Anima Indonesian Psychological Journal
2018, Vol. 33, No. 2, 125-146
https://doi.org/10.24123/aipj.v33i2.1583
0215-0158

Does Multidimensionality Cause DIF?

Ali Ridho
Faculty of Psychology
Universitas Islam Negeri Maulana Malik Ibrahim Malang

The differential item functioning (DIF) of an item that initially assumed unidimensional is frequently attributed to the assumption of multidimensionality. Therefore, it is important to test the assumption that multidimensionality causes an item to be functionally different between the disadvantaged group (i.e. focal group) and the benefited group (i.e. reference group) on the Aptitude Potential Test for New Student Selection in State Islamic University (AP SPMB-PTAIN). This study aims to: (a) explore and confirm the internal structure of AP SPMB-PTAIN; (b) identify items containing DIF based on the types of school the candidates attended (Madrasah Aliyah/MA, that is, secondary education managed by the Ministry of Religious Affairs, or regular high school/SMA); and (c) evaluate the multidimensionality effects on DIF. The data analyses (N = 10,000) showed that: (1) the internal structure of AP SPMB-PTAIN is semi-complex multidimensional; (2) 15 items contain DIF UIRT (12 items benefited high school graduates while three items benefited MA graduates); five items contain DIF MRT that benefited high school graduates; and (3) the multidimensionality difference between the focal and reference group did not appear to correspond to DIF.

Keywords: multidimensionality, item response theory, DIF, MA-SMA, aptitude test

Hadirnya keberfungsian butir yang berbeda (DIF) pada sebuah butir yang diasumsikan unidimensi sering diatribusikan pada asumsi multidimensionalitas. Oleh sebab itu, sangat menarik untuk membuktikan bahwa multidimensionalitas memicu butir menjadi DIF antara kelompok fokal (dirugikan) dan Seleksi Mahasiswa Baru Perguruan Tinggi Agama Islam Negeri (PA SPMB-PTAIN). Penelitian kelompok referensi (diuntungkan) pada Tes Potensi Akademik ini bertujuan untuk: (a) mengeksplorasi dan mengonfirmasi struktur internal PA SPMB-PTAIN; (b) mengidentifikasi butir-butir yang mengandung DIF berdasarkan kelompok lulusan (MA-SMA); dan (c) mengevaluasi efek multidimensionalitas pada DIF. Berdasarkan hasil analisis pada data respons (N = 10.000), hasil penelitian menunjukkan bahwa: (1) struktur internal PA SPMB-PTAIN bersifat multidimensi semi kompleks; (2) 15 butir mengandung DIF UIRT (12 menguntungkan lulusan SMA, tiga menguntungkan lulusan MA); lima butir mengandung DIF MIRT yang menguntungkan lulusan SMA; dan (3) perbedaan multidimensionalitas antara kelompok fokal dan referensi tidak terbukti terkait dengan DIF.

Kata kunci: multidimensionalitas, item response theory, DIF, MA-SMA, tes potensi akademik

Item response theory (IRT) has been widely applied in the field of psychology and education (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; McKinley & Mills, 1985). In its development, the IRT is able to accommodate participants' responses both in unidimensional IRT (UIRT) and multidimensional IRT (MIRT) models. The UIRT model is best used on the unidimensional test, while the MIRT model is more suitable for multidimensional tests. Problems may occur when multidimensional data are treated as unidimensional, which is common in test evaluation/data analysis practice today. Such an erroneous practice can threaten the validity of the measurement score. This paper focuses on the evidence based on internal structure as stated in the new release of the Standards for Educational and Psychological Testing, Differential Item Functioning (DIF) is considered as validity evidence based on internal structure (AERA, APA, & NCME, 2014). Although

Correspondence concerning this article should be addressed to Ali Ridho, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Jalan Gajayana 50 Malang, 65144 Indonesia. E-mail: aliridho@uin-malang.ac.id

published in the United States, the standard occupies an important position in the testing community worldwide (Zumbo, 2014).

Evidence of validity based on the internal structure can be obtained by observing the interactions of the dimensions as well as the load of constructs, dimensions, and items that reveal a particular construct (Figure 1). The structure of constructs can be unidimensional, multidimensional between items, and multidimensional within-items. In its early development, IRT was based on unidimensionality assumption; where only one latent attribute (θ) underlies the response of test participants (Hambleton et al., 1991; Lord, 1980). Another assumption in IRT related to unidimensionality is local independence (LI) (Lord, 1980) that is, given the item parameters are known, the probability of answering an item correctly is influenced only by one θ. Such item is unidimensional, revealing one θ. However, if there is another θ that is needed to explain the participants' performance - other than the θ that was the measurement goal-, the LI is automatically declined. This situation means the test takers need more than one θ to correctly answer the item or the item becomes multidimensional, revealing more than one θ.

To ensure the internal structure aligned with the developed construct conception, any test that is considered as the unidimensional needs to provide proof that participants answered the item solely based on a single θ which is the measurement goal. In fact, it is possible that the participants need more than one θ to correctly answer the test items. This means that the data might actually be multidimensional. It is problematic when the participant's multidimensional data are treated as unidimensional, and put into question the unidimensional assumption in UIRT.

If the internal structure has been verified in a population, the question is whether the structure applies equivalently when the population consists of important grouping variables, such as gender (female-male) or area (Java-outside Java, rural-urban). When dimensions loadings on items are equal between the compared groups, multiple group invariance, or construct equivalence can be claimed (Ridho, 2014; Stucky, Gottfredson, & Panter, 2012). Invariance or construct equivalence is proven by showing either equal dimensions' loadings between two groups at either the construct level (the equivalence of dimensional content) or the dimensional level (the equivalence of content item).

The academic aptitude of prospective students has become a common criterion in new students' selection. In an Academic Aptitude (AP) test that is claimed as unidimensional, there will only be one latent attribute (θ) that influences the performance of the participants in answering the correct items in the test. Two major groups of potential participants are high school (SMA) and Madrasah Aliyah (MA) graduates. If the items in AP test are equivalent (invariance),

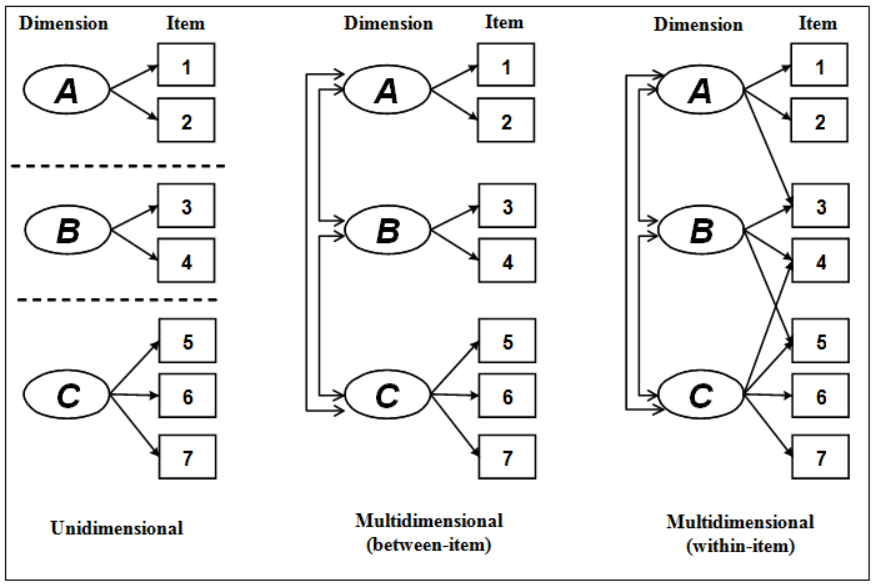

Figure 1. Graphical representation of Unidimensional, Multidimensional between-item, and Multidimensional within-item models.
[adapted from Cheng, Wang, and Ho (2009)]

$P(\theta)$ on each item will be influenced solely by the number of θ (academic potential attribute) from each participant, regardless of their type of school (SMA/ MA). When two groups from SMA and MA with the same ability (θ) have different P(θ) on an item, the principle of invariance (equivalence) is not proven. This phenomenon is known as differential item functioning (DIF), which indicate a disparity of performance between two groups that have equal ability. Hence, it can be said that in addition to academic aptitude there are other θs that were also represented by the items. DIF indicates the failure to reach validity between groups (Huang, Wilson, & Wang, 2016) which has been one of the themes in pursuing validity evidence based on internal structure.

