IS= Item Selection, DS= Data Structure, TC= Theta Correlation, *=Significant at alpha level of 0.05.

Item Pool Utilization

Item pool usage. The interaction of item selection method and data structure explained the largest proportion of item pool usage variance (η^2=.44), controlling for all other factors. It was the KL method's relationship to data structures that causes this interaction (see Figure 3). For the Ag, Vm, and V methods, pool usage increased as the data structure moves from independent cluster structure to approximate simple, complex, general structures. However, with the KL method, pool usage showed a sharp decline when used within the general data structure. With a few minor exceptions, as the size of theta correlation increased, pool usage increased.

Test length. The interaction of data structure and theta correlation explained a meaningful proportion of test length variance (η^2=.20), as did the main effect of data structure (η^2=.65). The main effect of data structure is clear in the Figure 4 and indicates that the complex data structure was consistently associated with the longest test length. This was particularly true when the correlation between thetas was low, andit is notable that test length was by far the longest for complex structures of poorly correlated dimensions, regardless of item selection method. This explained the interaction discussed above.

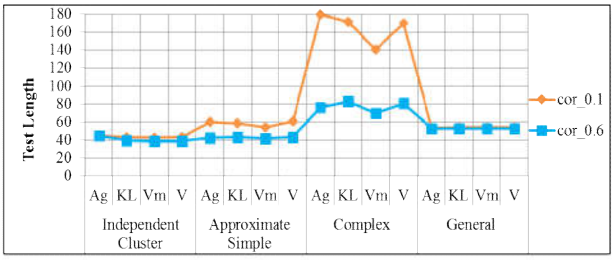

Figure 4 Test lengths across the conditions.

Measurement Acuracy

Absolute bias. The main effect of data structure explained the largest proportion of absolute bias variance (η^2=.91), controlling for all other factors. The results were presented in Figure 5. The main finding was that the type of data structure had a much larger effect on the size of absolute bias than did the method of item selection and theta correlation. Moreover, regardless of the item selection method, the magnitude of the absolute bias decreased as the data structure moves from independent cluster to approximate simple, to complex. Another finding was that with a few exceptions, as the theta correlation decreased, the size of absolute bias decreased. Again, the gap was more apparent under complex data structure, and followed by approximate simple. However, the graph shown in Figure 5 did not differ substantially when explored

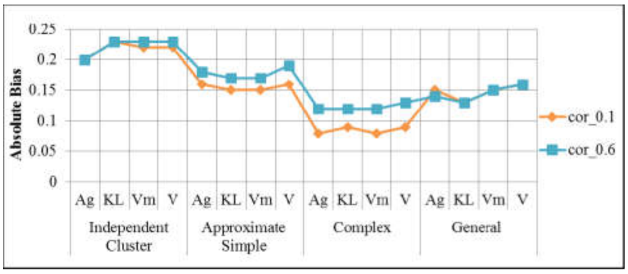

Figure 5 Size of absolute bias across the conditions.

within each of the levels of other item selection methods; the practical findings and interpretations were consistent across the Ag, KL, $V_m$and V item selection methods. Mean square error. The main effect of data structure explained the largest proportion of mean square error variance (η^2=.93), controlling for all other factors. The results were presented in Figure 6. The findings and interpretations for the outcome of mean square were the same with the findings and interpretations for the outcome of absolute bias. These include: a) the large impact of data structure on mean square error, b) the small effect of the size of the theta correlation magnitude on mean square error, c) the magnitude of mean square error varying as a function of data structure, and d) the negligible effect of item selection method on mean square error.

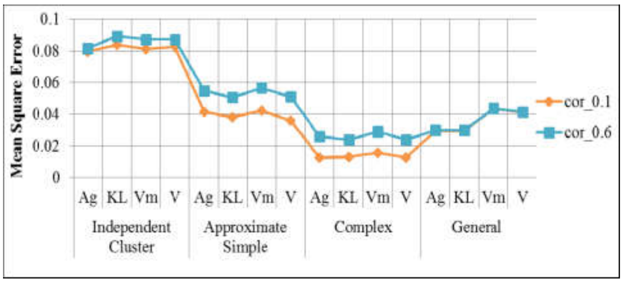

Figure 6 Size of mean square errors across the conditions.

