Cukurova Medical Journal

Araştırma Makalesi / Research Article

Investigation of Theoretical Exams in Medical Faculties in Terms of Differential Item Functioning: Sample of Mersin University Medical Faculty

Tıp Fakültelerinde Uygulanan Teorik Sınavlarda Kullanılan Testlerin Ayırt Edici Madde İşlevselliği Açısından İncelenmesi: Mersin Üniversitesi, Tıp Fakültesi Örneği

Hüseyin Selvi[1], İbrahim Başhan[2], Gönül Aslan[3]

[1]Medical Education Department, Mersin Üniversity,

[2]Medical Education Department, Mersin Üniversity,

[3]Clinical Microbiology Department, Mersin Üniversity,

Cukurova Medical Journal 2014 39(2):240-247.

ABSTRACT

Purpose: Current study investigated whether the measurement tools utilized during medical faculty preclinical period to form the basis of pass-fail decisions taken for students included Differential Item Functioning for foreign and Turkish students.

Material and Methods: A total of 205 1st Period (1st year) students 7 of which were foreign nationals attending Mersin University Medical Faculty participated in the study. All analyses are performed by Easy-DIF software and Mantel Haenszel method. And Differential Item Functioning identified this 6 items presented to the 10 expert for expert opinion

Results: It was observed as a result of analyses via the method of Mantel Haenszel based on 7 exams that contained 100 multiple choice items each that only 6 items out of 700 included significant levels of Differential Item Functioning. It was also seen that item averages of foreign national students for the rest of all 694 items were lower compared to other students; however this difference was not statistically significant.

Conclusion: According to the expert opinion for 6 items; long sentence structure, negative item root structure and some words that regarded as traditional play a role in the formation of Differential Item Functioning was found. It can be offered that we must not use long sentence structure, negative item root structure and some words that regarded as traditional for preparing exams to ensure validity of exams.

Key Words: Differential Item Functioning, Medical Faculty Student's Achivement, Foreing Nationality Students

ÖZET

Amaç: Bu çalışmada, tıp fakültesi klinik öncesi dönemde kullanılan ve öğrenciler için bir üst sınıfa geçti-kaldı kararlarına dayanak oluşturan ölçme araçlarının, yabancı ve Türk uyruklu öğrenciler açısından Değişen Madde Fonksiyonu (DMF) içerip içermediği araştırılmıştır.

Materyal ve Metod: Çalışmaya Mersin Üniversitesi Tıp Fakültesinde öğrenim gören 7'si yabancı uyruklu olmak üzere toplam 205 dönem I (1. Sınıf) öğrencisi katılmıştır. Analizler Easy_DIF yazılımı ve Mantel Haenszel yöntemi kullanılarak yapılmıştır.

Bulgular: Mantel Haenszel yöntemi kullanılarak her biri çoktan seçmeli 100'er madde içeren 7 sınav üzerinden yapılan analizler neticesinde toplam 700 madde içerisinden yalnızca 6 tanesinde anlamlı düzeyde Değişen Madde Fonksiyonu gözlenmiştir. Geriye kalan 694 maddenin tamamında ise yabancı uyruklu öğrencilerin madde ortalamalarının diğer öğrencilere oranla daha düşük olduğu; ancak bu farkın Değişen Madde Fonksiyonu açısından anlamlı düzeye ulaşmadığı gözlenmiştir.

Sonuç: DMF'li tespit edilen 6 madde ilgili 10 alan uzmanının görüşüne sunulmuş ve uzmanlardan; devrik ve uzun cümle yapısının, olumsuz madde kökünün ve geleneksel sayılabilecek bazı kelimelerin Değişen Madde Fonksiyonu oluşumunda rol oynadığına yönelik dönüt alınmıştır. Buradan hareketle Tıp Fakültelerinde yapılan sınavlarda kullanılan testlerin geçerliklerin garanti altına alınması ve yabancı uyruklu öğrencilerin mağdur edilmemesi amacıyla devrik ve uzun cümle yapısı, olumsuz madde kökü ve geleneksel kelimeler içeren maddeler yazılmamasına dikkat edilmesi gerekmektedir.