The presence of additional θ which contributes to $P(\theta)$ on an item makes the item no longer unidimensional, but multidimensional. Some researchers hold a premise that multidimensionality is the cause of differential item functioning (DIF) in a test claimed to be unidimensional. DIF might be caused by the presence of at least one additional θ that is measured by the items (e.g. Ackerman, 1994; Camilli, 1992, Gierl, 2005; Gierl, Bisanz, Bisanz, & Boughton, 2003; McDonald, 2000; Oshima & Miller, 1992; Roussos & Stout, 1996a). Hence, the identification of DIF on an item indicates the presence of an additional θ measured by the item. The inclusion of multidimensional items may make DIF function differently between groups when calibrated with unidimensional assumptions (Ackerman, 1991, 1994; Furlow, Ross, & Gagné, 2009). This study investigates whether multidimensionality on an item will trigger a DIF on it.

In order to prove the multidimensionality effect on DIF, multidimensional response data are required as the object of the research. For that purpose, the researcher used the data on Academic Aptitude test (AP) which have been proven to be multidimensional (Azwar & Ridho, 2012; Ridho, 2011). AP response data used in this research was AP Selection of New Student Admissions of State Islamic University (SPMB-PTAIN) in 2012. This test consists of two subtests; the verbal subtest (analogical, logical, and analytical components); and the quantitative subtest (arithmetic, comparative, and geometric shapes components). Both subtests are assumed to be unidimensional and form the structure of the measured potential.

This study aims to answer three questions about the validity of AP scores: (1) internal structure; (2) identification of AP items containing DIF based on school group (MA-SMA); and (3) multidimensionality effects on DIF.

Unidimensional Item Response Theory (UIRT)

As noted earlier, the IRT model is based on unidimensional assumptions and local independence (LI). These two assumptions will form an item characteristic curve (ICC) that is invariant in the grouping of certain participants. The probability of answering correctly on an item is modeled in a probability function, $P(\theta)$.

The proportion or probability of a student answering correctly on an item, $P_i(\theta)$, is not affected by the group from which they originated, rather, it is affected only by one aspect, that is, the ability level, θ (Figure 2). The shaded area can be occupied by group A or group B. Regardless the group of origin, when the ability is similar then the probability of answering correctly on a point will be similar. Invariance does not only occur in item parameters. The invariance item parameters will be followed by the invariance of participant's ability parameters, θ. When the characteristics of each item that makes up a test are known, a group of items where the participant's ability remain the same even though the estimates are based on different item groups. Such a condition is called participant parameter invariance.

In fact, in a multiple-choice test participant might correctly guess the answer, which adds pseudo guessing as a component in the probability of answering correctly. This is highly possible on multiple-choice items, so that Birnbaum (1968, in Lord, 1980) modified the logistics model of two parameters (2PL);

$$P(u_{ij}=1 \mid a_i,b_i,\theta_j)=\frac{e^{Da_i(\theta_j-b_i)}}{1+e^{Da_i(\theta_j-b_i)}} \qquad (1)$$

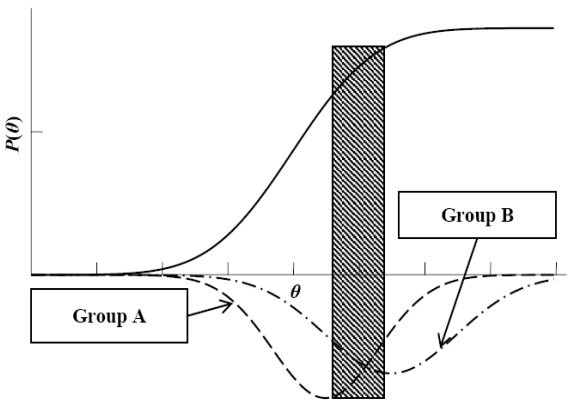

Figure 2. Parameter Invariance in IRT. (Hambleton, Swaminathan, & Rogers,1991)

where $P(u_{ij}=1| a_i,b_i,\theta_i)$ is the probability to answering item *i* correctly when test takers in ability level is at θ_j, and a_i, b_i are the parameters of difficulty level and item discrimination *i*; to a three parameters logistical model formed (3PL):

$$P(u_{ij}=1|a_i,b_i,c_i,\theta_j) = c_i + (1- c_i)\frac{e^{Da_i(\theta_j-b_i)}}{1+e^{Da_i(\theta_j-b_i)}} \quad (2)$$

by including an additional parameter that represents the contribution of guessing to the true-answer probability (c_i), this model would be suitable for larger sample sizes (Martín, Pino, & Boeck, 2006) as in this study.

Multidimensional Item Response Theory (MIRT)

The items in the test often measure composite abilities, which were not meant to be measured by the test developer in the blueprint. If an item is not sensitive enough to measure more than one latent attribute or participants vary in the same latent attribute, then the interaction between the item and the participant is unidimensional (Ackerman, 1992, 1994).

The concept of MIRT can be viewed as a special case of factor analysis or structural equation, or the development of UIRT (Reckase, 1997). Some possible models that may explain the interaction between the participants and the item are represented in Figure 1 (Cheng et al., 2009). Figure 1 (the left figure) shows the unidimensional structure with the latent attributes being measured are A, B, and C; the calibration of latent attributes is done by overriding the correlation between the latent attributes A, B, and C. In the middle, calibration is performed simultaneously by considering the correlation between latent attributes A, B, and C. Such model is called the multidimensional model with a simple structure. The model in Figure 1 (the right figure) is a complex multidimensional model in which there are items measuring more than one latent attribute.

As a development of UIRT, both in simple and complex form, MIRT can be divided into two types: compensatory model and noncompensatory model (Reckase, 2009). The compensatory model is based on the relation of a linear combination of latent attribute vector, θ. Meanwhile, the noncompensatory model separates latent attributes in response to the items and uses the UIRT model for each latent attribute. This research discusses compensatory model.

The probability of answering correctly on an item *i* is determined by the *m* latent attribute $\theta = (\theta_1, \theta_2, ..., \theta_m)$ where *m* is the number of dimensions used in the model. In the dichotomous response, this model can be represented in the form of:

$$P_i(\theta_1,\ldots,\theta_m)= P_i(u_1= 1|\theta_1,\ldots,\theta_m) = P_i(\theta) \quad (3)$$

The more general form can be seen in (4). The vector ξ represents the parameter of the item, **U** is the score on the item, *u* is a score (0 or 1), *f* is a function that showing the relationship between the participant's location vector (θ) and the probabiity of answering correctly.

$$P_i(U)= u_1 | \theta) = f(\theta,\xi_i,u) \quad (4)$$

Illustrations of UIRT development can be depicted from 2PL model references in equation (1). The model has components in the form of exponent $a(\theta - b)$. When elaborated, it produces $a\theta - ab$. If $(- ab)$ is replaced by d it will form the equation $a\theta + d$. By increasing the 2PL model several dimensions θ form a vector θ, which eventually create $a\theta' + d$. Vector a denotes the vector $1 \times m$ discrimination power parameter and θ is the vector $1 \times m$ that *m* denotes the dimension in the spatial coordinates. Intercept parameter symbolized by *d*. Form 2PL MIRT (M2PL) will be:

$$P(u_{ij} = 1 | \boldsymbol{\theta}_j,\mathbf{a}_i,d_i) = \frac{e^{\mathbf{a}_i\boldsymbol{\theta}'_j+d_i}}{1+e^{\mathbf{a}_i\boldsymbol{\theta}'_j+d_i}} \quad (5)$$

where u_{ij} = response of participant *j* on item *i* (0 or 1); θ_j = latent attribute vector; a_i = vector of discrimination power on item *i*; and d_i = easiness index, related to difficulties level of item *i*. Exponent *e* in this model can be spelled out so the interactions between vector *a* and θ:

$$a_i\theta'_j+d_i=a_{i1}\theta_{j1}+a_{i2}\theta_{j2}+\ldots+a_{im}\theta_{jm}+d_i= \sum_{l=1}^{m} a_{il}\theta_{jl}+d_i \quad (6)$$

In the UIRT model, the probability of answering correctly 0.5 will correspond to the parameter of item's difficulty level, *b*. In MIRT, the concept of probabi-

lity on answering correctly of 0.5 will correspond to the pair of vector θ in *m* dimension. For instance, when $m = 2$, there will be θ_1 and θ_2. For $P(\theta_1, \theta_2) = 0.5$ then $a\theta' + d$ should be equal to 0 since $e^0 = 1$ so $P(\theta_1, \theta_2) = 1 / (1 + 1) = 0.5$. For an item with $a_1 = 0.5$, $a_2 = 1.5$, and $d = -0.7$ then $P(\theta_1, \theta_2) = 0.5$ is obtained for the various pairs θ_1 and θ_2 so as to form a straight line as in Figure 3. The low latent attribute of dimension 1 can be compensated by the height of the latent attribute of dimension 2. That is why this model is referred to as the compensatory model.