Fidelity coefficient. Based on the factorial ANOVA findings, the main effect of data structure explained the largest proportion of mean square error variance (η^2=.91), controlling for all other factors. The results were presented in Figure 7. The main finding was that fidelity coefficients increased as the data structure moves from independent cluster to approximate simple, to complex. The effect of theta correlation was higher under independent cluster data structure (see Figure 7) across all item selection methods except the Ag. The item selection methods almost always produced the same results across all conditions except a few particular conditions. The exceptions occurred for the Ag method (a) under independent cluster structure and moderate theta correlation, and (b) under general structure and low theta correlation (see Figure 7).

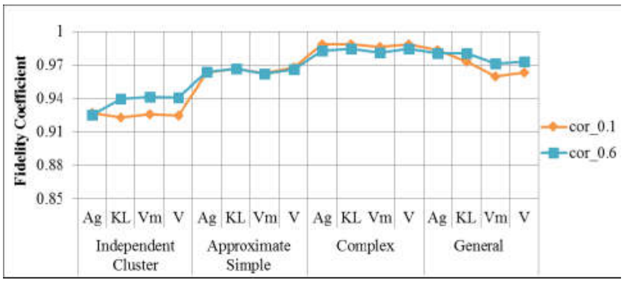

Figure 7 Fidelity coefficients across the conditions.

DISCUSSION

Previous research has tested MCAT algorithms (e.g., item selection, stopping, ability estimation… etc.) on item pools that suggest independent cluster multidimensional data structure. However in practice, some items measure nuisance abilities and/ or items will be intended to measure multiple domains that are hard to disentangle at the item level. When some of the non-zero relationships between observed variables and unobserved variables are ignored (i.e., fixed to zero), the correlation among the latent factors will often be overestimated, thereby leading to distorted findings from the model (Asparouhov & Muthen, 2009; Sass & Schmitt, 2010). We expect, and should appropriately model, multidimensional data structures, and research is needed on how these structures impact adaptive testing methods.

We hypothesized that the multidimensional structure of the item pool and the size of theta correlation would impact the performance of item selection methods in multidimensional data structures in a manner that is currently unknown to researchers and practitioners. The results showed that there were many apparent interaction effects between data structure, item selection method, and theta correlation magnitude in their effects on item pool utilization and measurement accuracy. The choice of item selection methods should depend on data structure, dimension correlation, and the most important outcomes for a particular application (e.g., item pool usage may be more important than measurement accuracy in a particular operational test). As a whole, the study showed the importance of considering more complex multidimensional data structures in operational MCAT applications. We discuss below some core findings.

On data structures. We found that the type of the underlying data structure had a large impact on both item pool utilization and measurement accuracy outcomes. Taking into account the totality of the results, approximate simple data structures were most often associated with the highest measurement efficiency. This was followed by general, independent cluster and then complex data structures. That said, there are other intricacies to the relationship between data structure and the study outcomes. For example, while the independent cluster structure is often used in practice (e.g., Wang & Chang, 2009; Yao, 2013), we found many instances in which other data structures, such as approximate simple and/or general, were associated with better outcomes. Also, when general structure is of interest to test developers, it is recommended to avoid using the KL method due to very low pool usage. Moreover, even though measurement accuracy and pool usage across the conditions was often higher under complex structures, it takes longer tests to achieve this no matter which item selection method is used. Thus, it is wise to avoid having item pools showing complex dimensional structures unless test length is of no concern.

On item selection method. The study found that the type of item selection method usually had only a minor effect on both item utilization and measurement accuracy. However, we offer a few suggestions that may be helpful in practice. The Ag method appears to be the best choice with respect to item pool usage because a) it was the method least affected by differences in the magnitude of theta correlation (see Figure 3), and b) it used either the same amount of different items as the other methods or higher than others (see Figure 3). When a complex dimensional structure is present, the Vm method is advised to minimize test lengths but all item selection methods could be expected to produce similar measurement accuracy. The KL method is not recommended for any of the studied testing conditions because it always produced the same or worse outcomes across all conditions. It was due to the fact that we did not control item exposure so, the KL method consistently selected high quality (e.g., informative) itemsfrom the pool especially under general data structure. Hence, we conclude that when this method is of an interest, item exposure rate should be taken under control.