Anahtar Kelimeler:Ayırt edici madde işlevselliği, Tıp Fakültesi başarı oranı, Yabancı uyruklu öğrenciler

INTRODUCTION

Training provided in medical faculties is mainly composed of two phases: preclinical and clinical. Preclinical period which coincides with the first three years basically consists of theoretical classes and practical implementations geared towards developing vocational skills. In the clinical period that lasts for another three years, students are basically given applied courses in addition to theoretical classes[1].

In order to graduate, physician candidates are required to obtain high academic achievement levels and necessary-sufficient vocational knowledge-skills at the end of the 6-year training.

Assessment and evaluation procedures are utilized throughout the training to ensure the acquisition of sufficient vocational information-skills and the students who are found unsatisfactory in this regard are required to repeat classes.

In order to be successful in the clinical period, students are required to obtain a passing grade from multiple choice committee exams that contain 2 scores (1-0) for each 100 items.

These exams are highly influential for pass-fail decisions taken for students. Making the right decisions on behalf of the students based on these exams is rather important for several principles such as justice and equal opportunities in education. In order for a measurement tool to facilitate making right decisions, it has to meet two basic requirements: Reliability and validity.

Reliability is the competence of the measurement tool to make assessments without error and it is a prerequisite for validity. Validity, in general, is the competence of the measurement tool to evaluate the quality that is needed to be assessed without confusing it with other qualities[2,3,4,5.]

When it is investigated in terms of measurement tools (theoretical exams) used in the preclinical period in medical training (Periods 1-2-3), reliability is affected by several variables such as the length of test, group homogeneity, average item difficulty and cheating behaviors. Validity is a wider term that includes reliability as well and is affected from many different variables such as the construct of the variables that needs to be measured, measurement tool's level of coverage in the related topic, the purpose of its use and bias.

Item and test bias, one of the variables that negatively affect validity, is a situation often observed in exams in which foreign nationals attend. Similarly, biased items can be seen in exams given to students who come from various parts of the country and who have different socio-cultural, economic etc backgrounds[6,7,8.]

The concept of bias has been defined:

- As "a systematic change that interferes with the assessment process" by Osterling (1983),
- As "different processing of the item or the test in different sub groups which causes systematic change" by Camilli & Shepard (1994) and
- As "the difference of answering test items correctly among individuals in different sum groups with the same competency levels by Angoff (1993)[7.]

As can be seen from the definitions above, bias can be defined as the situation in which the individuals in different sub groups (cultural, socio-economic, nationality etc) with the same mental competence cannot correctly answer a test item although they know the answer due to some characteristics based on the item or vice versa.

This situation is distinctly observed in foreign national students who know the correct answers to the questions however cannot understand the items due to incompetence in their vocabulary.

Studies regarding these types of situations (bias) are handled among validity studies and rather comprehensive methods along with advanced computer infrastructures-theories are used. Whether an item or a test has bias can be understood basically in two phases. The first phase identifies whether the item is processed differently in various sub groups and this phase is called Differential Item Functioning (DIF). The second phase consists of presenting the items with DIF to be viewed by experts and searches the reason for the divergent processing of the item by different sub groups.

One of the widely used methods in the first phase is the Mantel-Haenszel (M-H) method since it is based on Contingency Table methods and it does not include premises difficult to satisfy.