Differential Item Functioning (DIF)

To understand the phenomenon of DIF on an item, here is an illustration by Penfield and Lam (2000). The test participants can be grouped into two, namely reference group (R) and focal group (F). Group R is usually referred to the group that is allegedly benefited (SMA graduates), while group F is the group that is allegedly disadvantaged (MA graduate). If the performance on Group F is lower than R and it caused by lower θ, the condition can be considered normal. However, if θ group F is not lower than that of R, then the performance on a particular item is lower, it is likely that a DIF that benefited group R occurred. Theoretically, some experts (e.g. Angoff, 1993; Camilli, 1992; Embretson & Reise, 2000; Hambleton, 2006; Penfield & Camilli, 2007) suggests that an item can be identified as a DIF if there is an unequal probability in answering an item correctly in two groups of test takers with equal abilities, after being on the same ability continuum.

DIF detecting techniques can be conducted in several ways. Some of them are: (1) Mantel-Haenszel (MH) (Dorans, Holland, & Educational Testing Service, 1992; Mantel & Haenszel, 1959); (2) Lord's Chi-Square (LC) (Lord, 1980); (3) Signed Area (SA) and Unsign Area (UA) (Raju, 1990; Raju, Drasgow, & Slinde, 1993); (4) Logistic Regression (LR) (Swaminathan & Rogers, 1990); (5) Discriminant Analysis (DA) (Miller & Spray, 1993); (6) Likelihood Ratio Test (LRT) (Thissen, 2001; Thissen, Steinberg, & Wainer, 1993); (7) simultaneous item bias test (SIBTEST) (Bolt & Stout, 1996; Roussos & Stout, 1996b); and (8) Classification Trees (CT) (Vaughn & Wang, 2008, 2010) that are still under development. The various DIF detection methods are summarized by Santelices and Wilson (2012) into three categories: (1) parameter comparison; (2) area comparison; and (3) the ratio of probability (likelihood).

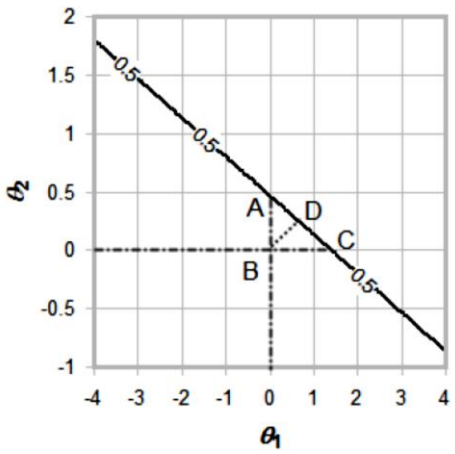

Figure 3. Plot of vector pair so that $P(\theta_1, \theta_2)$

DIF detection methods can be categorized into parametric and nonparametric. In nonparametric method there is no estimation of model parameters, whereas parametric method involves estimation of model parameters (Teresi & Fleishman, 2007). Both, the parametric and non-parametric needs a matching criterion (Patarapichayatham, Kamata, & Kanjanawasee, 2012; Scarpati, Wells, Lewis, & Jirka, 2011). In summary, the authors summarize it in Table 1.

Dimensionality

In the context of a test consisting of several subtests as in this study, the evaluation of the dimensionality determines whether or not a subscore needs to be reported. If the items in the two sub-types are unidimensional, then the scores of these subtests should be combined into one report. (Haberman & Sinharay, 2010; Stone, Ye, Zhu, & Lane, 2010; Yao, 2010,

Table 1
Summary of DIF Method's Characteristics

Method	Base
(1) Mantel-Haenzel (MH)	NP, observed score
(2) Lord's chi-Square (LC)	NP, latent score
(3) sign & unsigned area (SA-UA)	NP, latent score
(4) logistic regression (LR)	P, observed score
(5) determinant analysis (DA)	NP, observed score
(6) likelihood ratio test (LRT)	P, latent score
(7) simultaneous item bias test (SIBTEST)	NP, observed score
(8) classification trees (CT)	NP, observed score

Note. Description: NP = nonparametric; P = parametric.

2011). In relation to DIF detection in the context of MIRT, Snow and Oshima (2009) suggest a thoroughly tested dimension to determine the number of dimensions measured by the test. So, the multidimensionality can be detected early.

As stated before, there are two types of multidimensionality, that is, between and within-items. If an item is shown to be multidimensional within-items in the two groups compared, the DIF occurs conceptually because of the interaction between multidimensionality and grouping variables (MA-SMA). That is, DIF occurs when an item has different dimension load in two different groups. For example, in group MA an item has 0.4 dimension loading on the quantitative and 0.4 on the symbol; while in the SMA group it has 0.3 and 0.5 dimension loading respectively.

There are many ways to test dimensionality, for example, factor analysis (Deng, Wells, & Hambleton, 2008; Finch & Habing, 2007; Glanville & Wildhagen, 2007; Slocum-Gori & Zumbo, 2011), bifactor models (Brown, Finney, & France, 2011; Immekus & Imbrie, 2008; Reise, Morizot, & Hays, 2007), principal component analysis (Chou & Wang, 2010), conditional covariance (Finch & Habing, 2007; Jang & Roussos, 2007; Levy, Mislevy, & Sinharay, 2009). In addition, there are also methods of Mokken Scale Analysis for Polytomous Items (MSP), Dimensionality Evaluation To Enumerate Contributing Traits (DETECT), hierarchical cluster analysis (HCA), and dimensionality tests (DIMTEST) (van Abswoude, van der Ark, & Sijtsma, 2004a). In this research, the author use DETECT and DIMTEST methods.

In the test items that are scored dichotomously, Tate (2003) provides a review of methodological developments on how empirical procedures in evaluating the response structure of test items. He argued that evaluation procedures can be divided into two major groups: parametric and nonparametric. Table 2 is a summary of computer methods and programs that have been developed to help the computing process.

Method

Participants

A total of 53,637 participants took the SPMB-PTAIN in 2012 in two forms. The two forms were developed based on the same blueprint so that it was reasonable to assume the two forms are parallel. Participants were from different regions in Indonesia. Five provinces with the largest participants were Nangroe

Table 2
Methods and Programs for Dimensionality Test Computing

Approach	Computing Program
Parametric	
1. *Exploratory factor analysis of tetrachoric correlations*	MPlus
2. *Confirmatory factor analysis of tetrachoric correlations with robust weighted least squares estimation*	MPlus
3. *Nonlinear factor analysis*	NOHARM
4. *Chi-square test of NOHARM solution*	CHIDIM
5. *Full-information item factor analysis*	TESTFACT
6. *Selected indices for local item dependencies*	IRTNEW
Nonparametric	
7. *Hierarchical cluster analysis of item proximities*	HCA/CCPROX
8. *Test of essential dimensionality*	DIMTEST
9. DETECT *index of dimensionality*	DETECT

Note. Adapted from Tate (2003).

Table 3
Components and Descriptor PA SPMB-PTAIN 2012

Components	Descriptor	Number	Total
1. Analogic (θ_1)	The ability to uncover the relationship between two things, then creating analogical relationship to the relationship between two other things	1 - 15	15
2. Logic (θ_2)	The ability to make the most appropriate decision from two or more premises	16 - 23	8
3. Analytic (θ_3)	The ability to use facts or information presented in a discourse to draw a conclusion appropriately	24 - 38	15
4. Arithmetic (θ_4)	The ability to calculate addition, multiplication, and square	39 - 53	15
5. Comparison (θ_5)	The ability to compare values or quantities	54 - 64	11
6. Geometry (θ_6)	The ability to find geometric symbol pattern	65 - 75	11

Note. Source: Organizing Committee SPMB-PTAIN 2012.

Aceh Darussalam (14.72%), Central Java (11.23%), East Java (10.54%), Lampung (7.41%) and South Sulawesi (6.79%). In term of the origin of the school, 35.29% are high school (SMA) graduates and 42.15% are MA graduates. Participants in this study were 25,744 test takers (5,376 MA graduates, 5,308 graduates of high school, 9,636 men, 15,965 women) who were included in form 1. The data used were participants' response data on form 1, a 5,000 MA graduates and 5,000 SMA graduates that were selected randomly.