On theta correlation. We found that theta correlation played amoderaterole on measurement accuracy outcomes but a major role on test utilization outcomes. As the size of the theta correlation decreased from moderate to low, we generally obtained better measurement accuracy. We however needed more items to achieve this. In many cases, pool usage was higher and test length was lower when theta correlation was moderate. The effect of the theta correlation magnitude on these outcomes was greater under complex data structure. In practice, we do not have control over the magnitude of theta correlation between latent abilities as it is determined by the construct(s) of interest for testing. However, we know that before administering an adaptive test or when constructing an item bank for a high stakes test, items are always screened, pre-tested, and item statistics are computed (Dorans, *et al*., 2000). The retained items are then placed into the item pool. After this point, the nature of the theta correlations is often known. As a result, appropriate decisions aligned with that theta correlation can be based on some of our study findings.

We believe the findings in this study are important as they lay the foundations for evaluating a broad set of factors that predict MCAT outcomes. It is possible to extend this study by considering more factors. For example, the performance of item selection method depends on the quality of the item pool, which is related to the match of thetas and difficulty parameters (Choi & Swartz, 2009). This study can be replicated by having item pools that vary in quality. Other factors that could be considered include the size of the item pool, the balance of the item pool (e.g., having an unequal number of items associated with each dimension), the distribution of true thetas, and the number of dimensions. More importantly, we intentionally did not use item exposure control and content control because these two components have the potential to confound the effects of item selection methods (D. Weiss, personal communication, March 7, 2015). In order to evaluate the effect of these two confounds, future research can be conducted with three more conditions that would include (1) item exposure control only, (2) content control only, (3) either exposure or content controls.

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003).Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement, Issues and Practice, 22*(3), 37-53.

Asparouhov, T., &Muthén, B. (2009).Exploratory structural equation modeling. *Structural Equation Modeling, 16*(3), 397.

Beguin, A. A., &Glas, C. A. W. (2001).MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541-562.

Bolt, D. M. (2001).Conditional covariance-based representation of multidimensional test structure. *Applied Psychological Measurement, 25*(3), 244.

Bolt, D. M., &Lall, V. F. (2003). Estimation of compensatory and non compensatory multidimensional item response models using markov chain montecarlo. *Applied Psychological Measurement, 27*(6), 395-414.

Choi, S. W., Grady, M.W., & Dodd, B. G. (2011).A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71*(1), 37-73.

Choi, S. W., & King, D. R., (2015). R Package MAT: Simulation of Multidimensional Adaptive Testing for Dichotomous IRT Models. *Applied Psychological Measurement*, 1-2.

Choi, S. W., & Swartz, R.J., (2009).Comparison of CAT Item Selection Criteria for Polytomous Items. *Applied Psychological Measurement, 33*(6), 419-440.

Dorans, N.J., Eignor, D., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., &Thissen, D. (2000).*Computerized Adaptive Testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Douglas, J., Kim, H. R., Roussos, L., Stout, W., & Zhang, J. (1999).LSAT Dimensionality Analysis for the December 1991, June 1992, and October 1992 Administrations. Statistical Report. LSAC Research Report Series.

Finch, H., Stage, A.K., & Monahan, P., (2008).Comparison of Factor Simplicity Indices for Dichotomous Data-DETECT R, Bentler's Simplicity Index, and the Loading Simplicity Index. *Applied Measurement in Education*, *21*, 41-64.

Finkelman, M., Nering, M. L., & Roussos, L. A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement*, *46*(1), 84-103.

Florida Department of Education (January 2007). Florida Comprehensive Assessment Test: Technical Report For 2006 FCAT Test Administrations. San Antonio, TX. Retrieved from http://fcat.fldoe.org/pdf/fc06tech.pdf

Frey, A., & Seitz, N., (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation, 35*(2-3), 89-94.

Gierl, M. J., Leighton, J. P., & Tan, X. (2006). Evaluating DETECT classification accuracy and consistency when data display complex structure. *Journal of Educational Measurement, 43*(3), 265-289.