In this method, odds ratios obtained from all sub group members that are matching in terms of total scores are combined by taking group weights into consideration.

$$\alpha_{MH} = \frac{\sum_j \frac{p_{r_j} q_{f_j} n_{r_j} n_{f_j}}{n_j}}{\sum_j \frac{q_{r_j} p_{f_i} n_{r_j} n_{fi}}{n_j}} \qquad \beta_{MH} = \log_e(\alpha_{MH})$$

p_{r_j} =Item difficulty of the related item in *j* score level

q_{f_j} = Incorrect answer ratio of the related item in *j* score level

As can be seen from the formula, logarithm of the α value is obtained to attain β values. Positive β values show reference (Türkish National Students) whereas negative values show item effect on behalf of the focal (Foreign National Students) group. β values of “0” shows that there is no item effect[7].

Mantel-Haenszel (1959) developed Chi-Square statistics below in order to prove the statistical significance of α_{MH} and β_{MH} values that give item effect measurements[6].

$$MH\chi^2 = \frac{\left\{\sum_{j=1}^{s}(Aj - E(Aj)) - \frac{1}{2}\right\}^2}{\sum_{j=1}^{s} VAR(Aj)} \qquad VAR(Aj) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)} \qquad E(Aj) = \frac{n_{Rj} m_{1j}}{T_j}$$

Aj = Number of correct answers for the related item in reference group’s *j* score level

$E(Aj))$ = Expected number of correct answers for the related item in reference group’s *j* score level

T_j = Total number of individuals in *j* score level

m_{1j} = Total number of individuals with correct answers in focus and reference groups in *j* score level

m_{0j} = Total number of individuals with incorrect answers in focus and reference groups in *j* score level (including the blank answers as well).

Here, the value of $MH\chi^2$ shows χ^2 distribution in 1 degree of freedom for Uniform DIF[6].

By using the Mantel Haenszel method, current study also aimed to identify whether the assessment tools utilized during Medical Faculty Period 1 theoretical exams included items with DIF and to increase the quality (reliability and validity) of the measurement tools used in these exams.

Since students from different backgrounds (cultural, economic, social, nationality etc) are trained in Medical Faculties, it is important to determine whether items in faculty tests contain DIF to ensure making right decisions about passing or failing the medical students.

The study sought answers to the question "Are there any items with DIF in Medical Faculty Period 1 theoretical exams for foreign national students?"

MATERIALS and METHODS

This section includes information regarding the type of study, the working group and the data analysis.

Type of Study: Since the study investigated whether there were items with DIF in theoretical exams in terms of nationality variable, it can be termed as a descriptive study.

Working Group: Current study was implemented on a total of 205 1st Period students in Mersin University Medical Faculty who sat for the theoretical exams during 2012-2013 educational year. 121 of the participating students were males (59.02%) and 84 (40.08%) were females. Sampling process was not undertaken since all data was obtained. Table I provides the nationality and gender distribution for the participating students.

Table 1: Nationality And Gender Distribution For The Participating Students

Nationality Of Students	n	%	$\bar{X}$
Türkish National Students (Reference Group Grup)	198	96,6	65,92 (std:4,65)
Foreign National Students (Focal Group)	7	3,4	49,17 (std:7,61)

Item universe for the study was composed of a total of 700-item multiple choice test items used in the 7 theoretical exams (excluding the make-up exams) containing 100 items each that were implemented on Period I students during 2012-2013 educational year.

Data Analysis

First of all, 7 (exams) x 100 (items) x 205 (students) = 143500 pieces of data (obtained from all exams, items and individuals) were transferred to digital environment. In data description, central tendency (mean) and distribution (standard deviation) measures were utilized and exam reliability was calculated with the help of KR-20 coefficient. Since the distribution was not normal, differences among item averages in groups were investigated with the help of Mann-Whitney U Test.

Mantel-Haenszel (M-H) method was used in DIF analysis of the data. Since this method is based on Contingency Table methods (non-parametric), there was no need for the data to meet the premises in the parametric methods[6.] Since M-H method contains Chi-Square test in itself, the value of p<0.05 was considered to be statistically significant in this phase. All analyses utilized during the research were undertaken via Easy-DIF software[9].