The Instrument

The Academic Potential Tests used in this study consists of Verbal and Quantitative Subtests. The Verbal subtest consists of some components such as analogical (15 items), logical (8 items), analytical (15 items), while Quantitative Subtests consist of arithmetic (15 points), comparisons (11 items), and geometry (11 items). Table 3 presents the components, descriptor, serial number, and the number of PA test items. Participants' correct answer was scored 1 and wrong answer was scored 0.

Procedure

The written examination instruments of SPMB-PTAIN were responded simultaneously nationwide on 19th of June 2012, together with other test subjects. Test subjects are administered in exam classrooms throughout PTAIN in Indonesia. Each room contains 20 participants. The tests were administered in this order: (1) Academic Potential; (2) Basic Capabilities; (3) Islamic value; (4) Natural Science; and (5) Social Sciences. The time allocated for AP test was 60 minutes.

Data Analysis

The series of data analysis performed in this study is presented in Figure 4. The explanation is as follows:

(1) To determine characteristics of verbal subtest and quantitative subtest item - according to the Unidimensional Items Response Theory (UIRT) - a parameter item calibration in each subtest is conducted with Marginal Maximum Likelihood method. The procedure applied to every item in each subtest until data matching was gained, both in the item and the

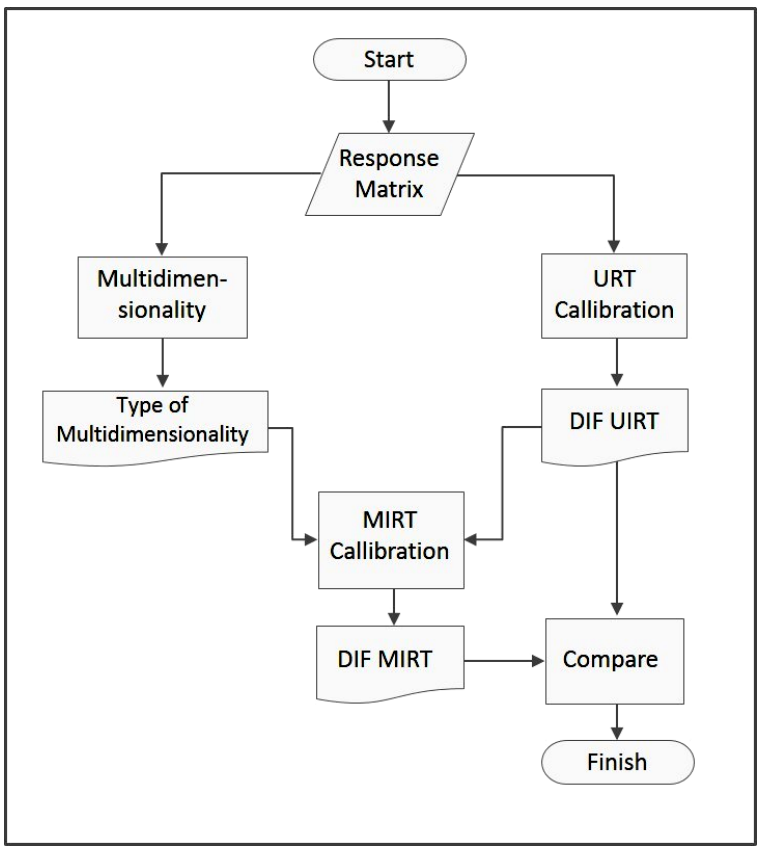

Figure 4. Series of data analysis.

Table 4
Dimensions that Revealed by the SPMB-PTAIN AP Test Items

Dimension	Component	Item Number	Total
1. vocabulary (θ_1)	1. Analogic	1, 2, 4, 5, 9, 11, 12, 13, 15	9
2. verbal (θ_2)	2. Logic	17, 18, 20, 21, 23	5
	3. Analytic	24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38	14
3. quantitative (θ_3)	4. Arithmetic	39, 40, 41, 42, 43, 44, 45, 47, 48, 49, 51, 52, 53	13
	5. Comparison	54, 55, 56, 57, 58, 59, 60, 61, 62, 63	10
4. symbol (θ_3)	6. Geometry	69, 70, 71, 72, 73, 74, 75	7

instrument with the help of BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003);

(2) Dimensional analysis in this research refers to Jang and Roussos (2007) suggestion which apply the exploratory and confirmatory technique on AP dimension structure both in MA and SMA group. Steps taken in each group are: (a) Testing whether the items proved to be unidimensional or multidimensional. This is conducted through exploratory procedures DIMTEST, DETECT, and HCA/CCPROX; and (b) Exploratory findings are used to develop hypotheses followed by confirmatory analysis with the help of DIMTEST. The choice of procedure is also based on research results showing that DIMTEST detection will generate large power to detect deviations from unidimensional assumption (Finch & Habing, 2007; Nandakumar, 1994);

(3) To know the item characteristics based on multidimensional item response theory (M IRT), a calibration is carried out with the help of BMIRT (Yao & Boughton, 2007). Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and Deviance Information Criterion (DIC) criteria is taken to investigate whether the response data is more appropriate for simple or complex models;

(4) To figure out the AP items that identified as DIF UIRT based on the origin school, the Likelihood Ratio Test (LRT) method was used with the help of IRTLRDIF (Thissen, 2001). DIF item justification was based on the effect of DIF UIRT (Meade, 2010). An item is categorized as DIF when the LRT test shows significant results;

(5) To figure out which AP items contains DIF MIRT based on the original school group, the Lord's Chi-Square (LCS) method was used with the help of BMIRT (Yao & Boughton, 2007). Identification of DIF based on MIRT point of view is justified with the reference of DIF MIRT (Suh, 2016).

Table 5
Goodness of Fit of MIRT Model between Simple and Complex Structures

Structure	AIC	BIC	DIC
S1	642744	931605	527394
S2	624241	913101	521726

Note. Description: S1 = simple structure; S2 = complex structure with θ_1 pair with θ_2 and θ_3 pair with θ_4; Akaike's Information Criterion (AIC); Bayesian Information Criterion (BIC); Deviance Information Criterion (DIC). The lower AIC, BIC, and DIC indicate the model is gaining more match with the response data.

Results

In the preliminary analysis, 17 items were not included in the subsequent analysis due to their number of r_{bis} that < 0.25. These items are item number: 3, 6, 7, 8, 10, 14, 16, 19, 22, 28, 46, 50, 64, 65, 66, 67, and 68. As such, the analysis carried out to 58 AP items.

Through DETECT exploratory and confirmatory techniques (Zhang & Stout, 1999) that followed by HCA/CCPROX (Roussos, Stout, & Marden, 1998) and DIMTEST (Stout & Nandakumar, 2006), there are four dimensions that contributed to the construction of AP SPMB-PTAIN: vocabulary *(θ_1)*, verbal (θ_2), quantitative (θ_3), and symbol (θ_4). The distribution of each item is presented in Table 4.

The measure of the model fit of UIRT is based on the difference of -2loglikelihood (G^2) in the three parameters model (3PL) and two parameters model (2PL) resulting in G^2 = 47091,0767 (df = 58, p < .01). Therefore, it is inferred that 3PL model better explained the data than the 2PL model. The fitness of MIRT is presented in Table 5. The three index of model matching (AIC, BIC, and DIC) showed that response data is more suitable when the selected structure is a complex structure where θ_1 pairs with θ_2 and θ_3 pairs with θ_4.

The results of the DIF test in the MS group are presented in Table 6. There are 12 items proven benefiting high school graduates (+) and three items benefiting MA graduates (-).

The explanatory dimension of every item is presented in Table 7 where four items with different dimensions are found.

Discussion

The main purpose of this study was to test whether multidimensionality was proven to cause DIF UIRT. The response data from the Academic Potential Test (AP) of the SPMB-PTAIN, whose participants are all adolescents who have just graduated from secondary education, are divided into two (MA graduates and SMA graduates). This study was designed to investigate whether DIF-detected items can be explained by the multidimensionality contained in it. In the UIRT model, the items are calibrated separately based on their respective subtests (Verbal and Quantitative).

Dimension

A UIRT calibration that treats AP as a unidimensional unity proved to be imprecise. Two-dimensional conception (Verbal and Quantitative) also left a problem. Exploration and confirmation based on the total data showed that the AP revealed four dimensions: vocabulary (θ_1), verbal (θ_2), quantitative (θ_3), and symbol (θ_4). Although the vocabulary dimension is adjacent to the verbal, the two dimensions cannot be regarded as a whole. The verbal ability turns out to be an enhanced vocabulary capability. Likewise, the quantitative dimension is adjacent to the symbol, but cannot also be regarded as a whole.