He, W., Diao, Q., & Hauser, C. (2013, April). *A Comparison of Four Item-Selection Methods for Severely Constrained CATs.* Paper presented at the NCME annual meeting, San Francisco, CA.

Hirsch, T.M., & Miller, T. R. (1991).Investigation of the Dimensional Structure of the P-ACT. Retrieved from https://www.act.org/research/researchers/reports/pdf/ACT_RR91-01.pdf

Jang, E. E., & Roussos, L. (2007).An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, *44*(1), 1-22.

Jin, R. (2010). *Sample size in exploratory factor analysis with ordinal data*.(Unpublished doctoral dissertation), University of Florida, Gainesville, FL.

Lee, Y.H., Ip, E. H., &Fuh, C.D. (2008).A Strategy for Controlling Item Exposure in Multidimensional Computerized Adaptive Testing. *Educational and Psychological Measurement*, *68*(2), 215-232.

Linacre, J. M. (1988). Simple and effective algorithms: Computer-adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

R Development Core Team, 2013.*R:* A language and environment for statistical computing, reference index version 2.2.1. R Foundation for Statistical Computing. Vienna: Austria. URLhttp://www.R-project.org.

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics*, *4*(3), 207-230.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25-36.

Reckase, M. D. (2009). Multidimensional item response theory. New York, NY: Springer.

Riley, B. B., Dennis, M. L., & Conrad, K. J. (2010). A comparison of content-balancing procedures for estimating multiple clinical domains in computerized adaptive testing: Relative precision, validity, and detection of persons with misfitting responses. *Applied Psychological Measurement, 34*(6), 410-423

Sands, W.A., Waters, B.K., &McBride, J.R. (Eds.). (1997). Computerized Adaptive testing: from inquiry to operation. Washington, DC: American Psychological Association.

Sass, D. A. (2010). Factor loading estimation error and stability using exploratory factor analysis. *Educational and Psychological Measurement*, *70*(4), 557-577.

Sass, D. A., & Schmitt, T. A. (2010).A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, *45*, 73-103.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*(2), 331-354.

Segall, D. O. (2005).Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*. New York, NY: Academic Press.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, *24*(4), 398-412.

van der Linden, W. J., &Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice.* Boston: Kluwer.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*(4), 575-588.

Wang, C., & Chang, H.H. (2009). Kullback-Leibler information in multidimensional adaptive testing: theory and application. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved from www.psych.umn.edu/psylabs/ CAT Central/

Wang, C., & Chang, H. H. (2011). Item selection in multidimensional computerized adaptive testing-gaining information from different angles. *Psychometrika*, *76*(3), 363-384.

Wang, C., Chang, H.H, &Boughton, K.A. (2012).Deriving Stopping Rules for Multidimensional Computerized Adaptive Testing. *Applied Psychological Measurement.37*(2) 99-122.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375.

Yao, L. (2010). Reporting valid and reliability overall score and domain scores. *Journal of Educational Measurement.47*(3), 339-360.

Yao, L. (2011b). Simu MCAT: Simulation of multidimensional computer adaptive testing [Computer software]. Monterey, CA: Defense Manpower Data Center.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and Applications. *Psychometrika*, *77*(3), 495-523.

Yao, L. (2013). Comparing the Performance of Five Multidimensional CAT Selection Procedures with Different Stopping Rules. *Applied Psychological Measurement, 37*(1), 3-23.

Yao, L. (2014). Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constraints. *Journal of Educational Measurement.51*(1), 18-38.

Yao, L., &Boughton, K. A. (2007).A Multidimensional Item Response Modeling Approach for Improving Subscale Proficiency Estimation and Classification. *Applied Psychological Measurement,31*(2), 31-83.

Zhang, J. (1996). *Some fundamental issues in item response theory with applications* (Order No. 9712500).Available from Pro Quest Dissertations & Theses Global. (304295412). Retrieved from http://search. proquest.com/docview/304295412?accountid=10920

How to cite this article:

Halil Ibrahim Sari.2017, Comparing Item Selection Methods In Multidimensional Computerized Adaptive Testing Under Different Types of Multidimensional Structures. *Int J Recent Sci Res.* 8(8), pp. 19269-19276.
DOI: http://dx.doi.org/10.24327/ijrsr.2017.0808.0659

****