DIF identified this 6 items presented to the 10 expert for expert opinion. Then simple consistency coefficient was determined for checking consistency between experts.

RESULTS

Table 2 provides the descriptive statistics, reliability coefficients and Mann-Whitney U test results regarding the committee and final exams composed of a total of 700 items given in Medical Faculty Period I. Table 3 presents the findings obtained during DIF identification studies.

Table 2: Descriptive Values and Reliability Coefficients for the Exams

		n	$\bar{X}$	Std.	Skewness	Kurtosis	KR-20	Mann-Whitney U	
								Z	p
Exam 1	Focal Group	7	54,35	14,7	-0,4	0,84	0,843	-10,35	0,00
	Ref. Group	198	77,93	9,4	-1,31	4,34			
Exam 2	Focal Group	7	46,44	20,06	0,01	-0,75	0,781	-4,41	0,00
	Ref. Group	198	58,95	15	-0,04	-0,01			
Exam 3	Focal Group	7	52,40	20,25	0,20	-0,17	0,870	-6,33	0,00
	Ref. Group	198	68,05	11,5	-0,40	0,03			
Exam 4	Focal Group	7	49,57	23,86	0,16	-0,93	0,833	-4,78	0,00
	Ref. Group	198	64,89	12,48	0,42	-0,21			
Exam 5	Focal Group	7	49,14	19,74	0,02	-0,54	0,876	-4,37	0,00
	Ref. Group	198	60,48	12,41	025	-0,61			
Final 1. Term	Focal Group	7	45,89	21,53	0,18	-0,51	0,603	-6,16	0,00
	Ref. Group	198	64,01	15,32	-0,25	-0,58			
Final 2. Term	Focal Group	7	46,45	20,48	0,01	-0,64	0,857	-7,09	0,00
	Ref. Group	198	67,16	14,35	-0,28	-0,74			

Table 3: DIF Identified Items And Their Distribution Among Exams

	DIF Identified Items	Mean Of Focal Group	Mean Of Reference Group	MH	p
Exam 1	Item 7	0,37	0,85	4,62	0,03
	Item 17	0,37	0,79	3,98	0,04
Exam 3	Item 15	0,37	0,76	5,33	0,02
	Item 67	0,37	0,82	5,39	0,02
	Item 90	0,50	0,79	4,64	0,03
Exam 5	Item 76	0,25	0,78	4,59	0,03

Investigation of Table 2 shows that KR-20 values of the exams change between 0,603 and 0,876. These values indicate that the exams are generally reliable. Table also shows that foreign national students in the focus group have lower grade point averages than those of Turkish students in the reference group. As Table 2 points out, this difference is significant for all exams (p<0,01).

Investigation of the data in Table 3 shows that 6 of the 700 items contain DIF according to Mantel Haenszel method. Two of these six items were observed in Committee I, three in Committee III and one in Committee V exams.

According to expert opinion, long sentence structure, not clear, overturned and negative item root structure and some of can be considered as traditional words (less used words by public) has been found that the formation of DIF.

Thus, these items have been judged to be biased because these items are less understandable by foreign national students.

Simple consistency coefficient between experts was found 0.70 for 5 items and 1.00 for 1 item. According to Erkuş (2012) 0.70 and over values of simple consistency coefficient is enough for consistency between experts[10.]

DISCUSSIONS

Alfred Binet who laid the foundation on bias research in 1910 mention in his studies that answers individuals from different social, cultural and economic contexts provide to some items are affected from several variables such as attention, education at home, knowledge of language, nationality etc in addition to mental skills[6]. Camilli & Shepard (1994) state that these items lower the validity of the exams and therefore should be identified[6.]

Current study investigated the theoretical exams that are used as a foundation to make decisions about medical faulty students' passing/failing their grades in terms of foreign national students and it was observed that 6 of the 700 items contained significant levels of DIF. It was observed that all but one of these 6 items has uniform DIF and this item contained DIF only at low skill levels and did not contain DIF at the other skill levels. In the rest of the 694 items foreign national students were seen to have lower item averages compared to other students however this difference did not reach significant levels in terms of DIF. In line with these findings, it can be claimed that exams in general do not contain items with DIF and therefore have high validity.