Table 6
Summary of DIF UIRT and MIRT

No	Items	UIRT	MIRT	No	Items	UIRT	MIRT
1	ANLG02	+		2	ANLG05	-	
3	ANLG09	+	+	4	ANLG13	-	
5	LOGI17	+	+				
6	LOGI18	+	+				
7	LOGI20	+	+				
8	METK39	+	+	12	METK52	-	
9	METK40	+					
10	METK41	+					
11	METK45	+					
13	KOMP55	+					
14	KOMP56	+					
15	KOMP61	+					

S	+	12	5
M	-	3	0
	Total	15	5

Note.
+ = benefited SMA;
- = benefited MA

Table 7
Results of Items Grouping Based on DETECT Exploratory

Dimension	Items	Total
	MA (n = 5,000)	
vocabulary (θ_1)	1, 2, 4, 5, 9, 11, 12, 13, 15, 17, 18, 20, 21, 23, ***37***	15
verbal (θ_2)	24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 38	13
quantitative (θ_3)	39, 40, 41, 42, 43, 44, 45, 47, 48, 49, 51, 52, 53, 55, 56, 59, 60, 61, 62, 63	20
symbol (θ_4)	***54***, ***57***, ***58***, 69, 70, 71, 72, 73, 74, 75	10
	SMA (n = 5,000)	
vocabulary (θ_1)	1, 2, 4, 5, 9, 11, 12, 13, 15, 17, 18, 20, 21, 23	14
verbal (θ_2)	24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38	14
quantitative (θ_3)	39, 40, 41, 42, 43, 44, 45, 47, 48, 49, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63	23
symbol (θ_4)	69, 70, 71, 72, 73, 74, 75	7

Note. Description: bold and italic item number is empirically identified and has different dimension content that differs with the concept design.

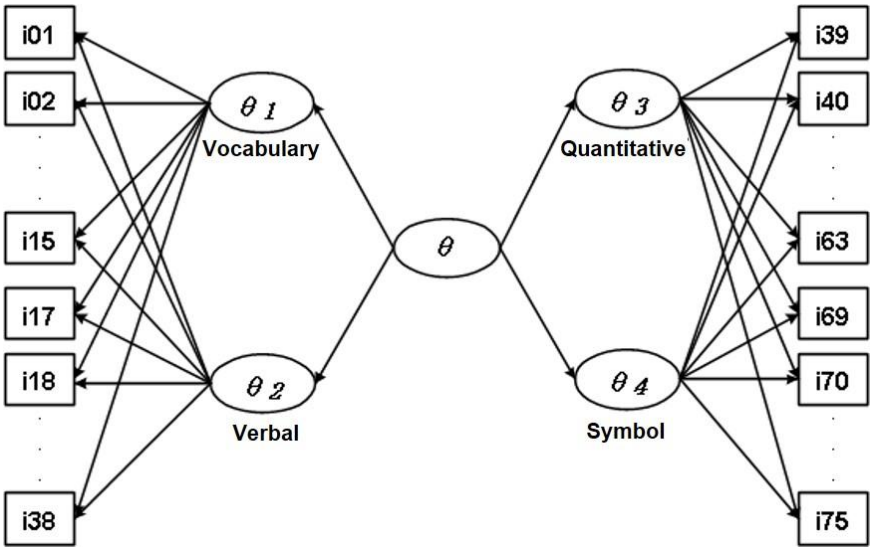

Figure 5. Semi complex structure of AP SPMB-PTAIN.

Three components: analogical, logical, and analytic; in principal measuring same ability, which is verbal reasoning (Allalouf, Hambleton, & Sireci, 1999; Enright, Tucker, & Katz, 1995). Furthermore, Enright et al. (1995) asserted that analogical component measures verbal skills at a simple level, while logical and analytic measure at a more complex level. Based on the exploratory DETECT analysis, the items in the analytic component have close proximity to the items in the logical component. In addition, on reference to the items of the SPMB-PTAIN AP instrument the completion of the items in the logical and analytical components require a verbal ability because it contains sentences that require verbal comprehension. Thus, some parts of the logical and analytical component variance were affected by participants' vocabulary capabilities (Diones, Bejar, & Chaffin, 1996). In summary, analogy (vocabulary) becomes the second dimension for logical and analytical (verbal).

Arithmetic and comparison were designed as one entity in quantitative dimensions. This is empirically proven by grouping items in these two components based on exploratory analysis of HCA/CCPROX or DETECT. On the other hand, empirically speaking the items in the geometry component were always grouped together, but separated from other clusters, either in the exploration procedure through HCA/CCPROX or DETECT. The naming of the symbolic reasoning is based on the fact that although in the SPMB-PTAIN AP test a script is referred to as a geometry component, the content represents more symbolic reasoning. When further examined, the ability of geometric visualization assists to solve some items in the quantitative dimension. Thus, the symbol dimension is the second dimension for the quantitative dimension.

Internal Structure

The results of dimensional exploration through DIMTEST, DETECT, and HCA/CCPROX indicated that participants' response to the SPMB-PTAIN AP test proved to be multidimensional. This multidimensionality is reinforced by previous research results which conclude that AP test is multidimensional (Azwar & Ridho, 2012; Ridho, 2011). These results are in line with the recommendations given by ETS in the usage of GRE test score scores. Because of its multidimensionality, the verbal and quantitative scores should not be added as each gives independent information to each other (ETS, 2007, p. 4).

The internal structure of the response data is more suitable to be identified as semi-complex. The items of vocabulary-verbal revealed both the vocabulary and verbal dimension; while quantitative-symbol items revealed the quantitative and symbol dimension altogether. Complexity visualization of this dimensional structure is presented in Figure 5. According to Adams, Wilson, and Wang (1997) also Cheng, Wang, and Ho (2009), this model may call within-item multidimensional.

The implication of the MIRT model match is for the scoring system. Some scoring methods that can be applied are: (a) UIRT model; (b) high order IRT

model (higher-order IRT, HO-IRT); (c) MIRT model; and (d) bifactor model (Yao, 2010). Taking into consideration the structure (Figure 5), in order to obtain the overall scores estimation of the academic potential of SPMB, a MIRT model method can be applied. If the estimation of academic potential based on MIRT structure gained, then the estimated score of each domain (dimension) can be gained as well. The overall score is obtained by using the maximum information function of the obtained MIRT model. Justification of this method is reinforced by the opinion of Kahraman and Thompson (2011) which coined the term of the unidimensional composite method. Furthermore, through the projection of multidimensional response chamber in each dimension, we can get the score information of each dimension, if necessary.

The scoring practice that used in AP is by taking into account to the number of correct scores every participant obtained. Given the internal structure of PA, the scoring method needs to be changed because the total score has not accommodated the proven multidimensionality. Therefore, scoring by MIRT model method can be used as a reference to produce a composite score that is reliable and has minimum measurement error. In addition, scores of each dimension can also be raised, when needed. In the context of AP SPMB, the application of this method will produce scores as follow: (1) composite AP; (2) vocabulary; (3) verbal; (4) quantitative; and (5) symbols.

DIF Identification

The process of identifying the DIF based on UIRT found 12 items benefiting SMA graduates and three items benefiting MA graduates. While based on MIRT, five items benefiting SMA graduates. Unfortunately, researchers have not been able to find out what item patterns favor high school graduates and what kind of patterns benefit the MA graduate group. To clarify which items are in favor of SMA and MA, the formulation of the items is enclosed in the appendix. Subsequently, examples of ICC item with UIRT model that benefited SMA graduates (Figure 6) and MA graduates (Figure 7) are presented, while MIRT model is presented in Figure 8.

Figure 6 shows the item ANLG09. It appears that SMA graduates have a higher probability of answer correctly than the MA graduates. Hence, it can be said that this item benefited SMA graduates.

The probability of answering METK52 item correctly as shown in Figure 7 suggests the item benefited SMA graduated in low to high (- 1.8 ≤ θ ≤ 1.4) capability. Meanwhile, on the high ability scale, the MA graduates benefited more. However, the SMA graduates are benefited in the wider scale of capability. Therefore, it can be said that this point only benefited the high-ability MA graduates.

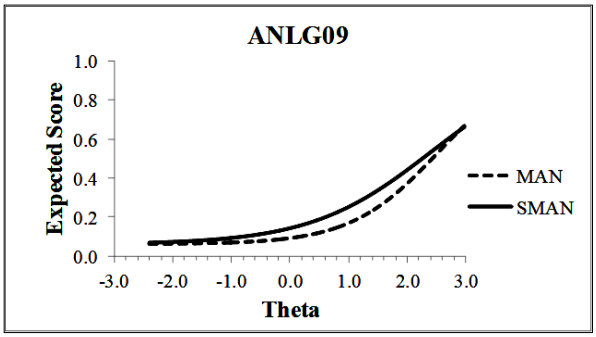

Figure 6. ICC ANLG09 (favor SMA graduates).

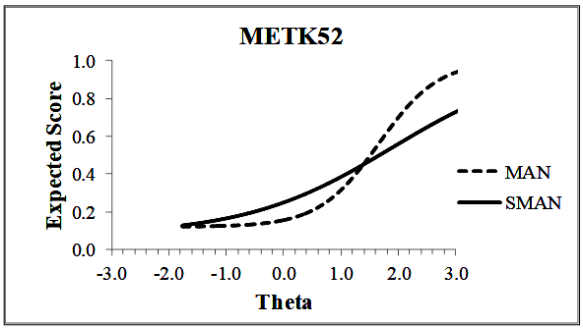

Figure 7. ICC METK52 (favor MA graduates).

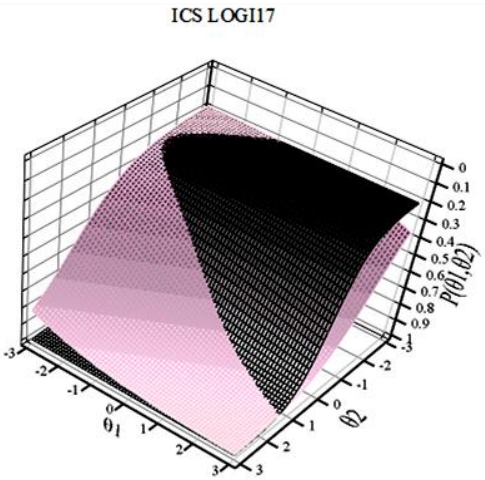

Figure 8. ICS Item ANLG17 in group M and S.

Note. dark ICS for SMA graduates; light ICS for MA graduates; θ_1 is vocabulary dimension; θ_2 is verbal dimension.

The probability of correct answer on item ANLG 17 formed a surface in each MA and SMA group as shown in Figure 8. The picture infers that the probability of correct answer in SMA group (dark ICS) appear to be higher in a certain area, while in another area the MA group (bright ICS) is higher. Thus, it can be stated that the ANLG items have different performance in the MS group. This condition also implies the existence of additional dimension variance (other than θ_1 and θ_2) which influence the performance of MA and SMA graduate participants in answering the item correctly.

Visualization of the ANLG17 item discrimination vector that formed by the MS group is presented in Figure 9. Item discrimination vectors form ∠ 59.55° for MA and ∠ 87.77° for SMA. This asserts that this item, in general, revealed the verbal dimension (θ_2). However, in the high school group, this item is more sensitive than in the MA group. The sensitivity difference triggered the emergence of DIF MIRT in ANLG17.

Multidimensionality Effect

With reference to Table 7, items 37 that originally included in verbal dimension is identified as vocabulary dimension in MA group. Items 54, 57, and 58 fall within the dimension of symbols, although initially they are in quantitative dimension. Based on the premise argued by Ackerman (1991) and Furlow et al. (2009), the item 37, 54, 57, and 58 will experience DIF due to the inclusion of multidimensional items that can make DIF function differently between groups when calibrated with unidimensional assumptions. Nevertheless, the results of data analysis presented in Table 6 showed that the four items did not experience DIF. The fourth item was also not detected in DIF MIRT. This fact shows that multidimensionality on items 37, 54, 57, and 58 is independent in both MA and SMA groups. This means that there is no interaction between grouping variable (MA and SMA) to the multidimensionality. This is why the four items do not perform DIF. In other words, it can be said that there is no correlation between the grouping variables and the additional dimensions that cause the DIF.

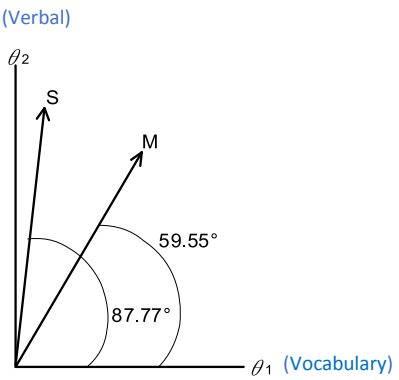

Figure 9. LOGI17 Visualization of the discrimination power vectors in the M and S groups.

In fact, the same item's complexity between groups can also performed DIF. This may occur due to the sensitivity of different items in uncovering the dimensions measured, related to the group being compared. It can also be said that there are other dimensions outside the four dimensions that have already been identified. These dimensions are not relevant to the AP construct and henceforth called irrelevant constructs (Haladyna & Downing, 2004; S. Messick, 1984).

The above findings are in line with Liaw (2015) who investigated comparisons of ability scores between focal and reference groups. The results showed that the primary abilities consistently favored the reference group, while at the same time also impaired the focal group whenever DIF items involved in the scoring.

DIF and Score Validity

Discussion on construct validity revolves around the issues whether or not the instrument or the items in measurement instrument works in accordance with the underlying theoretical conception. In line with the opinions of Grimm and Widaman (2012) which refer to the validity conception put forward by Messick (1995), DIF-free items indicate the fulfillment of internal validity evidence. When the model determined based on the internal structure of the AP SPMB, the DIF items have implications to the data matching with the pre-determined model. In other words, item's performance does not support the measurement model. As a result, the presence of DIF items make the matching of data with the model is reduced (Cheng et al., 2009). A clearer implication is the probability of an unequal correct answer in a group of participants with similar abilities. The inclusion of AP items containing DIF has implications on deviating the score from its original form. This is evidenced by the difference in estimated score generated based on the whole item when compared to the estimated score without involving the DIF items. This condition can be seen in Figure 10.

The magnitude of the academic potential of the estimated 10,000 participants (5,000 MA graduates and 5,000 SMA graduates), which then translate into

the ordinal variable, is presented in Figure 10. It appears that there is a stark difference between the estimates based on the overall items compared to the estimates based on the items that are DIF-free. The black line indicates the rank of the participant when the academic potential is estimated based on the whole items, while the red line (up and down) indicates an academic potential ranking estimate based on items that DIF free. The further implication of this is the highest order of scores changed when non-DIF items are excluded from capacity estimation. Thus, there is a possibility that among the 10,000 participants who supposed to enter the intended major eventually become out-of-reach due to the bias of the measurement results.

The preceding facts corroborate the opinions of some experts (e.g., Ackerman, 1994; Ackerman & Evans, 1994; Walker & Beretvas, 2003; Yen & Walker, 2007) who stated that DIF on the items causes inaccurate estimation on participant's parameter (latent attribute, θ). This implies the bias of items and tests, as well as the erroneous interpretation of scores. Figure 10 shows the ranking of participants becomes different when it involves the DIF items compared to DIF-free items. Capability estimation by involving items containing DIF yields a biased score. This resulted in the presence of both advantaged and disadvantaged participants, which asserts that the interpretation of the score is wrong. The erroneous interpretation of this score contaminates the validity (Messick, 1998; Yao & Li, 2010) so it can be said that DIF items have a marked effect on the validity of the test scores. Following Liaw's suggestion (2015), DIF items must be eliminated in the scoring.

Limitations

The data used in the research are empirical data, not followed by data simulation with scenarios of multidimensionality variations that may occur. The within item multidimensionality would be better evaluated if conducted simultaneously using multigroup confirmatory factor analysis (MGCFA). In addition, satisfactory explanations on what triggers DIF in the form of irrelevant constructs have not been satisfactorily disclosed. Beside of that, the model used in this study is limited to dichotomy model, where similar research is needed in the case of a politomous model.

Conclusion and Recommendations

The conclusions that can be drawn in this study are: SPMB-PTAIN AP items in 2012 revealed four dimensions: vocabulary (θ_1), verbal (θ_2), quantitative (θ_3), and symbol (θ_4). The construct structure of PA SPMB-PTAIN 2012 is semi-complex within-item multidimensional: verbal-vocabulary items reveal the vocabulary and verbal dimensions altogether; while the quantitative items - symbols both reveal the quantitative dimension and the dimensions of the symbol. Based on DIF UIRT, 12 items were found benefiting SMA graduates and three items benefiting MA graduates. According to MIRT, five items that benefited SMA graduates were found. Thus, multidimensionality differences between MA and SMA groups did not trigger the emergence of DIF (either UIRT or MIRT) on the items because the additional dimension unrelated to grouping variables (MA-SMA).