In their study Biomer et. al. (1998) stated that the ratio of DIF identified in items used in scales prepared for implementations on different nations were lower than those prepared to be used singly[11]. Since theoretical exams given in Period I are implemented in block hours of 100 items, the finding of lower DIF ratio is consistent with the findings of the mentioned study.

Literature suggests undertaking DIF analyses in terms of different variables and identifying the variables that cause DIF in order to have quality exams[6,7,8.]

Existence of items with DIF in medical faculty exams based on nationality variable was identified in the analyses undertaken in the current study.

As a result of expert opinions, long sentence structure, not clear, overturned and negative item root structure and some of can be considered as traditional words (less used words by public) has been found that the formation of DIF. Similarly, item writing suggestions of using simple and clear language, avoidance of words that can not be understood by students, the root of item and its options musn't extend by unnecessary sentences can be found generally in the literature[2,5]. ·In this context, especially in examinations which has participation of foreign students, developing the general rules of item writing studies are recommended.

Although the rate of items with DIF was found to be lower compared to total number of items, it is suggested to continue these studies in the framework of validity studies and to identify the variables that causes DIF in order to have quality exams.

Current study only investigated the existence of DIF in theoretical exams in terms of nationality variable. Similar studies should be undertaken in terms of different variables such as gender, ethnic background etc.

Mantel Haenszel method was used in DIF investigation. The method is often preferred since this method does not contain premises that are difficult to meet, includes chi-square test in itself and is rather easy to calculate[6,7.]. However, significant differences between exam averages of Turkish and foreign national students and the fact that foreign nationals have a higher rate of failure compared to Turkish students necessitate the use of different methods developed in the framework of Item Response Theory in these studies.

REFERENCES

1. Mersin University, Medical Faculty, Training Guide, 2012-2013.
2. Murphy, K.R. & Davidshofer, C.O. *Psychological testing: principles and applications*. Prentice Hall, New Jersey, USA. 2001.
3. Anastasi, A., Urbina, S., *Psychological testing.* 7th.ed. New Jersey: Prentice Hall. 1990
4. Thorndike, R.L., *Applied Psychometrics*. Houghton Mifflin Company, Boston. 1982
5. Aiken L., *Psychological testing and assessment*. Allyn&Bacon, USA. 2000.
6. Camili, G., Shepard, L.A. *Methods for identifying biased test items.*Sage Publication, London. 1994
7. Holland, P.W. & Wainer, H. (Editör). *Differential item functioning*. Lawrence Erlbaum Associates, Publishers, New Jersey. 1993.
8. Osterling, S.J. *Test item bias*. Sage Publication, London.1983.
9. Andres, G., Padilla, J.L., Hidalgo, M.D., Benito, J.G., Benitez, I. EASY-DIF: Software for Analyzing Differential Item Functioning Using the MantelHaenszel and Standardization Procedures. Applied Psychological Measurement. 2010.
10. Erkuş, A. Psikolojide ölçme ve ölçek geliştirme-ı: temel kavramlar ve işlemler. Ankara: Pegem Akademi.2012;85.
11. Biomer, J.B., Krehr, S., Ware, J.E., Damsqaard, M.T., Bech, F. Differential Item Functioning in the Danish Translationof the SF-36. Journal Of Clinical Epidemiology. 1998;51;1189-1202.

Yazışma Adresi / Address for Correspondence:

Dr. Hüseyin Selvi
Mersin Üniversity,
Specialist of Measure and Evaluation Medical Education Department,
TR-33343 MERSİN
e mail: hsyn_selvi@yahoo.com.tr

geliş tarihi/received :20.09.2013
kabul tarihi/accepted:21.10.2013