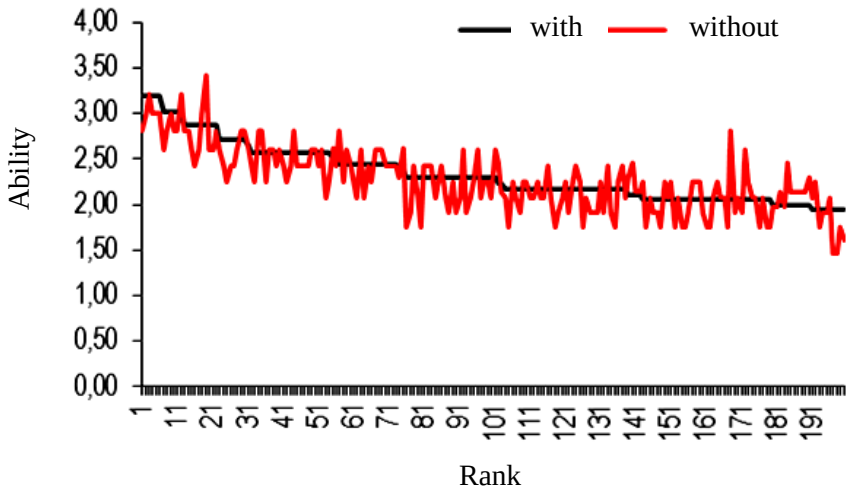

Figure 10. The amount of ability (θ) with and without the DIF item (n = 10,000).

Thus, the practical recommendations that can be given based on the results of this study are:

(1) For the scoring team, the SPMB-PTAIN scoring method should be converted from total score to composite based on the maximum information function of MIRT model so it can produce scores: (a) AP composite; (b) vocabulary; (c) verbal; (d) quantitative; and (e) symbols. In addition, scoring should be conducted after eliminating DIF-detected items;

(2) For the SPMB-PTAIN team of developers and test writers: (a) it is necessary to conduct a more in-depth content analysis of DIF-detected items so that improvements can be made to the future development of SPMB-PTAIN AP tests in order to improve construct validity the SPMB-PTAIN AP test, also ensure equality and fairness for those coming from MA - SMA; (b) DIF analysis should be used as one of the procedure for AP tests development process. Thus, when the test is widely used, the items are relatively free of DIF; and (c) the development of the items should only reveal the academic potential, not influenced by formal education background;

(3) For the SPMB committee, the development of the AP test should be submitted to professional agencies/bodies specialized in the testing field so that test handling becomes more professional;

(4) For psychometric researchers, since the findings in this study indicate that the claim of multidimensionality is the cause of the item containing an unfounded DIF, a further investigation on the source of DIF in the items is needed. In addition, the theme of items sensitivity in the context of MIRT needs to be explored more.

References

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 15*(1), 13-24. https://doi.org/10.1177/014662169101500103

Ackerman, T. A. (1992). *Assessing construct validity using multidimensional item response theory*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*(4), 255-278. https://doi.org/10.1207/s15324818ame0704_1

Ackerman, T. A., & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analyses. *Applied Psychological Measurement, 18*(4), 329-342. https://doi.org/10.1177/014662169401800404

Adams, R. J., Wilson, M., & Wang, W.-c. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement, 21*(1), 1-23.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In W. H. Angoff (Ed.), *Differential Item Functioning* (1st ed., pp. 3-24). New Jersey: Lawrence Erlbaum Associates.

Azwar, S., & Ridho, A. (2012). *Abilitas komposit dalam tes potensi*. Fakultas Psikologi Universitas Gadjah Mada.

Beretvas, S. N., Cawthon, S. W., Lockhart, L. L., & Kaye, A. D. (2012). Assessing impact, DIF, and DFF in accommodated item scores. *Educational and Psychological Measurement, 72*(5), 754-773. https://doi.org/10.1177/0013164412440998

Bolt, D. M., & Stout, W. F. (1996). Differential item functioning: Its multidimensional model and resulting sibtest detection procedure. *Behaviormetrika, 23*(1), 67-95. https://doi.org/10.2333/bhmk. 23.67

Brown, A. R., Finney, S. J., & France, M. K. (2011). Using the bifactor model to assess the dimensionality of the hong psychological reactance scale. *Educational and Psychological Measurement, 71*(1), 170-185. https://doi.org/10.1177/0013164410387378

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*(2), 129-147. https://doi.org/10.1177/014662169201600203

Cheng, Y.-Y., Wang, W.-C., & Ho, Y.-H. (2009). Multidimensional rasch analysis of a psychological test with multiple subtests: a statistical solution for the bandwidth--fidelity dilemma. *Educational and Psychological Measurement, 69*(3), 369-388. https://doi.org/10.1177/0013164408323241

Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals.

Educational and Psychological Measurement, 70 (5), 717-731. https://doi.org/10.1177/0013164410379322

Deng, N., Wells, C. S., & Hambleton, R. K. (2008). *A confirmatory factor analytic study examining the dimensionality of educational achievement tests*. Paper presented at the Northeastern Educational Research Association (NERA) Annual Conference, Rocky Hill, Connecticut.

Diones, R., Bejar, I. I., & Chaffin, R. (1996). The dimensionality of responses to SAT analogy items. *ETS Research Report Series, 1996*(1), i-31. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1996.tb01679.x

Dorans, N. J., Holland, P. W., & Educational Testing Service, P. N. J. (1992). *DIF detection and description: Mantel-haenszel and standardization*. Retrieved from http://ezproxy.lib.indiana.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED387526&site=eds-live&scope=site

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

Enright, M. K., Tucker, C. B., & Katz, I. R. (1995). A cognitive analysis of solutions for verbal, informal, and formal-deductive reasoning problems. *ETS Research Report Series, 1995*(1), i-31. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01641.x

ETS. (2007). Factors that can influence performance on the gre general test 2006-2007. *Test fairness and score use*. Retrieved from http://www.ets.org/Media/Tests/GRE/pdf/gre_0809_factors_2006-07.pdf

Finch, W. H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-based statistics for testing unidimensionality.*Applied Psychological Measurement, 31*(4), 292-307. https://doi.org/10.1177/0146621606294490

Furlow, C. F., Ross, T. R., & Gagné, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement, 33*(6), 441-464. https://doi.org/10.1177/0146621609331959

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24*(1), 3-14. https://doi.org/10.1111/j.1745-3992.2005.00002.x

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement, 40*(4), 281-306. https://doi.org/10.1111/j.1745-3984.2003.tb01148.x

Glanville, J. L., & Wildhagen, T. (2007). The measurement of school engagement: Assessing dimensionality and measurement invariance across race and ethnicity. *Educational and Psychological Measurement, 67*(6), 1019-1041. https://doi.org/10.1177/0013164406299126

Grimm, K. J., & Widaman, K. F. (2012). Construct validity. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (1st ed., pp. 621-642). Washington, DC, US: American Psychological Association.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 74*(2), 209-227. https://doi.org/10.1007/s11336-010-9158-4

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(11), S182-S188. https://doi.org/10.1097/01.mlr.0000245443.86671.c4

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

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology, 36*(2), 378-390. https://doi.org/10.1080/01443410.2014.946890

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using the full-information item bifactor analysis for graded response data: An illustration with the State Metacognitive Inventory. *Educational and Psychological Measurement, 68*(4), 695-709. https://doi.org/10.1177/0013164407313366

Jang, E. E., & Roussos, L. A. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach.

Journal of Educational Measurement, 44(1), 1-21.https://doi.org/10.1111/j.1745-3984.2007.00024.x

Kahraman, N., & Thompson, T. (2011). Relating unidimensional IRT parameters to a multidimensional response space: A review of two alternative projection IRT models for scoring subscales. *Journal of Educational Measurement, 48*(2), 146-164. https://doi.org/10.1111/j.1745-3984.2011.00138.x

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33*(7), 519-537. https://doi.org/10.1177/0146621608329504

Liaw, Y.-L. (2015). *When can multidimensional item response theory (MIRT) models be a solution for differential item functioning (DIF)? A Monte Carlo Simulation Study.* (Doctor of Philosophy), University of Washington, Seattle.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Martín, E. S., Pino, G. d., & Boeck, P. D. (2006). Irt models for ability-based guessing. *Applied Psychological Measurement, 30*(3), 183-203. https://doi.org/10.1177/0146621605282773

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*(2), 99-114. https://doi.org/10.1177/01466210022031552

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-57. https://doi.org/10.1177/014662168500900105

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(4), 728-743.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215-237. https://doi.org/10.1111/j.1745-3984.1984.tb01030.x

Messick, S. J. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Messick, S. J. (1998). *Consequences of test interpretation and use: The fusion of validity and values in psychological assessment.* Research Report No. 98-48. Princeton, NJ: Educational Testing Service.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*(2), 107-122. https://doi.org/10.1111/j.1745-3984.1993.tb01069.x

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses: Comparison of different approaches. *Journal of Educational Measurement, 31*(1), 17-35.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16*(3), 237-248. https://doi.org/10.1177/014662169201600304

Patarapichayatham, C., Kamata, A., & Kanjanawasee, S. (2012). Evaluation of model selection strategies for cross-level two-way differential item functioning analysis. *Educational and Psychological Measurement, 72*(1), 44-51. https://doi.org/10.1177/0013164411409743

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics 26: Psychometrics* (1st ed., pp. 125-167). Amsterdam: Elseveir.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5-15. https://doi.org/10.1111/j.1745-3992.2000.tb00033.x

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207. https://doi.org/10.1177/014662169001400208

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational & Psychological Measurement, 53*, 301-314. https://doi.org/10.1177/0013164493053002001

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412. https://doi.org/10.1177/014662168500900409

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: Springer.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19-31. https://doi.org/10.1007/s11136-007-9183-7

Ridho, A. (2011). *Multidimensionalitas Tes Potensi Akademik*. Paper presented at the Second International Conference of Indigenous and Cultural Psychology, Denpasar, Bali.

Ridho, A. (2014). *Invariansi sebagai Bukti Validitas Pengukuran.* Paper presented at the Pengembangan Instrumen Penilaian Karakter yang Valid, Solo.

Roussos, L. A., & Stout, W. F. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied psychological measurement, 20*(4), 355-371. https://doi.org/10.1177/014662169602000404

Roussos, L. A., & Stout, W. F. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*(1), 1-30.

Santelices, M. V., & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: An issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement, 72*(1), 5-36.

Scarpati, S. E., Wells, C. S., Lewis, C., & Jirka, S. (2011). Accommodations and item-level analyses using mixture differential item functioning models. *The Journal of Special Education, 45*(1), 54-62.

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research, 102*(3), 443-461. https://doi.org/10.1007/s11205-010-9682-8

Snow, T. K., & Oshima, T. C. (2009). A comparison of unidimensional and three-dimensional differential item functioning analysis using two-dimensional data. *Educational and Psychological Measurement, 69*(5), 732-747. https://doi.org/10.1177/0013164409332223

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23*(1), 63-86. https://doi.org/10.1080/08957340903423651

Stout, W. F. (1984). *A statistical procedure for assessing test dimensionality.* Retrieved from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat82session5.pdf

Stout, W. F. (1989). *A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation.* Cognitive Science Program. Champaign, IL: Department of Statistics - Univ. of Illinois

Stout, W. F. (2002). Psychometrics: From practice to theory and back (15 years of nonparametric multidimensional IRT, DIF/test equity, and skills diagnostic assessment). *Psychometrika, 67*(4), 485-518. https://doi.org/10.1007/BF02295128

Stout, W. F., & Nandakumar, R. (2006). DIMTEST 2.1 [Computer Software]. Missoula: Assessment System Corporation.

Stucky, B. D., Gottfredson, N. C., & Panter, A. T. (2012). Item-level factor analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (1st ed., pp. 683-697). Washington, DC, US: American Psychological Association.

Suh, Y. (2016). Effect size measures for differential item functioning in a multidimensional IRT model. *Journal of Educational Measurement, 53*(4), 403-430. https://doi.org/10.1111/jedm.12123

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*(3), 159-203. https://doi.org/10.1177/0146621603027003001

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research, 16*(Supplement 1), 33-42. https://doi.org/10.1007/s11136-007-9184-6

Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for diffe-*

rential item functioning [Computer software]. Chapel Hill: L. L. Thurstone Psychometric Laboratory, University of North Carolina.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning.* (1st ed., pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

van Abswoude, A. A. H., van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*(1), 3-24. https://doi.org/10.1177/0146621603259277

Vaughn, B. K., & Wang, Q. (2008). Classification based on tree-structured allocation rules. *Journal of Experimental Education, 76*(3), 315-340. https://doi.org/10.3200/JEXE.76.3.315-340

Vaughn, B. K., & Wang, Q. (2010). DIF trees: Using classification trees to detect differential item functioning. *Educational and Psychological Measurement, 70*(6), 941-952. https://doi.org/10.1177/0013164410379326

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40*(3), 255-275.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47*(3), 339-360. https://doi.org/10.1111/j.1745-3984.2010.00117.x

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement, 35*(1), 48-66. https://doi.org/10.1177/0146621610373095

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*(2), 83-105. https://doi.org/10.1177/0146621606291559

Yao, L., & Li, F. (2010). *A DIF detection procedure in multidimensional item response theory framework and its applications*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Colorado, Denver.

Yen, S. J., & Walker, L. (2007). *Multidimensional IRT models for composite scores*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Chicago, IL.

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 231-249.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG (Version 3). Lincolnwood, IL: Scientific Software International.

Zumbo, B. D. (2014). What role does, and should, the test standards play outside of the United States of America? *Educational Measurement: Issues and Practice, 33*(4), 31-33. https://doi.org/10.1111/emip.12052

(appendix follows)

Appendix

The Formulations of Items that Detected DIF Unidimensional (U) and Multidimensional (M) Based on MA and SMA Graduates Group

Item	U	M	Items Script
ANLG02	+		INSPIRATION (A) Suggestion (B) Creation (C) Ideas (D) Innovation (E) Revelation
ANLG09	+	+	ABSOLUTE (A) Liberal (B) Abstract (C) Relative (D) Free (E) Unstable
LOGI17	+	+	All students who study must pass the exam. Some of the students who passed the exam did not learn. So: (A) Every students study (B) Every students study and pass the exam (C) Half of students study and pass the exam (D) Every students study and did not pass the exam (E) Every students did not pass the exam
LOGI18	+	+	All students are required to prepare for a retrial. Some students get good grades in the retrial. So : (A) All students prepare and get good grades (B) Some students do not prepare and do not get good grades (C) Some students do not prepare themselves but get good grades (D) All students prepare (E) Not all students prepare themselves
LOGI20	+	+	All Tigers are meat eaters. Some animals are Tigers. So: (A) All meat eaters are Tigers (B) Some Tigers are meat eaters (C) Some animals are meat eaters (D) a, b, c is not the correct answer (E) (E) Not all animals are meat eaters
METK39	+	+	12! : 10! (A) 1,2 (B) 12 (C) 13 (D) 132 (E) 120
METK40	+		$4 \times 2^2 + 2^4 = ...$ (A) 50 (B) 80 (C) 32 (D) 36 (E) 16

(appendix continues)

Item	U	M	Items Script
METK41	+		$2 \times 3^2 + 3^4 = \ldots.$ (A) 40 (B) 64 (C) 69 (D) 96 (E) 99
METK45	+		45 x 0,5 = how many percent of 90? (A) 10 (B) 15 (C) 20 (D) 25 (E) 30
KOMP55	+		If x^2- 64 = 0 and y =8, then (A) x > y (B) x < y (C) x = y (D) 2x > y (E) X and y cannot be determined
KOMP56	+		If (1/3) compare to (5/7) then …. (A) 1 compare to 5 (B) 3 compare to 7 (C) 1 compare to 21 (D) 5 compare to 21 (E) 7 compare to 15
KOMP61	+		If x = the area of the square whose length is 21 cm and y = the area of the circle whose diameter 28 cm, then ...x > y (A) x < y ??? (B) x = y (C) 2x = 3y (D) x and y cannot be determined
ANLG05	-		CONTEMPORARY (A) Weird (B) Abstract (C) Ancient (D) Irregular (E) Current
ANLG13	-		WEATHER : METEOROLOGY (A) Physics: Astronomy (B) Book: Pedagogic (C) Descendants: Gerontology (D) Disease: Pathology (E) Fossils: Anthropology
METK52	-		If x is the rectangular side 25 cm^2 and y is the long side of a rectangle 50 cm wide with a short side 5 cm, what is xy? (A) 25 (B) 50 (C) 20 (D) 75 (E) 55

Note. Description: + = benefiting high school graduates; - = benefiting MA graduates; U = DIF is unimportant; M = DIF multidimensional.